The Maximum Order Complexity of Sequence Ensembles

CEES J.A. JANSEN
Philips Crypto B.V.

Abstract

In this paper we extend the theory of maximum order complexity from a single sequence to an ensemble of sequences. In particular, the maximum order complexity of an ensemble of sequences is defined and its properties discussed. Also, an algorithm is given to determine the maximum order complexity of an ensemble of sequences linear in time and memory. It is also shown how to determine the maximum order feedback shift register equivalent of a given ensmble of sequences, i.e. including a feedback function. Hence, the problem of finding the absolutely shortest (possibly nonlinear) feedback shift register, that can generate two or more given sequences with characters from some arbitrary finite alphabet, is solved. Finally, the consequences for sequence prediction based on the minimum number of observations are discussed.

1 Introduction

The notion of maximum order complexity of sequences was introduced in [8] as the length of the shortest feedback shift register that can generate a given (part of a) sequence, where the feedback function may be any function, mapping states onto characters. The import of maximum order complexity is that it tells exactly how many keystream characters have to be observed at least, in order to be able to generate the entire sequence by means of a feedback shift register of that length. Also maximum order complexity can be viewed as an additional figure of merit to judge the randomness of sequences.

We recall a number of results as given in [8, 9, 10]:

- the typical complexity profile closely follows the $2 \log l$ curve
- there exists a linear time and memory algorithm to determine the maximum order complexity profile of a given sequence; this algorithm in fact builds a directed acyclic word graph (DAWG) from the sequence
- generally, there exists a class of feedback functions which all give rise to one and the same sequence (even for periodic sequences)

D.W. Davies (Ed.): Advances in Cryptology - EUROCRYPT '91, LNCS 547, pp. 153-159, 1991.

- determining the maximum order feedback shift register equivalent of a given sequence of length l has an expected order of $2l^2 \log l$, where the feedback function is given by its algebraic normal form, however for DeBruijn sequences the order is $l \log l$

- the DAWG can efficiently be used for predicting successive characters of a given sequence after the least number of observations.

In this paper we consider the following problem. Suppose that a number of sequences with characters from some finite alphabet are given. What is the shortest feedback shift register that can generate all the given sequences, i.e. one FSR with one fixed, possibly nonlinear feedback function? To this end, the theory of maximum order complexity is generalized to include the multiple sequence case. In particular, the maximum order complexity of an ensemble of sequences is defined and its properties examined in Section 2. Section 3 deals with a multiple sequence DAWG to determine the ensemble complexity. Finally, in Section 4 it is considered how to determine the maximum order FSR equivalent of an ensemble of sequences, and how to resynthesize sequences with the DAWG.

2 The Maximum Order Complexity of an Ensemble of Sequences

Let $\mathcal{Z}$ denote an ensemble of N sequences $\underline{z}_i = (\alpha_{0,i}, \alpha_{1,i}, \ldots, \alpha_{l_i-1,i})$ of lengths l_i, $1 \leq i \leq N$, with characters $\alpha_{j,i} \in \mathcal{A}$, where the alphabet $\mathcal{A}$ is some finite set. How many sections (i.e. memory cells) should a feedback shift register at least have in order to generate all N sequences of $\mathcal{Z}$? So regardless of what the (memoryless) feedback function would have to be, linear or nonlinear. Analogous to the single sequence situation, the following definition of maximum order complexity is proposed:

Definition 1 *The maximum order complexity $c(\mathcal{Z})$ of an ensemble $\mathcal{Z}$ containing N sequences $\underline{z}_i = (\alpha_{0,i}, \alpha_{1,i}, \ldots, \alpha_{l_i-1,i})$ of lengths l_i, $1 \leq i \leq N$, with characters $\alpha_{j,i} \in \mathcal{A}$, where the alphabet $\mathcal{A}$ is some finite set, is defined to be the length L of the shortest feedback shift register for which there exists a memoryless feedback mapping, such that the FSR can generate all N sequences of $\mathcal{Z}$.*

This definition clearly implies that all the sequences of $\mathcal{Z}$ are uniquely identified by their first $c(\mathcal{Z})$ characters, assuming that all sequences of $\mathcal{Z}$ are distinct. In the case that only periodic sequences are considered, clearly any consecutive $c(\mathcal{Z})$ characters uniquely identify any of the sequences of $\mathcal{Z}$. The periodic case is in fact more general, as it also covers the situation that an observed sequence does not necessarily start with the first character of a given sequence. These observations and the fact that one can easily construct an example in which two sequences can be identified by a subsequence of length much shorter than $c(\mathcal{Z})$ characters, are the motivation for the following definition:

Definition 2 *The ensemble complexity of an ensemble $\mathcal{Z}$ containing N sequences $z_i = (\alpha_{0,i}, \alpha_{1,i}, \ldots, \alpha_{l_i-1,i})$, of lengths l_i, $1 \leq i \leq N$, with characters $\alpha_{j,i} \in \mathcal{A}$, where the alphabet $\mathcal{A}$ is some finite set, denoted by $c_\mathcal{E}(\mathcal{Z})$, is defined as the shortest length ℓ, such that any subsequence of length ℓ uniquely identifies each of the N sequences $z_i \in \mathcal{Z}$.*

Definition 2 has a clear Information Theoretic significance, as shown in [8, Ch. 7], i.e. it tells how many characters have to be observed, in order to know exactly to which sequence these characters belong.

One can easily see that the following inequalities hold for the ensemble complexity:

$$\log|\mathcal{Z}| \leq c_\mathcal{E}(\mathcal{Z}) \leq \max_i l_i, \quad \text{and} \tag{1}$$

$$\log|\mathcal{Z}| \leq c_\mathcal{E}(\mathcal{Z}) \leq \max_i p_i, \quad \text{for periodic sequences.} \tag{2}$$

Now let $\hat{c}(\mathcal{Z})$ denote the maximum of all the individual maximum order complexities, i.e.:

$$\hat{c}(\mathcal{Z}) := \max_{z_i \in \mathcal{Z}} c(z_i). \tag{3}$$

The relation between $c(\mathcal{Z})$, $c_\mathcal{E}$ and $\hat{c}$ is given in the next proposition.

Proposition 1 *Let $c(\mathcal{Z})$, $c_\mathcal{E}(\mathcal{Z})$ and $\hat{c}(\mathcal{Z})$ be as defined before. Then the inequality $\hat{c}(\mathcal{Z}) \leq c(\mathcal{Z}) \leq \max(\hat{c}(\mathcal{Z}), c_\mathcal{E}(\mathcal{Z}))$ holds for any ensemble $\mathcal{Z}$.*

The proof is easily obtained by using the definitions. A direct consequence of Proposition 1 is that, if $c_\mathcal{E} < \hat{c}$, then $c(\mathcal{Z}) = \hat{c}$. For the special case of periodic sequences it can easily be shown that $c(\mathcal{Z}) = \max(\hat{c}(\mathcal{Z}), c_\mathcal{E}(\mathcal{Z}))$.

Note that from the definition of maximum order complexity it can be seen that the highest value in the auto- or crosscorrelation function of one or two sequences is lowerbounded by the maximum order complexity of the sequence or the ensemble of two sequences.

Typical Complexities

Let us consider an ensmble of randomly chosen sequences. From the results of Arratia et al. [1, 2, 3] we know that the length of the longest subsequence which is common to all sequences in the ensemble tends towards $\sum_i \log l_i/(N-1)$. This value can therefore be seen as a lowerbound to the expected ensemble complexity. However, $c_\mathcal{E}$ does not deal with the longest common subsequence, but rather with the shortest length such that all subsequences uniquely identify all sequences. Hence, for $N = 2$ the lowerbound is significant, but for $N > 2$ it is not.

An obvious lowerbound for the expected value of $\hat{c}$ is the expected complexity of the longest sequence in the ensemble, i.e. $2 \max_i \log l_i$.

An upperbound to all complexities is obtained by considering all sequences in the ensemble as constituting one sequence of length $\sum l_i$, i.e. $2 \log \sum_i l_i$.

Statistical experiments seem to indicate that indeed all complexities are close to the upperbound.

The Class of Feedback Functions

In the case of an ensemble of sequences $\mathcal{Z}$ the maximum order feedback shift register equivalent is defined as the FSR of length $c(\mathcal{Z})$ and a feedback function such that the FSR can generate all the sequences $\underline{z}_i \in \mathcal{Z}$. As usual we restrict ourselves to periodic sequences of characters which are elements from some finite field $GF(q)$. This situation is typical for nonsingular nonlinear feedback shift registers, which may have a cycle structure with many distinct cycles

For finite field sequences it is customary to use the truth table to derive an analytical expression for the feedback function. The memory cells of the FSR provide for the argument values and hence the truth table is determined by all the occurring FSR states or equivalently all sequences $\underline{z}_i \in \mathcal{Z}$.

