Secure DNA data compression using algebraic curves

I. Soto[✉], I. Jiron, A. Valencia and R. Carrasco

A system that achieves compression using artificial DNA packaging with the support of two algebraic curves is presented, whereby the Hermitian channel code algorithm introduces gain and safety. Additionally, performance results are presented with a gain of 7 dB against uncoded quadrature phase shift keying and 1 dB against McEliece, for a bit error rate of 10^{-3}. The results of the security levels compared with the McEliece system are also presented.

Introduction: Currently, in wireless communication systems due to the large amount of information transmitted, it is necessary to optimise the bandwidth and increase the speed of data transmitted by compression of the messages and channel coding. However, there are problems associated with high-speed transmission in wireless communication systems because they experience a high degree of interference [1]. Therefore, methods and technologies need to be more efficient and stronger in terms of security to be resilient to the attacks.

In [2], Vardy showed that one of the strongest one-way functions for the intractability problem is implemented using distances for codes constructed over a field for the McEliece cryptosystem [3]. In [4], Atito *et al.* introduced the artificial DNA packing with chaotic map to use stenography and/or encrypt information for the purpose of improving and increasing security level, privacy and authentication. However, synchronisation can be lost by the use of different substrates with different computer clocks [4].

For that reason it has been proposed, using chaotic maps Galois fields (GF) with the DNA packing [5], to generate artificial DNA sequences with chaotic map over GFs to hide information. The generation of a very large carrier sequence (S_c) still persists, underestimating the problem of performance in terms of spectral efficiency [5]. However, as promoters (P) and terminators (T) can be unique, they can be found in the DNA strand by an attack called basics of side-channel, which analyses the pattern of power [6]. Moreover, eventually theoretical threads containing chaotic sequences between these markers can be broken by techniques which identify the Lyapunov exponent and other parameters of the chaotic map [7].

This Letter presents a communication system that achieves an increase in the data rate transmitted using wide bandwidth with high encryption by employing an artificial DNA algorithm with compression of information by a hyperelliptic curve, plus a mechanism of security, which provides encryption, robustness and efficiency using a Hermitian curve. It should be noted that in [8, 9] the Hermitian codes were used only for channel coding or data storage. In [10], a cryptographic system is constructed using a combination of a hyperelliptic and a Reed–Solomon code; in [11], those results were extended to an LDPC (low-density parity check) code. Alternatively, in this Letter the Hermitian code is used for the S_c channel encoding and security at the same time.

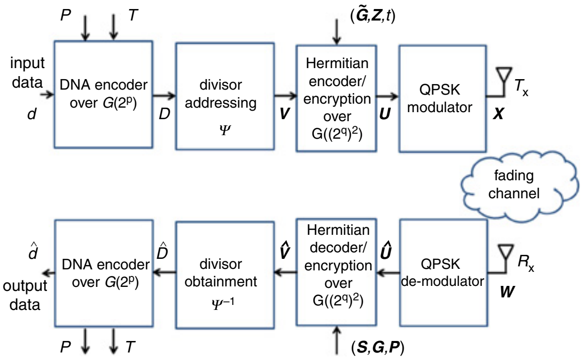

Fig. 1 *System description of artificial DNA over GF(2^p) and GF($(2^q)^2$)*

System description: Fig. 1 shows the system description of an artificial DNA communications system over GF(2^p) and GF($(2^q)^2$), where Tx is the transmitter side and Rx is the receiver side; it is assumed that the key ($\tilde{\boldsymbol{G}}$, $\boldsymbol{Z}$, $\boldsymbol{t}$) is public ($\boldsymbol{Z}$ is a random error vector and t are the correcting errors). Moreover, T and P were negotiated in advance, or can be modified after the private key ($\boldsymbol{S}$, $\boldsymbol{G}$, P) is computed. The transmitted data is a binary data sequence represented by $d = d_{\delta_0-1}, \ldots, d_1d_0$ with an arbitrary length $\delta_0 > = \rho_0 \times p$, with $\rho_0 \in \mathbb{N}$.

