Abstract – Recent cryptanalytic attacks have exposed the vulnerabilities of some widely used cryptographic hash functions like MD5 and SHA-1. Attacks in the line of differential attacks have been used to expose the weaknesses of several other hash functions like RIPEMD, HAVAL. In this paper we propose a new efficient hash algorithm that provides a near random hash output and overcomes some of the earlier weaknesses. Extensive simulations and comparisons with some existing hash functions have been done to prove the effectiveness of the BSA, which is an acronym for the name of the 3 authors.

Keywords: Cryptography, Hash Function, Random, BSA

1. INTRODUCTION

A cryptographic hash function takes an input string of arbitrary length and produces a message digest that is of a fixed, short length (e.g. 128 or 256 or 512 bits). The digest is sometimes also called the "hash" or "fingerprint" of the input. Hash functions are used in many situations where a potentially long message needs to be processed and/or compared quickly and also for security purposes. The most common application is the creation and verification of digital signatures.

The two most widely used cryptographic hash functions are the MD5 [1] and SHA-1 [2]. MD5 was designed by Rivest as a strengthened version of MD4. There had been a lot of tweaks and variants in the MD and the SHA series mostly by increasing the length of the message digest. [3] describes in details of the approach to find collisions in MD5 and break other hash functions like RIPEMD, HAVAL, MD4 and SHA-0 by using differential attacks. This has led to the recent development of many other cryptographic hash functions, each having its own strengths and weaknesses, aiming to be the “one” which is secure against birthday attacks, cube testers, differential cryptanalysis and several other attacks. The NIST hash function competition is an open competition held by the US National Institute of
Standards and Technology, for a new SHA-3 function to replace the older SHA-1 and SHA-2, which was formally announced on November 2, 2007 [6]. NIST selected 51 entries in Round 1 out of which the following 14 algorithms proceeded to Round 2: Blake[7], Blue Midnight Wish[8], CubeHash[9], Echo[10], Fugue[11], Groestl[12], Hamsi[13], JH[14], Keccak[15], Luffa[16], Shabal[17], SHA-vite-3[18], SIMD[19], Skein[20].

In this paper, we have proposed an efficient hash function, which is also a block cipher, where the set of initial values of the initialization vector and all the operations depend solely on the plaintext which leads to an excellent avalanche effect, where flipping a single bit in the plaintext changes more than about 50% bits in the ciphertext. Thus making the hash output nearly random. Extensive simulations have been carried out and comparisons have been made, on various factors like the Runs test, Chi-square test, Entropy, Mean Value and Serial Correlation for testing the randomness of the hash output, with the algorithms selected in Round 2 of the NIST SHA-3 competition to evaluate the performance of this algorithm. The rest of the paper is organized like this. Section 2 discusses some preliminary ideas about the Merkle-Damgard construction of hash functions, the need for randomness testing and a brief note on randomness tests. Section 3 discusses in details about the proposed BSA hash function. Section 4 gives the simulation results of the various tests conducted on the BSA and its performance comparison with some existing algorithms followed by concluding remarks.

2. PRELIMINARIES

2.1 Merkle -Damgard Construction

![Figure 1. Iterative Chaining of Merkle-Damgard Construction](image)

In the figure:
- IV = initial value
- CV = chaining variable
- $Y_i$ = $i$th input block
- f = compression algorithm
- L = number of input blocks
- n = length of hash code
- b = length of input block
Ralph Merkle [5] and Ivan Damgard [4] proposed an iterative chaining function for block ciphers. In this method the input to each compression function will be an initialization vector / initialization value and a chaining variable and the output will go to the next stage. They independently proved if the compression function is collision resistant, then so will be the hash function. In order to strengthen the above construction they further proposed that the padding should contain the length of the original message. This is called length padding or Merkle–Damgard strengthening. A finalization function is often used in the last stage to further compress the hash output or to increase the avalanche effect. Most widely used cryptographic hash functions like Sha-1 and MD5 use this method. At the heart of the BSA lies this form.

