Exploiting Linear Hull in Matsui's Algorithm 1 (extended version)*,**

Andrea Röck and Kaisa Nyberg

Aalto University School of Science,
Department of Information and Computer Science
P.O. Box 15400, FI-00076 Aalto, Finland
fistname.lastname(replace ö by o)@tkk.fi

Abstract. We consider linear approximations of an iterated block cipher in the presence of several strong linear approximation trails. The effect of such trails in Matsui's Algorithm 2, also called the linear hull effect, has been previously studied by a number of authors. However, the effect on Matsui's Algorithm 1 has not been investigated until now. In this paper, we fill this gap and examine how to exploit the linear hull in Matsui's Algorithm 1. We develop the mathematical framework for this kind of attacks. The complexity of the attack increases with the number of strong linear trails. We show how to reduce the number of trails and thus the complexity using related keys. Further, we illustrate our theory by experimental results on a reduced round version of the block cipher PRESENT.

Keywords: block cipher, linear cryptanalysis, linear hull, key recovery, Matsui's Algorithm 1

1 Introduction

Linear cryptanalysis of an iterated block cipher as originally presented by M. Matsui in [10] is based on strong correlations between a linear combination of plaintext bits and a linear combination of ciphertext bits. Matsui also showed that given a sufficient amount of data such correlations can be observed from the data and gave the relationship between the strength of the correlation and the data requirement. Thus estimation of the data complexity is reduced to the problem of estimating the correlation of the linear approximation. Matsui's solution was to identify a strong linear approximation trail by chaining approximations from round to round over the cipher and computing the total correlation as a product of the round correlations based on what he called the Piling-up lemma. According to this method only the sign, but not the magnitude of the correlation, depends on the secret key. Matsui presented a cryptanalysis method, called Algorithm 1, which by observing the sign of the correlation allows determining one bit of the secret key of the block cipher DES. The data complexity of this attack is determined by the magnitude of the correlation, which is the same for all keys.

Daemen et al. [6] noted that, for fixed input and output bit linear combinations, there may exist several approximation trails which give non-negligible correlations as calculated using the Piling-up lemma. Moreover, all such trail correlations contribute to the magnitude of the total correlation in a manner which depends on the secret key. For such ciphers, Matsui's algorithms do not work as expected. After the invention of linear cryptanalysis, the design principles of block ciphers include criteria such as the Wide Trail Strategy [7] to split linear approximations into several small approximation trails. Typical examples of block ciphers designed to be immune against Matsui's Algorithm 1 are AES [7] and PRESENT [4].

The set of linear trails contributing to the total correlation of a linear approximation was called the linear hull in [12], where it was also shown how to calculate the average value of the squared total correlation over the keys using the linear hull. This value gives a good estimate of the data complexity of a linear distinguisher for a large proportion of the keys, while, as noted recently also by S. Murphy [11], there may exist keys, which give total correlations with negligible magnitude and thus distinguishing

* The research described in this paper has been funded by the Academy of Finland under project 122736 and was partly supported by the European Commission through the ICT program under contract ICT-2007-216676 ECRYPT II.

** This is an extended version of the paper presented at WCC 2011.

attacks and Algorithm 2 are not effective The impact of linear hulls for Algorithm 1 type of cryptanalysis was briefly addressed in [14], but has remained unexplored so far. The goal of this paper is to fill this gap in the theory of block ciphers and investigate under which circumstances it is possible to determine information about the secret key by observing the value of the correlation of a linear approximation from the cipher data.

The first assumption we make is that the total correlation of a linear approximation is essentially determined by a number of about equally strong approximation trails. We develop a mathematical framework for the statistical analysis of the varying correlation values for key alternating block ciphers with linear key schedule. The number of bits of information of the secret key obtained in this manner is logarithmic to the number of trails. Subsequently we will show that we can reduce the number of active trails using a related key attack, which can lead to a reduced complexity. By using several related keys, we are able to increase the amount of secret key information that we learn. The data requirement of these attacks will be inversely proportional to the least correlation of the approximation trails that determine the total correlation. A suitable test bed of the new cryptanalysis method developed in this paper is provided by a reduced seven-round version of the block cipher PRESENT. The correlations of its linear approximations are determined by a number of equally strong trails, while the contribution of the remaining trails is negligible.

Finally, let us note that the new attack frameworks presented in this paper exploit a single linear approximation and are therefore essentially different from the multidimensional linear attacks and other attacks that exploit multiple linear approximations simultaneously [3], [8].

The rest of the paper is structured as follows: First, we introduce linear hulls in Section 2, and show the transition from trail-correlations to key-mask correlations in Section 3. In Section 4 we describe a direct way of exploiting the information from the linear hull. Subsequently, we show in Section 5 how we can refine the attack by using a related key approach with an arbitrary difference. We illustrate in Section 6 how we can exploit specific differences to learn on the average significantly more bits of information and in Section 7 we give a summary of the attack complexities. Finally in Section 8, we give some empirical results on a seven-round version of the block cipher PRESENT [4].

2 Linear Hull

Let $\mathcal{E}_K(x)$ denote the block cipher encryption of plaintext $x \in \mathbb{Z}_2^n$ with key $K \in \mathbb{Z}_2^\ell$. A linear approximation of a block cipher with mask $(u, v, w) \in \mathbb{Z}_2^{2n+\ell}$ is a Boolean function defined as

$$(x, K) \mapsto u \cdot x \oplus v \cdot K \oplus w \cdot \mathcal{E}_K(x) \ . \tag{1}$$

The most difficult task in linear cryptanalysis is finding linear approximations with correlation of large absolute value, and in particular, determining an adequate estimate of the correlation. Let us now assume that the block cipher is a key-alternating iterated block cipher with round function $G(x, K_i) = g(x \oplus K_i)$, where x is the data input and K_i is the key input to the round. With a fixed key K the iterated block cipher is a composition of a number, say R, of round functions.