The language symbols are encoded by the DNA codes in Table 1, whereby the DNA basis elements are represented by a couple of bits, but using polynomial notation adenine (A) = 0, thymine (T) = 1, cytosine (C) = α and guanine (G) = α^2. The sender first generates a random DNA carrier $S_c = s_{\delta_1-1}, \ldots, s_1s_0$ with an arbitrary length $\delta_1 = \rho_1 \times p$, with $\rho_1 > \rho_0$, $\rho_1 \in \mathbb{N}$. If there is a desire to transmit the short word 'for', by the ASCII code it would be 24 bits, but by Table 1 it is reduced to 18 bits. Assume that the carrier S_c has $\delta_1 = 5 * 5 = 25$ bits, 18 bits of which are data d, i.e. $\delta_0 = 18$, and the remaining 7 bits are T_1 and P_1, but concatenated as T_1 and P_1 and so on for longer word sequences.

Table 1: DNA encoder

Symbols	Code	Symbols	Code	Symbols	Code
A	$0\ \alpha^2\ \alpha$	K	$0\ \alpha\ 1$	U	$0\ \alpha^2\ \alpha^2$
B	$\alpha^2\ \alpha\ 1$	L	$\alpha^2\ \alpha^2\ \alpha$	W	$\alpha\ \alpha^2\ 1$
C	$\alpha\ 1\ 0$	M	$\alpha\ \alpha^2\ \alpha$	V	$\alpha^2\ 1\ \alpha$
D	$1\ 0\ \alpha^2$	N	$1\ 0\ 1$	X	$1\alpha^2\ 0$
E	$\alpha^2\ 0\ \alpha$	O	$\alpha^2\ 0\ 1$	Y	$1\ \alpha^2\ \alpha^2$
F	$0\ 1\ 1$	P	$0\ 0\ \alpha^2$	Z	$\alpha^2\ \alpha\ \alpha^2$
G	$\alpha^2\ \alpha\ \alpha$	Q	$\alpha^2\ 1\ 1$		$\alpha^2\ \alpha\ \alpha$
H	$\alpha\ 0\ 1$	R	$\alpha\ 0\ \alpha^2$	.	$\alpha\ \alpha\ 0$
I	$1\ 0\ \alpha$	S	$1\ 1\ \alpha$	—	—
J	$\alpha^2\ \alpha^2\ 0$	T	$\alpha\ 0\ 0$	—	—

We will use the hyperelliptic curve HyC:$v^2 + (u^2 + u)v = u^5 + u^3 + 1$ and $g = 2$ over GF(2^5) [10, 11]. From the Hasse–Weil theorem [12], HyC has $K = \left(\sqrt{2^p} + 1\right)^{2g} = \left(4\sqrt{2} + 1\right)^4 \simeq 1964$ reduced divisors. For HyC, a reduced divisor is $D_{165} = (\alpha^0u^2 + \alpha^{17}u + \alpha^{30},\ \alpha^{23}u + \alpha^{12})$ which has the cursor number $\Psi(D_{165}) = 165$. All of the coefficients of the polynomial representation $a(u) = \alpha^0u^2 + \alpha^{17}u + \alpha^{30}$ and $b(u) = \alpha^{23}u + \alpha^{12}$ belong to the Galois field GF(2^5). Then, the following carrier array S_c = 00001 10011 10010 01111 01110 is embedded in the coefficients of the polynomial representation for D_{165}, with δ_1 = 25 bits and δ_0 = 18 bits. From left to right, the first block is the coefficient of u^2, whereas the second block is the coefficient of u, and so on. The cursor number for D_{165} is represented as a binary with $k = \lceil \log_2 K \rceil = \lceil 10.94 \rceil \simeq 11$ bits. In this case, a binary array with k = 10 bits is used for the cursor number. Therefore, 165 = 00 10 10 01 01, which represents the message $\boldsymbol{V} = (0\ \alpha\ \alpha\ 1\ 1)$ in the field GF(2^2).