The 2 inputs to the compression function, as describe above, are an intermediate hash value, which is the output of the previous compression function, and a message block. If the compression function is an ideal one, the intermediate hash values should be random. For an ideal hash function the output should be random. Thus one of the measures to compare or evaluate the performance of a hash function is to perform randomness tests on the hash output to find the level of randomness of the hash function.

2.2 Tests for Randomness [21]

a) Chi-square Test
This is used to test the validity of a distribution assumed for a random phenomenon. The test evaluates the null hypotheses $H_0$ (that the data are governed by the assumed distribution) against the alternative (that the data are not drawn from the assumed distribution).

b) Run Test
This test is based on based on the frequency of run-lengths (a run is a sequence of consecutive digits)

c) Frequency Test (Mean Value Test)
This test checks that each symbol occurs with equal frequency (for a binary string, proportion of 0’s and 1’s should be 0.5 each)

d) Serial-Correlation test
Correlation coefficients appear frequently in statistics; if we have \(n\) quantities \(U_0, U_1, U_2, \ldots, U_{n-1}\), the correlation coefficient between them is a measure of the amount \(U_{j+1}\) depends on \(U_j\).

e) Entropy Test

The information density of the contents of the file, expressed as a number of bits per character. An entropy value of ‘1’ (for a binary bit stream) indicates that the file is extremely dense in information—essentially random. Hence, compression of the file is unlikely to reduce its size.

3. DESCRIPTION OF THE BSA HASH FUNCTION IN DETAILS

3.1 Description of the Components and Essential Operations in the BSA function

All the operations have been performed in the binary form of the message. The 1’s and 0’s used in this function denote the ‘1’ and ‘0’ bits in the binary message.

3.1.1 Padding

a) The first block to the hash function is length padded. The size of the block is 512 bits. The leftmost 4 words contain the number of 1’s in the message. The rightmost 4 words contain the number of 0’s in the message. The remaining middle 8 words contain the length of the message. The size of each word is 32 bits. Hence, there is a length limitation of the input message. The input message can contain a maximum of \(2^{128}\) 1’s and \(2^{128}\) 0’s, thus a total of \(2^{256}\) bits.

b) The input plaintext message is broken into 448 bit blocks. The 448-bit block message \(M_i\) is extended to a 512-bit block \(B_i\) by padding. Here we define two 16-bit strings \(L_i\) and \(R_i\) where,

\[
L_i = \text{bit position of 1’s in } M_i \text{ xor-ed to each other}
\]

\[
R_i = \text{bit position of 0’s in } M_i \text{ xor-ed to each other}
\]

(bit positions vary from 1-448 in each block). Thus each block input to the hash function is defined as, \(B_i = L_i \oplus R_i \oplus M_i \oplus R_i \oplus L_i\), making the total block length 512 bits.
c) In case the last block in the input message is less than 448 bits, then it is padded with 1’s or 0’s if the total number of 1’s in the entire plaintext is odd or even accordingly.

3.1.2 Crossover

The crossover point or pivot in each block is chosen as the index corresponding to the number of 1’s in that block. Thus the index can vary from 1 to 511 (no crossover for all or no 1’s). The strings on both the sides of that index are swapped to perform the crossover. The figure below shows a string $x_1x_2$, which is, crossed over a chosen crossover point $i$ giving the resultant string $x_2x_1$.

![Figure 3. Crossover](image)

3.1.3 Two Parallel Iterative Chains

The BSA consists of 2 iterative chains, which gives it added strength and security and also speed due to parallelism. Each chain, similar to the Merkel-Damgard construction in Figure 1, has an iterative structure consisting of several compression functions. The final result is obtained by concatenating the output of the 2 chains. The output of each chain is a 256-bit value and hence the final hash output is 512-bits in size.

3.1.4 Initialization Vector

There are 2 initialization vectors corresponding to each of the 2 Iterative Chains. The initialization vectors IV1 is taken as the 256 bits (in order) at odd places in Block 1 and IV2 is taken as the 256 bits (in order) at even places in Block 1.