Definition 1. *For a binary random variable X on $\{0, 1\}$ the* correlation *is defined as*

$$c(X) = 2\Pr(X = 0) - 1 \ .$$

For any Boolean function $f : \mathbb{Z}_2^n \to \{0, 1\}$ we can then define the correlation $c(f(x))$ *as*

$$c(f(x)) = 2^{-n}\Big(\#\{x \in \mathbb{Z}_2^n : f(x) = 0\} - \#\{x \in \mathbb{Z}_2^n : f(x) = 1\}\Big) \ .$$

Then the correlation $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x))$ over a key-alternating block cipher can be calculated by the following theorem:

Theorem 1. *([7], [13]) Let g be the round function of an R-round key-alternating iterated block cipher $\mathcal{E}_K$ with round keys $(K_1, K_2, \ldots, K_R)$. Then for any $u \in \mathbb{Z}_2^n$ and $w \in \mathbb{Z}_2^n$ it holds that*

$$c(u \cdot x \oplus w \cdot \mathcal{E}_K(x)) = \sum_{\substack{u_2, \ldots, u_R \\ u_1 = u,\, u_{R+1} = w}} (-1)^{u_1 \cdot K_1 \oplus \cdots \oplus u_R \cdot K_R} \prod_{i=1}^{R} c(u_i \cdot x \oplus u_{i+1} \cdot g(x)) \ . \tag{2}$$

The sequences $u_1 = u, u_2, \ldots, u_R, u_{R+1} = w$, over which the summation is taken, are called (*linear approximation*) *trails* from u to w and the product $(-1)^{u_1 \cdot K_1 \oplus \cdots \oplus u_R \cdot K_R} \prod_{i=1}^{R} c(u_i \cdot x \oplus u_{i+1} \cdot g(x))$ is called the *trail-correlation* of the trail $(u_1, \ldots, u_{R+1})$. The goal of classical linear cryptanalysis, as first proposed by Matsui [10], is to find masks u and w such that for almost all keys K this correlation is large in absolute value. Matsui's Algorithm 1 seeks to determine the bit $v \cdot K$ of information of the key K based on the sign of the observed correlation $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x))$. This will succeed under two conditions. First, the observed correlation $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x))$ for the fixed unknown key K must be large, and secondly a good theoretical estimate of the sign of the correlation $c(u \cdot x \oplus v \cdot K \oplus w \cdot \mathcal{E}_K(x))$ must be available. These conditions are satisfied, if the correlation is large in absolute value and the sum on the right hand side of the Equation (2) is dominated by a single term with $v = (u_1, u_2, \ldots, u_R)$. This is the classical setting for performing Matsui's Algorithm 1. Known examples of ciphers admitting single dominant correlation trails are DES and SERPENT [2]. An extreme example of the opposite case is the block cipher PRESENT [4], which due to its regular permutation layer splits all correlations to a large number of terms without a single dominant trail.

To illustrate such a behaviour let us consider a small example presented in [7], see also [14]. In this example, the correlation (2) is assumed to take the form $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x)) = (-1)^{\gamma \cdot K} c_\gamma + (-1)^{\lambda \cdot K} c_\lambda$ where c_γ and c_λ are the correlations of the linear trails γ and λ, and $c_\gamma \approx c_\lambda$. Assume we aim at determining the value of $(-1)^{\lambda \cdot K}$ using $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x))$ as an estimate of the trail correlation $(-1)^{\lambda \cdot K} c_\lambda$. When observing the correlation $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x))$ from the data, three values are possible. They are $-c_\lambda - c_\gamma$ or 0 or $c_\lambda + c_\gamma$ depending on the key K. In the first and the third case $(-1)^{\lambda \cdot K}$ will be correctly determined, while for about half of the keys we observe correlation $0 \approx c_\lambda - c_\gamma \approx c_\gamma - c_\lambda$ which does not give any useful information for the classical Algorithm 1.

Taking another look at this example reveals that given the trail correlations c_λ and c_γ and observing the value of the correlation $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x))$ from the data, we can extract quite a lot of information of the key. Indeed for half of the keys we get two bits $\lambda \cdot K$ and $\gamma \cdot K$ of information and for the other half of the keys, we get the information that $(\lambda \oplus \gamma) \cdot K = 1$. Thus, contrary to the classical linear cryptanalysis, also correlations equal to zero are meaningful.

As the number of trails grows, the more values the correlations may take and the distinct values of the correlation (2) split the set of keys into mutually disjoint sets. Thus if sufficiently separated, the distinct values of the correlations, and hence key classes, may be identified from the data. In this paper we present a new type of linear cryptanalysis attack based on this observation.

3 From Trails to Key-Masks

Equation (2) sums over all possible trails and involves several linear combinations of round key bits. In this section we transform it to an expression that is technically easier to handle.

We start by reducing the number of terms in (2) by including only the trails whose correlations are above a certain threshold τ.

Assumption 1. The influence of trails with trail-correlation essentially smaller than τ is negligible.

Thus for fixed input and output masks u, w, we can define the set of *strong trails*:

$$\mathcal{T} = \left\{ (u_1, \ldots, u_{R+1}) : u_1 = u, u_{R+1} = w, \left| \prod_{i=1}^{R} c(u_i \cdot x \oplus u_{i+1} \cdot g(x)) \right| \geq \tau \right\} \ .$$

By Assumption 1 it suffices to take the sum in (2) over the set $\mathcal{T}$.

Next we define the key K which will be the target of our attack. Let $K_{\mathrm{M}} \in \mathbb{Z}_2^k$ be the original master key, from which all the round keys are derived. If the key schedule is linear we set $K = K_{\mathrm{M}}$. In the case where the key schedule is non-linear we start with $K = K_{\mathrm{M}}$ and add to K a new bit for each round key bit that depends in a non-linear way from K_{M} and is not yet in K. In the end we have a key $K \in \mathbb{Z}_2^\ell$, for some positive integer ℓ, and a linear relation between K and $K_1, \ldots, K_R$. In the example considered in Section 8 the length of K_{M} is 80 and $\ell = 104$.

Due to the linear relation between K and the round keys $(K_1, \ldots, K_R)$, there exists a linear function f which maps the round-masks $u_1, \ldots, u_R$ to a single mask for K, i.e. for all keys

$$f(u_1, \ldots, u_R) \cdot K = u_1 \cdot K_1 \oplus \cdots \oplus u_R \cdot K_R \ .$$

This allows us to combine all the strong trails which map to the same value $f(u_1, \ldots, u_R) = v$. We call the vectors v the *key masks* and define, for all $v \in \mathbb{Z}_2^\ell$, the *key-mask correlation* as

$$\rho(v) = \sum_{\substack{(u_1, \ldots, u_{R+1}) \in \mathcal{T} \\ f(u_1, \ldots, u_R) = v}} \prod_{i=1}^{R} c(u_i \cdot x \oplus u_{i+1} \cdot g(x)) \ .$$

Remark 1. Depending on the size of $\mathcal{T}$ and the round key schedule the number of terms in this sum varies. Typically for practical ciphers, however, each v originates from a single strong trail.