In the Hermitian encoder/encryption processes block, the code generator matrix is given by

$$\boldsymbol{G}_{5,8} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & \alpha & \alpha & \alpha^2 & \alpha^2 \\ 0 & 1 & \alpha & \alpha^2 & \alpha & \alpha^2 & \alpha & \alpha^2 \\ 0 & 0 & 1 & 1 & \alpha^2 & \alpha^2 & \alpha & \alpha \\ 0 & 0 & \alpha & \alpha^2 & \alpha^2 & 1 & 1 & \alpha \end{pmatrix}$$

The columns of $\boldsymbol{G}_{5,8}$ are obtained by evaluating all monomial in Ho = $\{1,\ x,\ y,\ x^2,\ xy\}$, at rational points of the curve HeC:$y + y^2 = x^3$ over GF(2^2) [8]. The process continues with the construction of the 5 × 5 scrambler matrix $\boldsymbol{S}_5$ and the 8 × 8 permutation matrix $\boldsymbol{P}_8$, which are generated randomly. Thus, we can calculate the public key as follows:

$$\tilde{\boldsymbol{G}}_{5,8} = \boldsymbol{S}_5 \cdot \boldsymbol{G}_{5,8} \cdot \boldsymbol{P}_8$$

$$= \begin{pmatrix} \alpha^2 & 1 & 0 & 0 & 0 \\ 0 & \alpha^2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \alpha \end{pmatrix} \cdot \boldsymbol{G}_{5,8} \cdot \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & \alpha^2 & 0 & \alpha & \alpha^2 & 0 & \alpha \\ 1 & 1 & 0 & \alpha & \alpha^2 & 0 & \alpha & \alpha^2 \\ 0 & 1 & 1 & 0 & \alpha^2 & 0 & 1 & \alpha \\ \alpha^2 & \alpha^2 & 0 & \alpha & 1 & 0 & \alpha & 1 \\ \alpha & 1 & 0 & \alpha & \alpha^2 & 0 & \alpha^2 & 1 \end{pmatrix}$$

Then, the message $\boldsymbol{V} = (0\ \alpha\ \alpha\ 1\ 1)$ is encrypted by the random error vector $\boldsymbol{Z} = \begin{pmatrix} \alpha^2 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha \end{pmatrix}$ as

$$\begin{aligned} \boldsymbol{U} = \boldsymbol{V} \cdot \tilde{\boldsymbol{G}}_{5,8} + \boldsymbol{Z} &= \begin{pmatrix} \alpha^2 & \alpha & \alpha & \alpha^2 & \alpha & 0 & 0 & \alpha \end{pmatrix} + \boldsymbol{Z} \\ &= \begin{pmatrix} 0 & \alpha & \alpha & \alpha^2 & \alpha & 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

Then, $\boldsymbol{U}$ is represented as the binary array $\boldsymbol{U} = 00\ 10\ 10\ 11\ 10\ 00\ 00\ 00$. Finally, the binary array $S_c = 00001\ 10011\ 10010\ 01111\ 01110$ with $\delta_1 = 25$ bits has been compressed to the binary array $\boldsymbol{U} = u_{\delta_2} - 1, \ldots, u_1 u_0 = 00\ 10\ 10\ 11\ 10\ 00\ 00\ 00$ with $\delta_2 >= \rho_2 \times p$, with $\rho_2 \in \mathbb{N}$ and $\delta_2 = 16$ bits.

The divisor addressing block generates a cursor $\boldsymbol{V}$ using a map Ψ over the set of all reduced divisors that belong to HyC. The Hermitian encoder/encryption block generates the codeword $\boldsymbol{U}$ from the keys and the cursor $\boldsymbol{V}$ using the Hermitian curve HeC over GF($(2^q)^2$). The modulated codeword X is sent and is corrupted by the wireless channel and transformed into W using quadrature phase shift keying (QPSK) modulation.