3.1.5 Compression Functions

Each compression function consists of 16 Rounds (the number of Rounds is adjustable). In each Round a different bitwise operation is used. A function lookup table is used to determine the operation to be used in each Round. The Block number, which goes as one of the inputs
to each compression function, and the Round number give the index to the table. Each compression function operates on a 256-bit value and outputs a 256-bit value.

3.2 Working of the Algorithm in Details

Step 1: Preprocessing

The input message is divided into 448 bit blocks. The last block may need to be padded (Sec 3.1.1.c). The first block is a dummy block that is length padded (Sec 3.1.1.a).

Step 2: Block Processing

Each 448-bit block needs to be preprocessed to make it a 512-bit block (Sec 3.1.1.b) Bᵢ. Crossover (Sec 3.1.2) is performed in the resultant block Bᵢ. The resulting block Cᵢ, after crossover, is also 512-bits in size. The words in Cᵢ in even position, taken in order, form Xᵢ₁ and the words in odd position, taken in order, form Xᵢ₂.

Step 3: Chain Processing

As described in Section 3.1.3, there are 2 iterative chains. In chain 1, the compression functions, used iteratively, are denoted by Fᵢ₁ and that used in chain 2 are denoted by Fᵢ₂.

For the First Block C₁, the input to F₁₁ is X₁₁ and (IV₁ ⊕ X₁₁) and the input to F₁₂ is (~ X₁₂) and (IV₂ ⊕ X₁₂). The output hash value of F₁₁ is denoted by r₁₁ and that of F₁₂ is denoted by r₁₂.

For any other Block Cᵢ (obtained in Step 2):

a. The input to Fᵢ₁ is Xᵢ₁ and (Xᵢ₁ ⊕ rᵢ₋₁).
b. The input to $F_{i2}$ is $(\sim X_{i2})$ and $(X_{i2} \oplus r_{(i+1)2})$.

Step 4: Compression Function Processing

Each compression function $F_{i1}$ or $F_{i2}$ consists of 16 rounds each. The function to be used in each round is determined by using the Block number and Round number as an index to a function lookup table. The function lookup table (Fig. 6) gives the possible input and output bits for a bitwise operation for a specific Round. For example, in the 5th Round of any compression function, if the inputs bits for a bitwise operation are 00, 01, 10, 11 then the outputs bits are 0, 1, 0, 1 respectively i.e. it represents a bitwise XOR operation.

\[ F_{i1} \]
\[ F_{i2} \]

![Figure 5: Compression Function Processing](image)

![Figure 6: Function Lookup Table](image)

Step 4.1: Chain 1

Step 4.1.1: Basic Operation

The basic function number for Block $i$ and Round $j$ in compression function $F_{i1}$ is determined by $(i + j) \mod 16$. This is used as an index to the function lookup table. Let the bitwise Round function be $f_1$ for the $j$th Round for a Block $i$ in chain 1. We have taken the number of Rounds and the number of functions in the lookup table, both as 16.
We define 3 variables $a$, $b$ and $c$ such that $a = (i+j) \mod 256$, $b = (i^*j) \mod 256$ and $c = (i^2) \mod 256$.

The inputs for Round 0 are $X_{i1}$ and $r_{(i-1)1}$, and for any other Round the inputs are $r_{(i-1)1}$ and the output from the previous Round.

We now define the following variables:

- $\text{Reflection11} = f_1(r_{(i-1)1}\text{ in Gray code}, X_{i1}\text{ in reverse order of bits})$.
- $\text{Reflection21} = f_1(r_{(i-1)1}\text{ in reverse order of bits}, X_{i1}\text{ in Gray code})$.
- $\text{Reflection31} = f_1(r_{(i-1)1}, X_{i1})$
- $\text{Rotation11} = \text{Rotate Right (}X_{i1}\text{ by a bits)}$
- $\text{Rotation21} = \text{Rotate Right (}X_{i1}\text{ by b bits)}$
- $\text{Rotation31} = \text{Rotate Left (}r_{(i-1)1}\text{ by c bits)}$

$\text{Result11} = \text{Reflection11} \oplus \text{Rotation11}$

$\text{Result21} = \text{Reflection21} \& \text{Rotation21}$

$\text{Result31} = \neg (\text{Reflection31} \& \text{Rotation31})$

The final result of this round, $r_{i1} = (\text{Result11} \oplus \text{Result21} \oplus \text{Result31})$.