We define by $\mathcal{V} = \{v \in \mathbb{Z}_2^\ell : |\rho(v)| > 0\}$ the set of strong key masks. Using Assumption 1, we can now approximate the correlation $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x))$ using the sum

$$c(u \cdot x \oplus w \cdot \mathcal{E}_K(x)) = \sum_{v \in \mathcal{V}} (-1)^{v \cdot K} \rho(v) \ . \tag{3}$$

The set $\mathcal{V}$ can be represented as a $|\mathcal{V}| \times \ell$ matrix, where each vector $v \in \mathcal{V} \subset \mathbb{Z}_2^\ell$ is represented as a row of the matrix. We will denote this matrix by V. In our analysis, we will distinguish between two cases: independent and dependent. In the *independent case*, all vectors in $\mathcal{V}$ are independent, where as in the the *dependent case*, the vectors in $\mathcal{V}$ are not all independent, which means $|\mathcal{V}| > \operatorname{rank}(V)$. Thus, in the dependent case we have to take into account the linear dependencies between different trails.

4 Direct Attack

As discussed in Section 2, the correlation $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x))$ of a linear approximation with input/output masks u and w depends on the key K. In this section, we develop a statistical method to obtain information about the key using this fact.

Let $\mathcal{C} = \left\{c(u \cdot x \oplus w \cdot \mathcal{E}_K(x)) : K \in \mathbb{Z}_2^\ell\right\}$ be the set of possible outcomes of the correlation for masks u and w. We will denote by $d = \min_{c_1 \neq c_2 \in \mathcal{C}} |c_1 - c_2|$ the minimal distance between two elements of $\mathcal{C}$. This value affects the complexity of the attack. As a last value we define the constant $\tilde{c} = 2^{-n} \gcd_{v \in \mathcal{V}} (2^n \rho(v))$. From Definition 1 we know that $2^n \rho(v)$ is always an integer value. Then all strong key-mask correlations can be written as integer multiples of $\tilde{c}$. Let $\mathcal{I} \subset \mathbb{Z}$ denote the set of all integer multipliers of $\tilde{c}$ such that $\mathcal{C} = \{i\tilde{c}\}_{i \in \mathcal{I}}$. Then by (3) it must hold that $d \geq 2\tilde{c}$. The variable $\tilde{c}$ makes the notation easier, however, the important value is d.

Remark 2. In the case of PRESENT, the strong key-mask correlations are always $\pm\tilde{c}$.

4.1 Statistical Test

We divide the set of all keys into $|\mathcal{C}|$ disjoint subclasses $\mathcal{K}(c) = \left\{K \in \mathbb{Z}_2^\ell : c(u \cdot x \oplus w \cdot \mathcal{E}_K(x)) = c\right\}$ for all $c \in \mathcal{C}$. Then we use the basic m-ary hypothesis testing problem with $m = |\mathcal{C}|$ to determine the value c and, consequently, the key-class $\mathcal{K}(c)$.

Notation. We denote random variables by capital letters $X, Y, \ldots$ and their realizations $x \in \mathcal{X}, y \in \mathcal{Y}, \ldots$ by lower case letters. A sequence of independent and identical distributed (i.i.d.) random variables is denoted by a bold letter, e.g. $\mathbf{X} = X_1, X_2, \ldots, X_N$, where N is the length of the sequence. The discrete probability distribution of a random variable is denoted by $p = (p_\eta)_{\eta \in \mathcal{X}}$. For a given sequence $\mathbf{x}$, let $N(\eta|\mathbf{x}) = \#\{i : x_i = \eta\}$ denote the empirical frequency of η in the sequence, with $\eta \in \mathcal{X}$. If the sequence is clear from the context we will use sometimes $N_\eta = N(\eta|\mathbf{x})$. Then the empirical distribution q is given by $q_\eta = N_\eta / N$ for $\eta \in \mathcal{X}$.

The result of the linear approximation $(u \cdot x \oplus w \cdot \mathcal{E}(x))$ is either zero or one, thus we have $\mathcal{X} = \{0, 1\}$ and $p = (p_0, p_1)$. For our statistical analysis we want to solve a basic m-ary hypothesis testing problem, where we want to decide which of the $m = |\mathcal{C}| = |\mathcal{I}|$ different hypotheses H_i with $i \in \mathcal{I}$ is true. Hypothesis H_i states that the i.i.d. random variables X_j, $1 \leq j \leq N$, have correlation $i\tilde{c}$, thus $p_0^i = \frac{1}{2}(1 + i\tilde{c})$ and $p_1^i = \frac{1}{2}(1 - i\tilde{c})$. The a priori distribution of H_i is given by $\pi^i = \Pr(H_i) = 2^{-\ell} |\mathcal{K}(i\tilde{c})|$. We use the decision function $\delta : \mathcal{X}^N \to \mathcal{I}$ to solve the testing problem. This function assigns each sequence $\mathbf{x}$ to a hypothesis. We denote $P_{ij} = \Pr(\delta(\mathbf{X}) = i | H_j)$ the error probability of δ choosing H_i if H_j is true.

Optimal Test Statistic. We use a Bayesian approach, with $m = |\mathcal{C}|$ different hypotheses. For a given hypothesis H_i, the sequence $\mathbf{X}$ is binomial distributed, i.e.

$$\Pr\left(N(0|\mathbf{X}) = N_0 \middle| H_i\right) = \binom{N}{N_0} (p_0^i)^{N_0} (p_1^i)^{N_1} \quad .$$

Let $q = (q_0, q_1)$ be the empirical distribution, with $q_\eta = \frac{N(\eta|\mathbf{x})}{N}$, $\eta = 0, 1$. The previous equation only depends on q and is defined as the *likelihood function*

$$\mathcal{L}(i; q) = \binom{N}{Nq_0} (p_0^i)^{Nq_0} (p_1^i)^{Nq_1} \quad .$$

Then for a given sequence $\mathbf{x}$, the optimal decision function outputs i for which the probability of the sequence is maximal, i.e. which maximizes

$$\Pr\left(H_i \middle| N(0|\mathbf{x})\right) = \frac{\pi^i \Pr\left(N(0|\mathbf{x}) \middle| H_i\right)}{\Pr\left(N(0|\mathbf{x})\right)} \quad . \tag{4}$$