In general not all q^c possible FSR states occur in a particular ensemble. Consequently, there exists a class $\Phi_{\mathcal{Z}}$ of feedback functions which all give rise to the same ensemble of sequences $\mathcal{Z}$. The number of feedback functions in this class is given by the following proposition:

Proposition 2 *Let $\Phi_{\mathcal{Z}}$ denote the class of feedback functions of the maximum order feedback shift register equivalent of the ensemble $\mathcal{Z}$ of periodic sequences $\underline{z}_i$ over $GF(q)$, where $\mathcal{Z}$ has maximum order complexity c and the $\underline{z}_i$ are periodic with periods p_i. Then the number $|\Phi_{\mathcal{Z}}|$ of functions in the class $\Phi_{\mathcal{Z}}$ satisfies:*

$$|\Phi_{\mathcal{Z}}| = q^{q^c - \sum p_i}.$$

This result is in contrast with linear complexity where the feedback function is unique for periodic sequences. Hence, $\Phi_{\mathcal{Z}}$ in general contains more than one function and one is able to search for functions exhibiting certain properties such as nonsingularity, the least order product function or the function with the least number of terms.

Proposition 2 also confirms that $c(\mathcal{Z}) \geq \log_q \sum p_i$. For an ensemble, containing only dual sequences ([8, pg. 44]) $\Phi_{\mathcal{Z}}$ contains exactly one feedback function.

The above considerations have an interesting impact on generators, capable of generating a number of N binary periodic sequences of period 2^n, such as DeBruijn sequences [6] and Run Permuted sequences [8, 11]. To generate an ensemble $\mathcal{B}$ of these sequences, a FSR of length $c(\mathcal{B})$ is needed, with

$$c(\mathcal{B}) \geq n + \log N. \tag{4}$$

Inequality (4) shows that for DeBruijn sequences the FSR length increases with at least one bit each time the number of sequences is doubled. However, for an ensemble of Run Permuted sequences of equal complexities, the complexity of the ensemble could be equal to the complexities of the individual sequences.

To generate an ensemble $\mathcal{D}$ of all $2^{2^{n-1}-n}$ DeBruijn sequences of order n requires a FSR of length at least 2^{n-1} according to (4). However, it can be shown that $c(\mathcal{D}) = 2^n - 2$, which makes it infeasible to generate $\mathcal{D}$ by means of one FSR.

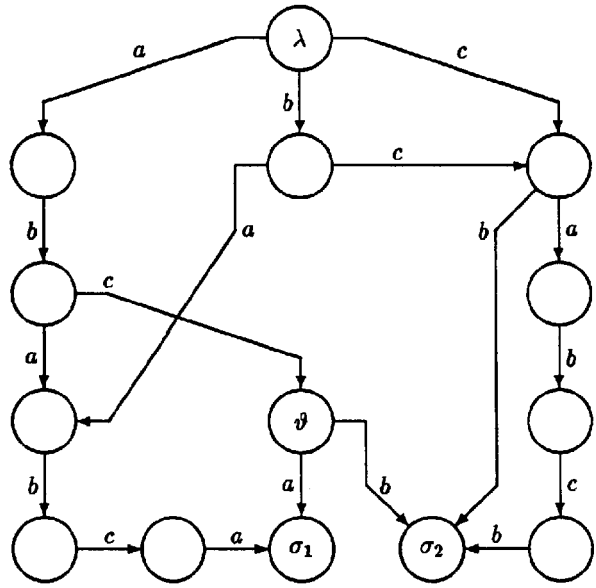

Figure 1: DAWG of *ababca* and *bcabcb*.

3 A Multiple Sequence DAWG

In [4] Blumer et al. describe a linear-time and -memory algorithm to build a *Directed Acyclic Word Graph* (DAWG) from a given string of letters, using a mechanism of suffixpointers. This DAWG is then used to recognize all substrings (or words) in the string. In [9, 10] it was shown how to use this algorithm for determining the maximum order complexity profile of a single given sequence.

Blumer et al. generalized their algorithm to build a DAWG from two or more given strings of letters, as described in [5]. We have subsequently adapted Blumer's algorithm to determine $c(\mathcal{Z})$ and $c_{\mathcal{E}}(\mathcal{Z})$ – in fact their profiles – linear in time and memory.

Example 1 Consider the following two sequences over $\{a, b, c\}$: $\underline{z}_1 = (a, b, a, b, c, a)$ and $\underline{z}_2 = (b, c, a, b, c, b)$.
Their respective MOC profiles are (0,1,1,1,3,3) and (0,1,1,1,1,3). The profile of $c(\underline{z}_1 \cup \underline{z}_2)$ is (0,1,1,1,3,3,3,3,3,3,3,4), whereas the $c_{\mathcal{E}}(\underline{z}_1 \cup \underline{z}_2)$ profile is (2,3,4,4,4,4). The corresponding DAWG (without the suffix pointers) is depicted in Figure 1. The profile of $c(\underline{z}_1|\underline{z}_2)$ is (0,1,1,1,3,3,3,3,3,3,3,7), where $\underline{z}_1|\underline{z}_2$ denotes the concatenation of $\underline{z}_1$ and $\underline{z}_2$.