Step 4.1.2: Adjusting the bits

Let $\text{Count1}$ and $\text{Count0}$ be the number of 1’s and 0’s in $r_{i1}$.

Case 1: If $\text{Count1} > \text{Count0}$,

Then first check the $7^{th}$ bit of the $1^{st}$ word of $r_{i1}$. If it is 1 then flip it and check the $17^{th}$ bit, else do the same in the next word of $r_{i1}$.

If the $17^{th}$ bit is a 1, then flip it and check the $27^{th}$ bit of that word or else check the next word. Continue in this fashion till either $\text{Count1} = \text{Count0}$ or the end of $r_{i1}$ is reached.

Case 2: If $\text{Count0} > \text{Count1}$,

Then first check the $1^{st}$ bit of the $1^{st}$ word of $r_{i1}$. If it is 0 then flip it and check the $11^{th}$ bit, else do the same in the next word of $r_{i1}$.

If the $11^{th}$ bit is a 0, then flip it and check the $21^{st}$ bit of that word or else check the next word. Continue in this fashion till either $\text{Count0} = \text{Count1}$ or the end of $r_{i1}$ is reached.

Step 4.2: Chain 2

The basic function number for Block $i$ and Round $j$ in compression function $F_{i2}$ is determined by $(i + (j+49) \mod 50) \mod 16$. This is used as an index to the function lookup table. Let the
bitwise Round function be \( f_2 \) for the \( j \)th Round for a Block \( i \) in chain 2. We have taken the number of rounds and the number of functions in the lookup table, both as 16. We define 3 variables \( a, b \) and \( c \) such that \( a=(i+j) \mod 256 \), \( b=(i^j) \mod 256 \) and \( c=(i^j) \mod 256 \).

The inputs for Round 0 are \( X_{i2} \) and \( r_{(i-1)2} \), and for any other Round the inputs are \( r_{(i-1)2} \) and the output from the previous Round.

We now define the following variables:
- Reflection12 = \( f_2 \) (\( r_{(i-1)2} \) in Gray code, \( X_{i2} \) in reverse order of bits).
- Reflection22 = \( f_2 \) (\( r_{(i-1)2} \) in reverse order of bits, \( X_{i2} \) in Gray code).
- Reflection32 = \( f_2 \) (\( r_{(i-1)2} \), \( X_{i2} \)).
- Rotation12 = Rotate Right (\( X_{i2} \) by a bits)
- Rotation22 = Rotate Right (\( X_{i2} \) by b bits)
- Rotation32 = Rotate Left (\( r_{(i-1)2} \) by c bits)

Result12 = Reflection12 \( \lvert \) Rotation12
Result22 = Reflection22 \& Rotation22
Result32 = \(~\) (Reflection32 \& Rotation32)

Final result of this round, \( r_{i2} = \sim (\text{Result12} \oplus \text{Result22} \oplus \text{Result32}) \).

**Step 5: Iterate and Final Result**

Repeat steps 2-5 for every 448-bit block in the input plaintext message.

The final hash output is given by concatenating the output hash value of each chain i.e. the final hash value is \((r_{N1} \; r_{N2})\), if \( N \) is the total number of blocks in the input plaintext message.

**Step 6: END**

### 4. SIMULATION RESULTS

Extensive simulations have been done to compare and evaluate the performance of the BSA. The tests performed have been categorized as:

i) Collision Tests  
ii) Avalanche Effect Test  
iii) Randomness Tests

i) Collision Tests
20 lakh random strings were generated each having 0-35,000 bits. The randomness of each string was verified using the randomness tests (Section 2.2). The level of significance in Runs and Chi-square tests were taken at .05% level of significance.