Which i maximizes (4), depends only on the empirical probability q. Thus for a given q the *optimal decision function* searches for the i which maximizes $\pi^i \mathcal{L}(i; q)$. By taking the logarithm and ignoring factors that are the same for all i's we get the following result:

Lemma 1. *[9] The optimal decision function is given by:*

$$\delta(\mathbf{x}) = \arg\max_{i \in \mathcal{I}} \left[\log_2(\pi^i) + N_0 \log_2(p_0^i) + N_1 \log_2(p_1^i)\right] \quad , \tag{5}$$

and leads to a total error probability of

$$P_e = \sum_{i \in \mathcal{I}} \pi^i \sum_{j \in \mathcal{I}, j \neq i} P_{ij} \quad . \tag{6}$$

Complexity. In this section we study the data complexity N that we need for a fixed error probability P_e. In the analysis we use the following two assumptions.

Assumption 2. The distributions of p^i and p^j are *close*, i.e. for $i, j \in \mathcal{I}$ there exists an $0 < \varepsilon < 1/2$ such that $|p_\eta^i - p_\eta^j| \leq \varepsilon p_\eta^j$ for all $\eta \in \{0, 1\}$.

Assumption 3. For all $i \in \mathcal{I}$, $1 \gg |i\tilde{c}|$. From this follows that $\tilde{c}^{-1} \gg |i|$ and we can approximate $\tilde{c}^{-2} + i^2$ by $\tilde{c}^{-2}$, for $i \in \mathcal{I}$.

Both assumption are true for all practical cases.

Lemma 2. *For the optimal decision function described in Lemma 1 and a fixed error probability P_e, the data complexity is upper bounded proportional to*

$$N = 8 \ln(2) \frac{\log_2(|\mathcal{C}| - 1) - \log_2 P_e}{d^2}. \tag{7}$$

Lemma 2 shows that the data complexity is proportional to d^{-2}, where d is the minimal difference between two correlations in $\mathcal{C}$, but only logarithmic in $|\mathcal{C}|$.

Proof. We start by considering a binary decision between two hypotheses H_i and H_j and finally deduce N for the total test.

The Chernoff theorem [5] states that the error probability P_{ij} is given by

$$P_{ij} = \mathcal{O}\left(2^{-N D^*(p^i, p^j)}\right), \tag{8}$$

independent of the a priori distributions, where

$$D^*(p^i, p^j) = -\min_{0\le\lambda\le 1} \log_2 \left(\sum_{\eta\in\mathcal{X}} (p^i_\eta)^\lambda (p^j_\eta)^{1-\lambda} \right)$$

is the *Chernoff-information* between the distributions p^i and p^j.

Baignères and Vaudenay [1] showed that if p^i is close to p^j, the Chernoff-information can be approximated by

$$D^*(p^i, p^j) \approx \frac{1}{8\ln(2)} C(p^i, p^j), \tag{9}$$

where $C(p^i, p^j) = \sum_{\eta\in\mathcal{X}} (p^i_\eta - p^j_\eta)^2 / p^j_\eta$ is the capacity between the two distributions. Due to Assumption 2, we can use the previous equation and due to Assumption 3 we can approximate the capacity by

$$C(p^i, p^j) = \frac{(i-j)^2}{\tilde{c}^{-1} - j^2} \approx (i-j)^2 \tilde{c}^2 \ .$$

Together with Equations (8) and (9) we get that for a fixed error P_{ij}, the data complexity is proportional to

$$N_{ij} = 8\ln(2) \frac{-\log_2 P_{ij}}{(i-j)^2 \tilde{c}^2} \ .$$

We now fix the pairwise error probabilities to $P_{ij} = P_e/(|\mathcal{C}|-1)$. To achieve this value we need a data complexity of

$$N = 8\ln(2) \max_{i\neq j\in\mathcal{I}} \frac{\log_2(|\mathcal{C}|-1) - \log_2 P_e}{(i-j)^2\tilde{c}^2} = 8\ln(2) \frac{\log_2(|\mathcal{C}|-1) - \log_2 P_e}{d^2_{\mathcal{C}}} \ .$$

From Equation (6) we know that the total error will be

$$\sum_{i\in\mathcal{I}} \pi^i \sum_{j\in I, j\neq i} \frac{P_e}{(|\mathcal{C}|-1)} = P_e \sum_{i\in\mathcal{I}} \pi^i = P_e \ .$$

□

Gained Information. The test will tell us in which key-class the secret key lies. A question remains: How much information do we gain by this knowledge?

For a probability distribution, the average information learned by guessing the outcome correctly is given by its Shannon entropy [15]. Thus, in our case we learn on average

$$h = -\sum_{i\in\mathcal{I}} \pi^i \log \pi^i \tag{10}$$

bits of information. In the independent case where all $|\rho(v)| = \tilde{c}$, the hypotheses are binomial distributed, i.e. $\mathcal{I} = \{-|\mathcal{V}| + 2j\}_{0\le j\le|\mathcal{V}|}$ and

$$\pi^i = \Pr(H_i) = \binom{|\mathcal{V}|}{\frac{|\mathcal{V}|+i}{2}} 2^{-|\mathcal{V}|} \ .$$

Then, the entropy, and thus the average information, is given by $\frac{1}{2}\log_2\left(\frac{\pi e}{2}|\mathcal{V}|\right) + \mathcal{O}\left(\frac{1}{|\mathcal{V}|}\right)$. From this follows that the gained information increases only logarithmically with the number of different paths.

If we have more variation in $\rho(v)$ in the independent case and always in the dependent case, we have to consider the a priori distribution for the specific set $\mathcal{V}$. In the next section we show an efficient way of finding these values without computing (3) for all $K \in \mathbb{Z}_2^\ell$. In any case, the entropy will be smaller or equal to $\log_2|\mathcal{I}| = \log_2|\mathcal{C}|$.

4.2 Efficient Computation of the Key-classes and the A Priori Probabilities

In this section we show how to compute the set $\mathcal{C}$, the different key classes and their a priori probabilities without evaluating (3) for all $K \in \mathbb{Z}_2^\ell$.