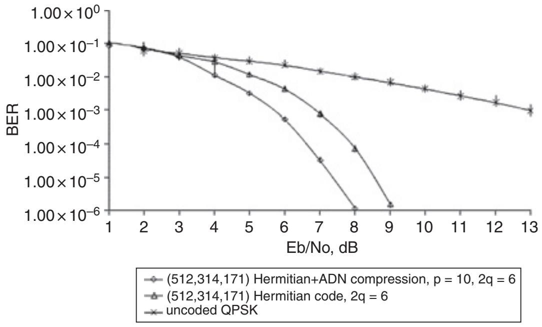

Fig. 2 *Performance of artificial DNA communication system against McEliece and uncoded QPSK over Rayleigh fading channel*

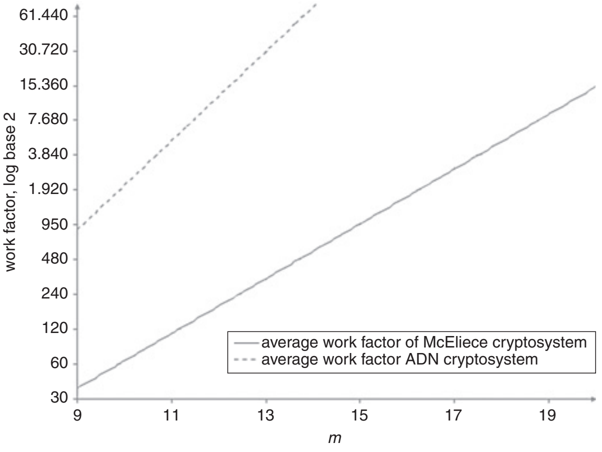

Fig. 3 *Comparison of attacks work factor (log$_2$) for DNA cryptosystem against McEliece cryptosystem*

On the receiver side Rx, the estimated transmitted codeword $\hat{U}$ is obtained after demodulation, and feeds in the Hermitian decoder/decryption block. The private key is used to retrieve the estimated cursor $\hat{V}$. In the divisor obtainment block, the inverse of Ψ, denoted by Ψ^{-1}, is applied to estimate the reduced divisors $\hat{D}$ between $T_{i_1} \& P_{i_1}$ and $T_{i_2} \& P_{i_2}$ with $i_1,\ i_2 \in \mathbb{N}$. Finally, the DNA-decoder block removes P and T and decodes the estimated binary data sequence $\hat{d} = \hat{d}_{\delta_0 - 1}, \ldots, \hat{d}_1 \hat{d}_0$ using Table 1, with $\delta_0 = 18$ bits and $\rho_1 > \rho_0 > \rho_2$.

Simulation results: Fig. 2 shows the artificial DNA packing performance against McEliece and uncoded QPSK over the Rayleigh fading channel. The code used is (512, 314, 17) from [8], the concatenation of an artificial DNA, with a hyperelliptic and a Hermitian curve with $p = 10$ and $q = 3$, generates a gain over a Rayleigh fading channel of 7 dB compared with the uncoded QPSK systems GF($(2^q)^2$) with $q = 3$, for a BER = 1×10^{-3}. Thus shows that for each δ_0 bits at the input, the system saves $\beta(\delta_0 - \delta_2)$ bits in the communication channel, with $\beta > 0$. Hence only δ_2 bits are transmitted to recover $\hat{d} = \hat{d}_{\delta_0 - 1}, \ldots, \hat{d}_1 \hat{d}_0$.

Fig. 3 shows a comparison of attacks work factor (log$_2$) for the McEliece algorithm over GF(2^m) [3] against the proposed DNA system, which shows the increase of the work factor on the proposed DNA system when compared with the attacks on the traditional McEliece algorithm. The proposed DNA system outperforms the security level McEliece algorithm by 1000% for a field exponent of $m = 13$.