Further simulations have been done with 20 lakh strings, each differing from the other by a single bit hamming distance, each having 10000 bits. No collision has been found in the hash outputs of the strings, using BSA, in both cases.

ii) Avalanche Effect Test
In this test, 1 lakh strings each differing from the other by a single bit (Hamming Distance =1) were tested using BSA. The hash outputs of the input strings were compared on the basis of their relative hamming distance. The minimum hamming distance between the hash outputs was found to be 206 bits, the maximum was 288 bits and the average hamming distance was 250 bits. For random strings of arbitrary length, the average hamming distance between the hash outputs was 253 bits.

The BSA was found to perform better than the 12 algorithms (listed in the Section 1 that proceeded to the Second Round in the NIST SHA-3 competition) when we compared the average hamming distance of their hash outputs, on input strings with hamming distance 1. Table 1 below, gives the hamming distance between the hash codes of each algorithm for input strings C0 (Hex) and 80 (Hex) with hamming distance 1. Table 2 shows their BSA hash output in hex.

iii) Randomness Tests
For an ideal hash function the output should be as random as possible. In Section 2.2 we have specified a few tests for evaluating the randomness of a hash function. For an ideal hash function, the entropy of the output hash bitstream should be 1, the serial correlation - coefficient should be 0.0 and the mean value should be 0.5. We have taken the help of a widely used software for conducting the randomness tests, developed by John Walker called the ENT [22] for performing the Entropy, Mean value and Serial Correlation tests.
<table>
<thead>
<tr>
<th>Algorithm Name</th>
<th>Hamming Distance between output Hash</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSA</td>
<td>231</td>
</tr>
<tr>
<td>FUGUE</td>
<td>187</td>
</tr>
<tr>
<td>GROESTL</td>
<td>164</td>
</tr>
<tr>
<td>SHABAL</td>
<td>162</td>
</tr>
<tr>
<td>SHAVITE-3</td>
<td>154</td>
</tr>
<tr>
<td>HAMSI</td>
<td>152</td>
</tr>
<tr>
<td>JH</td>
<td>151</td>
</tr>
<tr>
<td>BLUE</td>
<td>142</td>
</tr>
<tr>
<td>ECHO</td>
<td>136</td>
</tr>
<tr>
<td>SKEIN</td>
<td>136</td>
</tr>
<tr>
<td>BLAKE</td>
<td>136</td>
</tr>
<tr>
<td>LUFFA</td>
<td>130</td>
</tr>
<tr>
<td>KECCAK</td>
<td>124</td>
</tr>
</tbody>
</table>

**Table 1**: Hamming Distance between Hash Codes of Algorithms for Input Strings C0 and 80

<table>
<thead>
<tr>
<th>Input Message</th>
<th>C0 (hex)</th>
<th>80 (hex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSA Hash output</td>
<td>1dd86640 5931a2e3</td>
<td>834c4014 4d773438</td>
</tr>
<tr>
<td></td>
<td>932c2494 ce21cf63</td>
<td>89943f2c 2d28570d</td>
</tr>
<tr>
<td></td>
<td>62d4823b b31a9858</td>
<td>999bc461 91812aa9</td>
</tr>
<tr>
<td></td>
<td>12502e02 e420873e</td>
<td>83ee1484 e34040a5</td>
</tr>
<tr>
<td></td>
<td>c9316b5b 6b6f8f66</td>
<td>610bc1fb 2fa1c41f</td>
</tr>
<tr>
<td></td>
<td>3e8b6d2d 9e6c488e</td>
<td>3d8b2f3c a8a9258c</td>
</tr>
<tr>
<td></td>
<td>a549f56e cdab4f49</td>
<td>8e8179bd bf95592e</td>
</tr>
<tr>
<td></td>
<td>3379cb87 4c756713</td>
<td>737ee5aa c4b5c91c</td>
</tr>
</tbody>
</table>