Let t be the dimension of the vector space $\mathrm{span}(\mathcal{V}) \subset \mathbb{Z}_2^\ell$. We first choose a basis $\mathcal{B} = (b_0, \ldots, b_{t-1})$ of $\mathrm{span}(\mathcal{V})$ and denote by B the $t \times \ell$ matrix containing all the basis vectors and by B^T its transpose. Then we can represent every vector v by a t-tuples $\mathbf{v} = (v_0, \ldots, v_{t-1}) \in \mathbb{Z}_2^t$ with $v = \sum_{i=0}^{t} v_i b_i = \mathbf{v}B$. In the following we will always use $\mathbf{v}$ to denote the t-bit value and v to denote the corresponding ℓ bit value in $\mathcal{V}$. Let $\boldsymbol{\mathcal{V}} = \{\mathbf{v} \in \mathbb{Z}_2^t : v \in \mathcal{V}\} \subset \mathbb{Z}_2^t$, then we can write (3) as:

$$c(u \cdot x \oplus w \cdot \mathcal{E}_K(x)) = \sum_{\mathbf{v} \in \boldsymbol{\mathcal{V}}} (-1)^{\left(\sum_{i=0}^{t-1} v_i b_i\right) \cdot K} \rho(v) = \sum_{\mathbf{v} \in \boldsymbol{\mathcal{V}}} (-1)^{\mathbf{v} \cdot (KB^T)} \rho(v) \ . \tag{11}$$

We see that the correlation depends only on the t-bit value $\mathbf{K} = (KB^T)$. Thus, to obtain $\mathcal{C}$, the key classes and the a priori probabilities it is sufficient to consider only all $\mathbf{K} \in \mathbb{Z}_2^t$ instead of all $K \in \mathbb{Z}_2^\ell$.

The direct computation of (11) for all $\mathbf{K} \in \mathbb{Z}_2^t$ can be done in $\mathcal{O}\left(|\mathcal{V}| 2^t\right)$. However, if we extend the sum in (11) to all $\mathbf{v} \in \mathbb{Z}_2^t$ and set $\rho(v) = 0$ for $\mathbf{v} \notin \boldsymbol{\mathcal{V}}$, we can use a fast Walsh-Hadamard transform, which reduces the complexity further to $\mathcal{O}\left(t 2^t\right)$. To store the key classes we need $\mathcal{O}\left(2^t\right)$ memory.

5 Related-Key Approach

In this section we show how the number of terms in (3) can be reduced using a related key attack. If such an attack can be repeated using a number of different related keys, more refined information about the key will be possible to achieve as will be shown in the next section.

For the basic related key setting we consider the correlation differences between the keys K and $K \oplus \alpha$,

$$\Delta(K, \alpha) = c(u \cdot x \oplus w \cdot \mathcal{E}_K(x)) - c(u \cdot x \oplus w \cdot \mathcal{E}_{K \oplus \alpha}(x)) = \sum_{v \in \mathcal{V}} (-1)^{v \cdot K} \rho(v) - \sum_{v \in \mathcal{V}} (-1)^{v \cdot (K \oplus \alpha)} \rho(v) \ .$$

Many terms in the sum cancel out. Thus, the idea behind the related key approach is that we can reduce the number of v over which we have to sum. We define this reduced set by $\mathcal{V}_\alpha = \{v \in \mathcal{V} : v \cdot \alpha = 1\}$. Then we have

$$\Delta(K, \alpha) = 2 \sum_{v \in \mathcal{V}_\alpha} (-1)^{v \cdot K} \rho(v) \ . \tag{12}$$

We denote $\mathcal{C}_\alpha = \left\{\Delta(K, \alpha) : K \in \mathbb{Z}_2^\ell\right\}$, the set of all possible correlation differences, and $\mathcal{I}_\alpha$ with $\mathcal{C}_\alpha = \{i\tilde{c}\}_{i \in \mathcal{I}_\alpha}$, $\mathcal{K}_\alpha(c)$ and π_α^i the corresponding index set, key classes and a priori probabilities, respectively. We define again by $d_\alpha = \min_{c_1 \neq c_2 \in \mathcal{C}_\alpha} |c_1 - c_2|$ the minimal differences between two values in $\mathcal{C}_\alpha$. Note that due to the multiplication factor 2, $d_\alpha \geq 4\tilde{c}$.

5.1 Statistical Test

This time, instead of using the binomial distribution, we approximate the outcome of a sequence $\mathbf{X}$, with correlation c, by a normal distribution, i.e.

$$\Pr\left(N(0|\mathbf{X})\right) \sim \mathcal{N}\left(\frac{N}{2}(1+c), \frac{N}{4}(1+c)(1-c)\right) = \mathcal{N}\left(\frac{N}{2}(1+c), \frac{N}{4}(1-c^2)\right) \ .$$

Let $\mathbf{X}$ be a sequence with correlation c_1 and $\mathbf{Y}$ be a sequence with correlation c_2. We assume that $N(0|\mathbf{X})$ and $N(0|\mathbf{Y})$ are independently distributed. Then their difference is distributed with

$$\Pr\left(N(0|\mathbf{X}) - N(0|\mathbf{Y})\right) \sim \mathcal{N}\left(\frac{N}{2}(c_1 - c_2), \frac{N}{4}(2 - c_1^2 - c_2^2)\right) \ .$$

Since $2 \gg c_1^2 + c_2^2$, we will approximate the variance by $N/2$.

Optimal Test Statistic. Let $\mathbf{X}$ be the sequence corresponding to $c(u \cdot x \oplus w \cdot \mathcal{E}_K(x))$ and $\mathbf{Y}$ the sequence corresponding to $c(u \cdot x \oplus w \cdot \mathcal{E}_{K \oplus \alpha}(x))$.

Assumption 4. The random variables $N(0|\mathbf{X})$ and $N(0|\mathbf{Y})$ are independently distributed.

Hypothesis H_i states that $N(0|\mathbf{x}) - N(0|\mathbf{y})$ is distributed according to $\mathcal{N}\left(\frac{N}{2}(i\tilde{c}), \frac{N}{2}\right)$. For a given outcome $\mathbf{x}, \mathbf{y}$, we use again a test statistic that outputs hypothesis H_i for which the probability

$$\Pr\left(H_i \middle| N(0|\mathbf{x}) - N(0|\mathbf{y})\right) = \frac{\pi_\alpha^i \Pr\left(N(0|\mathbf{x}) - N(0|\mathbf{y}) \middle| H_i\right)}{\Pr\left(N(0|\mathbf{x}) - N(0|\mathbf{y})\right)} \tag{13}$$

is maximized. If we take the natural logarithm of (13) and discard all parts that do not depend on H_i we get the following optimal decision function:

$$\delta_\alpha(\mathbf{x}, \mathbf{y}) = \arg\max_{i \in \mathcal{I}_\alpha} \left[\ln(\pi_\alpha^i) - \frac{\left(N(0|\mathbf{x}) - N(0|\mathbf{y}) - \frac{N}{2} i\tilde{c}\right)^2}{N} \right] . \tag{14}$$

Complexity. Similar to Section 4, we can give the following lemma.