Conclusion: In this Letter, we have presented an algorithm that uses the combination of artificial DNA packing and two algebraic curves, which can be used in future wireless communication systems. Moreover, the concatenation of an artificial DNA, for a BER = 1×10^{-3} a gain of about 1 dB is achieved by the code in the proposed DNA communication system compared with the McEliece communication system with GF($(2^q)^2$) and $q = 3$, since S_c has no gaps, which means more information is transmitted in a shorter time and with spectral efficiency. The security of the new security scheme is NP (non-deterministic polynomial time) complete. It is based on the minimum distance problem of the combination of two algebraic curves.

Acknowledgment: The authors acknowledge the financial support of the Center for Multidisciplinary Research on Signal Processing (Project Conicyt/ACT1120), VRIDT/UCN and Project USACH/Dicyt number 061413SG.

Submitted: *28 May 2015* E-first: *19 August 2015*
doi: 10.1049/el.2015.1226
One or more of the Figures in this Letter are available in colour online.

I. Soto (*Electrical Engineering Department, USACH, Avd. Ecuador 3519, Estacion Central, Santiago, Chile*)

✉ E-mail: ismael.soto@usach.cl

I. Jiron (*Mathematics Department, Universidad Catolica del Norte, Antofagasta, Chile*)

A. Valencia (*Department of Mechanical Engineering, Universidad de Chile, Santiago, Chile*)

R. Carrasco (*Electrical and Electronic Engineering, Newcastle University, Newcastle upon Tyne, United Kingdom*)

References

1 Hwang, T., Yang, C., Wu, G., Li, S., and Li, G.Y.: 'OFDM and its wireless applications: a survey', *IEEE Trans. Veh. Technol.*, 2009, **58**, (4), pp. 1673–1694

2 Vardy, A.: 'The intractability of computing the minimum distance of a code', *IEEE Trans. Inf. Theory*, 1997, **43**, (6), pp. 1757–1766

3 Hamdaoui, Y, and Sendrier, N.: '(2013) A non asymptotic analysis of information set decoding. IACR cryptology ePrint archive. Available at http://www.eprint.iacr.org/2013/162

4 Atito, A., KhalifaS, A., and Rida, Z.: 'DNA-based data encryption and hiding using playfair and insertion techniques', *J. Commun. Comput. Eng.*, 2011, **2**, (3), pp. 44–49

5 Bashier, E., Ahmed, G., Othman, H.-E., and Shappo, R.: 'Hiding secret messages using artificial DNA sequences generated by integer chaotic maps', *Int. J. Comput. Appl.*, 2013, **70**, (15), pp. 1–5

6 Chevallier-Mames, B., Ciet, M., and Joye, M.: 'Low-cost solutions for preventing simple side-channel analysis: side-channel atomicity', *IEEE Trans. Comput.*, 2004, **53**, (6), pp. 760–768

7 Barreto, G.de.A., and Araujo, A.F.R.: 'A self-organizing NARX network and its application to prediction of chaotic time series'. Int. Joint Conf. on Neural Networks, Washington, DC, USA, July, 2001, Proc. IJCNN '01, 2001, vol. 3, pp. 2144–2149

8 Carrasco, R., and Johnston, M.: 'Non-binary error control coding for wireless communication and data storage' (John Wiley & Sons, Chichester, UK, 2008)

9 Chen, L., Johnston, M., and Tian, G.Y.: 'Iterative detection-decoding of interleaved Hermitian codes for high density storage devices', *IEEE Trans. Commun.*, 2014, **62**, (10), pp. 3401–3409

10 Jiron, I., Soto, I., Carrasco, R., and Becerra, N.: 'Hyperelliptic curves encryption combined with block codes for Gaussian channel', *Int. J. Commun. Syst.*, 2006, **19**, pp. 809–830

11 Valencia, C., Soto, I., and Carrasco, R.: 'Secure data compression with sphere packing', *Electron. Lett.*, 2007, **43**, (23), p. 8

12 Koblitz, N.: 'Algebraic aspect of cryptography', *Algorithms Comput. Math.*, 1998, **3**, pp. 127–128, ISBN 3-540-63446-0