**Table 2**: BSA Hex Code of Strings C0, 80
<table>
<thead>
<tr>
<th>Algorithm Name</th>
<th>Entropy Value</th>
<th>Mean Value</th>
<th>Serial Correlation Value</th>
<th>Runs Test</th>
<th>Chi square Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blake</td>
<td>0.98847</td>
<td>0.4369</td>
<td>0.0040</td>
<td>94.9</td>
<td>98</td>
</tr>
<tr>
<td>Blue Midnight Wish</td>
<td>0.98856</td>
<td>0.4371</td>
<td>0.0039</td>
<td>94.4</td>
<td>98</td>
</tr>
<tr>
<td>Echo</td>
<td>0.98854</td>
<td>0.4371</td>
<td>0.0038</td>
<td>96.5</td>
<td>99</td>
</tr>
<tr>
<td>Fugue</td>
<td>0.98853</td>
<td>0.4370</td>
<td>0.0031</td>
<td>95.2</td>
<td>99</td>
</tr>
<tr>
<td>Grostl</td>
<td>0.98850</td>
<td>0.4370</td>
<td>0.0035</td>
<td>95.4</td>
<td>98</td>
</tr>
<tr>
<td>JH</td>
<td>0.98856</td>
<td>0.4371</td>
<td>0.0034</td>
<td>94.9</td>
<td>99</td>
</tr>
<tr>
<td>Shabal</td>
<td>0.98857</td>
<td>0.4371</td>
<td>0.0032</td>
<td>95.1</td>
<td>98</td>
</tr>
<tr>
<td>Shavite-3</td>
<td>0.98844</td>
<td>0.4368</td>
<td>0.0036</td>
<td>94.7</td>
<td>98</td>
</tr>
<tr>
<td>Hamsi</td>
<td>0.98850</td>
<td>0.4370</td>
<td>0.0035</td>
<td>95.4</td>
<td>99</td>
</tr>
<tr>
<td>Keccak</td>
<td>0.98852</td>
<td>0.4370</td>
<td>0.0041</td>
<td>95.1</td>
<td>99</td>
</tr>
<tr>
<td>Luffa</td>
<td>0.98863</td>
<td>0.4373</td>
<td>0.0036</td>
<td>94.6</td>
<td>99</td>
</tr>
<tr>
<td>Skein</td>
<td>0.98858</td>
<td>0.4372</td>
<td>0.0036</td>
<td>95.2</td>
<td>99</td>
</tr>
<tr>
<td>BSA(Input String ≤ 2Kb)</td>
<td>0.98892</td>
<td>0.4381</td>
<td>0.0006</td>
<td>94.3</td>
<td>98.5</td>
</tr>
<tr>
<td>BSA(any Input String size)</td>
<td>0.98854</td>
<td>0.4371</td>
<td>0.0027</td>
<td>93.5</td>
<td>97.5</td>
</tr>
</tbody>
</table>

Table 3: Randomness Test Results between Hash Codes of Different Hash Functions
Further, we have performed the Runs test and Chi-square tests ourselves at 0.05% and 0.01% level of significance. All the tests were conducted on 2600 strings, each having 0-35,000 bits. Table 3 gives the comparison between different algorithms and the BSA on the results of the randomness tests conducted on the output hash bitstream of each algorithm. The entropy value is measured per bit, the Runs test and Chi-square test results of each algorithm indicate the percentage of the total number of hash outputs that are accepted by those tests at 0.05% and 0.01% level of significance.

5. Conclusion and Future Work

It is clearly evident from the simulation results that BSA performs the best when all the criteria for randomness and avalanche effects are combined. Further tests are underway in evaluating the strength of the BSA against differential cryptanalysis and other cryptographic attacks. The BSA showed promising results for all the tests conducted so far.

References:
[17] Jean-François Misarsky; Emmanuel Bresson, Anne Canteaut, Benoît Chevallier-Mames, Christophe Clavier, Thomas Fuhr, Aline Gouget, Thomas Icart, Jean-François Misarsky, Maria Naya-Plasencia, Pascal Paillier, Thomas Pornin, Jean-René Reinhard, Céline Thuillet, Marion Videau "Shabal, a Submission to NIST’s Cryptographic Hash Algorithm Competition", Candidate to the NIST Hash Competition, 2008