Lemma 3. *For the optimal decision function* (14) *and a fixed error probability P_e, the data complexity is upper bounded proportional to*

$$N = 16 \ln(2) \frac{\log_2(|\mathcal{C}_\alpha| - 1) - \log_2 P_e}{d_\alpha^2}$$

Note that in comparison with Lemma 1 we have the factor $16 \ln(2)$ instead of $8 \ln(2)$. However, due to the multiplication by 2 in (12), in most of the cases we will have $d_\alpha \geq 2d$, which leads to a slightly smaller data complexity in the related key approach.

Proof. Like in the previous section we start with the decision problem between two hypotheses H_i and H_j. Hypotheses i and j state that the outcome is distributed accordingly to, respectively, $\mathcal{N}\left(\frac{N}{2}(i\tilde{c}), \frac{N}{2}\right)$ and $\mathcal{N}\left(\frac{N}{2}(j\tilde{c}), \frac{N}{2}\right)$, where we approximate the variance by $N/2$. Then the error probability P_{ij} is given by $P_{ij} = \frac{1}{\sqrt{\pi N_{ij}}} e^{-\frac{N_{ij}}{16}(i-j)^2\tilde{c}^2}$ and we have

$$N_{ij} = \left[-\frac{\ln(2)}{2} \log_2(\pi N_{ij}) - \ln(2) \log_2(P_{ij}) \right] \frac{16}{(i-j)^2\tilde{c}^2} < \frac{-16 \ln(2) \log_2(P_{ij})}{(i-j)^2\tilde{c}^2} .$$

We now fix $P_{ij} = P_e/(|\mathcal{C}_\alpha| - 1)$ and obtain

$$N = 16 \ln(2) \max_{i \neq j \in \mathcal{I}_\alpha} \frac{\log_2(|\mathcal{C}_\alpha| - 1) - \log_2 P_e}{(i-j)^2\tilde{c}^2} = 16 \ln(2) \frac{\log_2(|\mathcal{C}_\alpha| - 1) - \log_2 P_e}{d_\alpha^2} .$$

This concludes the proof by the same arguments as in the proof of Lemma 2. □

Gained Information. We can apply the same method to evaluate the gained Information as in Section 4, however we use the set $\mathcal{V}_\alpha$ instead of $\mathcal{V}$. Thus by the related key approach we obtain

$$h_\alpha = -\sum_{i \in \mathcal{I}} \pi_\alpha^i \log \pi_\alpha^i \tag{15}$$

bits of entropy.

Efficient computation. We can use the same method as in Section 4.2. Instead of $t = \dim(\mathrm{span}(\mathcal{V}))$ we can consider $t_\alpha = \dim(\mathrm{span}(\mathcal{V}_\alpha))$, which in most cases will be smaller than t and thus reduces the complexity.

6 Using Multiple Related Keys

In this section we show how to use several related keys to obtain more information about the key K. It may take a lot of offline analysis to determine the optimal selection of the related key differences to be used in the attack.

To analyze the situation let us use the same approach as in Section 4.2. This means that we have a basis $\mathcal{B}$ for span($\mathcal{V}$) and we consider only the t-bit values $\mathbf{K} = KB^T$ of the keys instead of all $K \in \mathbb{Z}_2^\ell$. Thus, we might write $\Delta(\mathbf{K}, \alpha_i)$ for (12) instead of $\Delta(K, \alpha_i)$. In the independent case we can set directly $\mathcal{B} = \mathcal{V}$.

We now choose t differences $\alpha_0, \ldots, \alpha_{t-1} \in \mathbb{Z}_2^\ell$ in such a way that they form a dual basis for $\mathcal{B}$, i.e.

$$\alpha_i \cdot b_j = \begin{cases} 1 & \text{for } i = j \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{16}$$

Since all basis vectors are independent, we can always solve this system. From (12) follows that

$$\Delta(K, \alpha_i) = 2 \sum_{v \in \mathcal{V}_{\alpha_i}} (-1)^{v \cdot K} \rho(v) \ .$$

Knowing the corresponding correlation difference $\eta_i = \Delta(K, \alpha_i)$ will give us the key class $\mathcal{K}_{\alpha_i}(\eta_i)$. If we combine the results for all $0 \le i \le t-1$ we can increase our knowledge. Let $\boldsymbol{\eta} = (\eta_0, \ldots, \eta_{t-1})$, then we know that the key must be in

$$\mathcal{K}_{\mathcal{B}}(\boldsymbol{\eta}) = \bigcap_{0 \le i \le t-1} \mathcal{K}_{\alpha_i}(\eta_i) \ .$$

Note that the set $\mathcal{K}_{\mathcal{B}}(\boldsymbol{\eta})$ depends only on the choice of the basis $\mathcal{B}$ but not on the choice of the α_i as long as they satisfy (16). The question is now, how many keys are in each $\mathcal{K}_{\mathcal{B}}(\boldsymbol{\eta})$ and how much entropy can we gain by this method. This value depends of the set $\mathcal{V}$ and the choice of $\mathcal{B}$, and can be evaluated in $\mathcal{O}\left(t^2 2^t\right)$ by computing $\Delta(\mathbf{K}, \alpha_i)$ for all $\mathbf{K} \in \mathbb{Z}_2^t$ and $0 \le i \le t-1$. We need $\mathcal{O}\left(t 2^t\right)$ memory to store the definitions of all $\mathcal{K}_{\mathcal{B}}(\boldsymbol{\eta})$. Let $\mathcal{C}_{\mathcal{B}} = \{\boldsymbol{\eta} = (\eta_0, \ldots, \eta_{t-1}) : \Delta(\mathbf{K}, \alpha_i) = \eta_i, 0 \le i \le t-1, \mathbf{K} \in \mathbb{Z}_2^t\}$. Then the probability of $\boldsymbol{\eta}$ is $p_{\boldsymbol{\eta}} = 2^{-t} |\mathcal{K}_{\mathcal{B}}(\boldsymbol{\eta})|$ and we will learn on average

$$h_{\mathcal{B}} = -\sum_{\boldsymbol{\eta} \in \mathcal{C}_{\mathcal{B}}} p_{\boldsymbol{\eta}} \log_2 p_{\boldsymbol{\eta}} \ . \tag{17}$$

Since $|\mathcal{C}_{\mathcal{B}}| \le 2^t$, we can never achieve more than t bits of entropy. However, the example in Section 8 shows that we can get close to t bits of entropy. We could even find masks for which the entropy reaches t bits.