As for the single sequence case, Blumer's algorithm can be used for various other purposes by postprocessing the DAWG. For example, to find a subsequence in any of the given sequences is an operation which is linear in the subsequence length. It is

also possible to generate (any part of) any of the given sequences, based on the least number of observed characters and simultaneously determining – also after the least number of observed characters – to which of the sequences the characters belong.

The aforementioned applications can be run on a simple DOS compatible personal computer for sequence lengths of several thousands of characters.

4 Synthesis of the MOFSR Equivalent

To actually construct the shortest feedback shift register which can generates a given ensemble of sequences of lengths l_i one first determines its maximum order complexity and then one determines a feedback function in the class Φ_Z. The first operation has order proportional to $L := \sum l_i$ and is expected to yield a complexity value of $2 \log L$. The second operation can be performed with a technique called the Algebraic Normal Form Transform (see [7, 8]), which is a fast transform resulting in the algebraic normal form of the feedback function, assuming that the non-specified truth table entries are either all zeroes or ones. This ANF transform has order $c2^c$ for the binary case, where c denotes the number of truth table variables. Hence, the expected order of the MOFSR synthesis procedure is $L^2 \log L$.

Example 2 Consider the ensemble S of the two dual sequences $\underline{s}_1 = (11010)^\infty$ and $\underline{s}_2 = (11110010000)^\infty$. In this case $c(S) = 4$ and the truth table is completely specified. The feedback function is $F(x_0, x_1, x_2, x_3) = 1 + x_0 + x_1 + x_2 + x_1x_3 + x_2x_3$.

5 Conclusions

We have generalized the theory of maximum order complexity to the case of multiple sequence ensembles. To this end, new definitions of maximum order complexity and ensemble complexity were introduced and the consequences for some special sequence sets, such as DeBruijn sequences and Run Permuted sequences, examined. The described results provide a better understanding of nonlinear feedback shift registers and the complexity of sequences.

The practical import of this theory has been enhanced by the adaptation of Blumer's generalized algorithm, which allows for an efficient determination of the complexity of a given ensemble of sequences, and can be applied for sequence prediction.

References

[1] R. Arratia and M. S. Waterman. "Critical Phenomena in Sequence Matching", *The Annals of Probability*, vol. 13, no. 4, pp. 1236–1249, 1985.

[2] R. Arratia and M. S. Waterman. "An Erdös–Rényi Law with Shifts", *Adv. in Math.*, vol. 55, pp. 13–23, 1985.

[3] R. Arratia, L. Gordon and M. S. Waterman. "An Extreme Value Theory for Sequence Matching", *The Annals of Statistics*, vol. 14, no. 3, pp. 971–993, 1986.

[4] A. Blumer, J. Blumer, A. Ehrenfeucht, D. Haussler and R. McConnell. "Linear Size Finite Automata for the Set of all Subwords of a Word: An Outline of Results", *Bul. Eur. Assoc. Theor. Comp. Sci.*, no. 21, pp. 12–20, 1983.

[5] A. Blumer, J. Blumer, D. Haussler, R. McConnell and A. Ehrenfeucht. "Complete Inverted Files for Efficient Text Retrieval and Analysis", *JACM*, vol. 34, no. 3, pp. 578–595, July 1987.

[6] H. Fredricksen. "A survey of full-length nonlinear shift register cycle algorithms", *SIAM Rev.*, vol. 24, pp. 195–221, April 1982.

[7] C. J. A. Jansen and D. E. Boekee. "The Algebraic Normal Form of Arbitrary Functions of Finite Fields", *Proceedings of the Eighth Symposium on Information Theory in the Benelux, Deventer, The Netherlands*, pp. 69–76, May 1987.

[8] C. J. A. Jansen. *Investigations On Nonlinear Streamcipher Systems: Construction and Evaluation Methods*, PhD. Thesis, Technical University of Delft, Delft, 1989.

[9] C. J. A. Jansen and D. E. Boekee. "The Shortest Feedback Shift Register That Can Generate A Given Sequence", *Proceedings of Crypto '89, Santa Barbara, USA.*

[10] C. J. A. Jansen and D. E. Boekee. "On the Significance of the Directed Acyclic Word Graph in Cryptology", *Proceedings of Auscrypt '90, Sydney, Australia.*

[11] C. J. A. Jansen "On the Construction of Run Permuted Sequences", *proceedings of Eurocrypt '90, Århus, Denmark.*