Note that when computing the entropy $h_{\mathcal{B}}$, it is not allowed to sum over all h_{α_i} since the results from the different α_i's are not independent. In general, the correct value $h_{\mathcal{B}}$ is much smaller than $\sum_{i=0}^{t-1} h_{\alpha_i}$.

7 Complexity of the attacks

All attacks in this work can be separated in three phases. In the *precomputation* phase we compute the correlation for each key, in the *online* phase we obtain the empirical bias of the plaintext/ciphertext pairs for the secret key and in the *post-computation* phase we choose one key-class. A summary of the complexity for the different attacks is given in Table 1. The memory complexity in the multiple related key attack can be reduced to $\mathcal{O}\left(2^t\right)$ if we redo the computation of the key classes separately for each difference in the post-computation phase. However, this would increase the time complexity of the last phase to $\mathcal{O}\left(t^2 2^t\right)$.

8 Results from Experiments on Reduced Round PRESENT

The attacks presented in this paper were tested on a seven-round version of the block-cipher PRESENT [4] with an 80-bit key. PRESENT has a specific property which makes it very suitable for our purposes. This property is that for each possible input and output mask there are several strong trails, each consisting

Table 1. Complexity of the different attacks

	precomputation			online		post-computation
		time	memory		data	time
direct attack	t	$\mathcal{O}\left(t2^t\right)$	$\mathcal{O}\left(2^t\right)$	$d \geq 2\tilde{c}$	$N = 8\ln(2)\frac{\log_2(\lvert\mathcal{C}\rvert-1)-\log_2 P_e}{d^2}$	$\mathcal{O}\left(\lvert\mathcal{C}\rvert\right)$
related key	$t_\alpha \leq t$	$\mathcal{O}\left(t_\alpha 2^{t_\alpha}\right)$	$\mathcal{O}\left(2^{t_\alpha}\right)$	$d_\alpha \geq 4\tilde{c}$	$N_\alpha = 16\ln(2)\frac{\log_2(\lvert\mathcal{C}_\alpha\rvert-1)-\log_2 P_e}{d_\alpha^2}$	$\mathcal{O}\left(\lvert\mathcal{C}_\alpha\rvert\right)$
mult. rel. key	t	$\mathcal{O}\left(t^2 2^t\right)$	$\mathcal{O}\left(t2^t\right)$	$d_{\alpha_i} \geq 4\tilde{c}$	$N = \max_{0\leq i\leq t-1} N_{\alpha_i}$	$\mathcal{O}\left(\sum_{i=0}^{t-1}\lvert\mathcal{C}_{\alpha_i}\rvert\right)$

of round approximations with an absolute correlation of 2^{-2}. Thus, all strong trails over r rounds have a trail-correlation of absolute value 2^{-2r}. For seven rounds, all strong trails map to separate key-masks, thus, $\rho(v) = \pm 2^{-14}$ for all $v \in \mathcal{V}$. We can set $\tilde{c} = 2^{-14}$ and know that $d \geq 2^{-13}$ and $d_\alpha \geq 2^{-12}$.

We are using the original *non-linear* key-schedule of PRESENT. In the 80-bit key case, at every round except for the first one, 4 bits are transformed by an S-box. Thus to construct the target key K, from which all the round-keys depend in a linear way, we have to extend the original 80-bit key K_M by the bits that are the output from the S-box transformations. For 7 rounds the key K consists of $\ell = 104$ bits. In the related-key approach we must be careful not to use a difference for K which cannot be achieved by a linear difference in K_M.

We only used 1-bit masks for the input and the output, where the output mask is applied directly after the last S-box layer. Let u, w be the bit-position of, respectively, the input and the output mask, then we denote the mask pair by (u, w). The simple structure of the cipher allows to obtain the exact set $\mathcal{V}$ including $\rho(v)$ for a given mask quite fast. For example for the 1-bit mask pair $(53, 37)$ we could obtain $\mathcal{V}$ for 20 rounds in less than 14 seconds.

For our tests, we chose the masks $(53, 37)$, fixed a basis and computed $\alpha_0, \ldots, \alpha_{14}$. Our choice leads to the following values: $|\mathcal{V} = 24|$, $|\mathcal{C}| = 13$, $t = 15$, $h = 3.21$, $h_\mathcal{B} = 14.25$.

As we have seen, different approaches lead to different amounts of average learned information. To acknowledge this fact, we define the new notion of *achieved entropy* which is the entropy of a test multiplied by its success probability. Note that we achieve the full entropy $h_\mathcal{B}$ only if we determine all values η_i, $0 \leq i \leq t-1$, correctly.

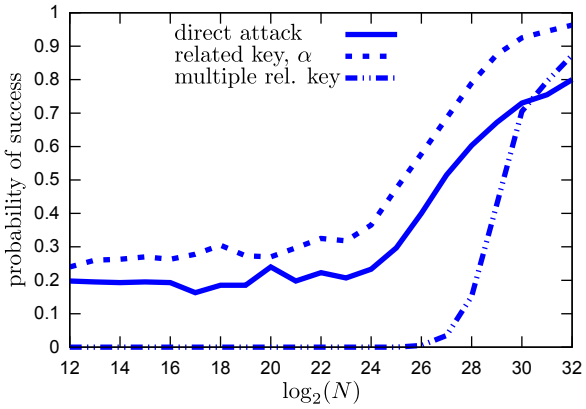
In Fig. 1 we consider three different cases: The direckt attack (Section 4), the related key attack for a single difference α (Section 5, $|\mathcal{V}_\alpha| = 9$, $|\mathcal{C}_\alpha| = 10$, $t_\alpha = 9$, $h_\alpha = 2.63$), and the multiple related key approach (Section 6). In all three cases we give the probability of success and the achieved entropy for 400 random keys and up to $N = 2^{32}$ plaintext/ciphertext pairs. Since the number of keys is not very high, the graphs still show some uneven behaviour.

We see that the success probability of the single related key attack is always larger than the one for the generic attack. This comes from the fact that $|\mathcal{C}_\alpha| < |\mathcal{C}|$, thus we have less choices and a higher probability to be correct, but also from the fact that $d_\alpha > d$. We only determine the full $\boldsymbol{\eta}$ correctly if we determine all the η_i correctly, thus the success probability of the third graph increases later, but in the end it benefits form the fact that $d_{\alpha_i} > d$.

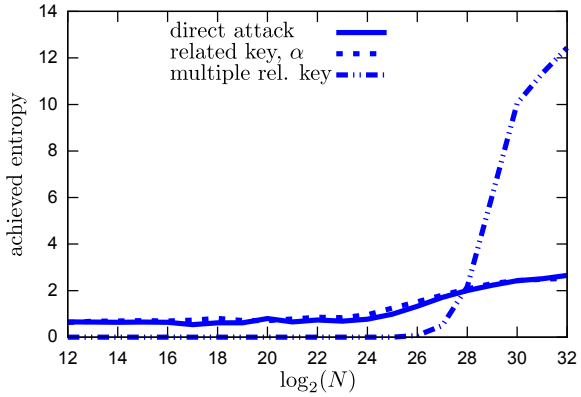
When considering the achieved entropy, we see that for some time the single related key approach leads to better results than the direct approach. For $N \geq 2^{28}$, the multiple related key approach leads to the highest achieved entropy.

9 Conclusion

In this paper, we have described some new approaches of extracting information of the observed correlation of a linear approximation. We have tested the attack algorithms on the PRESENT block cipher and seen that they work as expected. We would expect the attacks to be applicable to any iterated block cipher which has a linear key schedule and linear approximations originating from a relatively small number of about equally strong approximation trails. PRESENT has this behaviour over a small number of rounds, but as the number of rounds grows the number of trails and the rank of the trail matrix will become prohibitive. Also the data complexity determined by the trail correlation will exceed the size of the cipher's code book after about 16 rounds. The existence of a more practical example of a cipher with a suitable linear trail structure remains an open question. For the direct attack, the linear key-schedule is not necessary.



(a) Probability of success



(b) Achieved entropy

Fig. 1. Empirical results for 400 random keys

The direct attack uses the full value set of the correlation. We have also seen how to reduce this value set by inserting a difference, known to the attacker, in the secret key. This can lead to slightly smaller data complexity and to a reduced time and memory complexity. Similarly, it is possible to consider a related key fault attack by flipping one bit in a known position of the round key. If physically feasible, such an attack would work for any key alternating block cipher and give one bit of information of the round keys provided that the targeted trail correlation is sufficiently large.

We described a way how to exploit a linear hull in Matsui's Algorithm 1, which has not been analyzed until now. We showed that the data complexity is inversely proportional to the square of the smallest trail correlation. In Algorithm 2 the average data complexity is inversely proportional to the sum of the squares of the trail correlations, which makes the data complexity in general smaller than for Algorithm 1. However, our approach for Algorithm 1 works for all keys and not just for a subset and if the number of trails is small, the difference of the complexity between the two approaches is not very big. As the two algorithms target on different sets of key-bits, Algorithm 1 on the inner round-keys, Algorithm 2 on the external round-keys, Algorithm 1 is not an alternative to Algorithm 2 but typically used in addition to it, which makes these two algorithms not directly comparable.

Acknowledgements

We wish to thank the anonymous reviewers for helpful comments and Tommi Larjomaa, who was supported by the MATINE project SYMKRYPTO number 776, for his initial feasibility studies on the subject.

References

1. Baignères, T., Vaudenay, S.: The complexity of distinguishing distributions. In: ICITS 2008. pp. 210–222. LNCS, Springer, Heidelberg (2008)
2. Biham, E., Anderson, R., Knudsen, L.: Serpent: A new block cipher proposal. In: Vaudenay, S. (ed.) FSE 1998. LNCS, vol. 1372, pp. 222–238. Springer, Heidelberg (1998)
3. Biryukov, A., Cannière, C.D., Quisquater, M.: On Multiple Linear Approximations. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 1–22. Springer, Heidelberg (2004)
4. Bogdanov, A., Knudsen, L.R., Leander, G., Paar, C., Poschmann, A., Robshaw, M.J.B., Seurin, Y., Vikkelsoe, C.: PRESENT: An ultra-lightweight block cipher. In: Paillier, P., Verbauwhede, I. (eds.) CHES 2007. LNCS, vol. 4727, pp. 450–466. Springer, Heidelberg (2007)
5. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley-Interscience, New York, NY, USA (1991)
6. Daemen, J., Govaerts, R., Vandewalle, J.: Correlation matrices. In: Preneel, B. (ed.) FSE 1994. LNCS, vol. 1008, pp. 275–285. Springer, Heidelberg (1995)
7. Daemen, J., Rijmen, V.: The Design of Rijndael: AES - The Advanced Encryption Standard. Springer, Heidelberg (2002)
8. Hermelin, M., Nyberg, K.: Dependent linear approximations - the algorithm of Biryukov and others revisited. In: CT-RSA 2010. LNCS, vol. 5985, pp. 318–333. Springer, Heidelberg (2010)
9. Levy, B.C.: Principles of Signal Detection and Parameter Estimation. Springer, Heidelberg (2008)
10. Matsui, M.: Linear cryptanalysis method for DES cipher. In: Helleseth, T. (ed.) EURORYPT 1993. LNCS, vol. 765, pp. 386–397. Springer, Heidelberg (1994)
11. Murphy, S.: The effectiveness of the linear hull effect. Report RHUL-MA-2009-19, Departmental Technical Report (2009)
12. Nyberg, K.: Linear approximation of block ciphers. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 439–444. Springer, Heidelberg (1995)
13. Nyberg, K.: Correlation theorems in cryptanalysis. Discrete Applied Mathematics 111(1-2), 177–188 (2001)
14. Nyberg, K.: Linear cryptanalysis using multiple linear approximations. Early Symmetric Crypto (ESC 2010) seminar, Remich, Luxembourg, 11-15 January 2010 (2011), `https://cryptolux.org/mediawiki.esc/images/5/52/Esc_nyberg.pdf`
15. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL, USA (1949)