

Descifrado de Cesar por Medio de Análisis de Frecuencia para Tres Idiomas

Decrypting Cesar by Means of Frequency Analysis for Three Languages

Bárbara Emma Sánchez Rinza , Marlon Paul Flores Sayago, T. Rocha-Rinza 

Benemérita Universidad Autónoma de Puebla, UNAM, Facultad de Ciencias de la Computación, Instituto de Química
14 sur y avenida San Claudio

* Correo-e: brinza@hotmail.com

PALABRAS CLAVE: RESUMEN

Algoritmo de Cesa, análisis de frecuencias, cifrar, descifrar, texto llano.

En este trabajo se cifrar por medio del algoritmo de Cesar y se descifrar por fuerza bruta, mediante un análisis de frecuencias, se introduce el análisis de frecuencia para 3 idiomas para que el criptograma, se pueda descifrar en estos 3 idiomas elegidos, los textos llanos pueden estar en español, inglés, y portugués, la interfaz contará con tres secciones (texto plano, texto cifrado y texto descifrado), un área de salida de datos tipo consola y un área de comandos. Estos deben permitir las acciones de cargar texto desde un archivo, limpiar los campos, realizar el cifrado y descifrado, seleccionar un idioma de trabajo y realizar el ataque por análisis de frecuencias. Ya que el análisis de frecuencia varía dependiendo del idioma que utilizemos.

KEYWORDS: ABSTRACT

Cesa algorithm, frequency analysis, encrypt, decrypt, plain text.

In this work, it is encrypted by means of Cesar's algorithm and decrypted by brute force, through a frequency analysis, the frequency analysis is introduced for 3 languages so that the cryptogram can be deciphered in these 3 chosen languages, the plain texts They can be in Spanish, English, and Portuguese, the interface will have three sections (plain text, ciphertext and decrypted text), a console-type data output area and a command area. These should allow the actions of loading text from a file, clearing the fields, performing encryption and decryption, selecting a working language and performing the attack by frequency analysis. Since the frequency analysis varies depending on the language we use.

Recibido: 12 de julio de 2021 • **Aceptado:** 11 de septiembre de 2021 • **Publicado en línea:** 15 de febrero de 2022

1 INTRODUCCIÓN

La seguridad informática es una disciplina que se encarga de proteger la integridad y la privacidad de la información almacenada en un sistema informático.

Un sistema informático puede ser protegido desde un punto de vista lógico (con el desarrollo de software). Las amenazas pueden proceder desde vía remota (los delincuentes que se conectan a Internet e ingresan a distintos sistemas). En la actualidad existe mucha delincuencia cibernética, que realiza algoritmos para robar información valiosa, es por tal motivo que se realizó este trabajo, para poder romper este tipo de criptogramas.

Un sistema seguro debe ser íntegro (con información modificable sólo por las personas autorizadas), confidencial (los datos tienen que ser legibles únicamente para los usuarios autorizados), irrefutable (el usuario no debe poder negar las acciones que realizó) y tener buena disponibilidad (debe ser estable).

De todas formas, como en la mayoría de los ámbitos de la seguridad, lo esencial sigue siendo la capacitación de los usuarios. Una persona que conoce cómo protegerse de las amenazas sabrá utilizar sus recursos de la mejor manera posible para evitar ataques o accidentes. En otras palabras, puede decirse que la seguridad informática busca garantizar que los recursos de un sistema de información sean utilizados tal como una organización o un usuario lo ha decidido, sin intromisiones. Pero antes de ir más adelante revisaremos algunos conceptos importantes.

La palabra “criptografía” proviene etimológicamente del griego Kriptos (ocultar), Graphos (escritura), “ocultar la escritura”. En un sentido más amplio significa aplicar alguna técnica para hacer ininteligible un mensaje [1]. La criptografía es una herramienta muy útil cuando se desea tener seguridad informática; puede ser también entendida como un medio para garantizar las propiedades de confidencialidad, integridad y disponibilidad de los recursos de un sistema [2,3,4]. La criptología consiste en permitir el intercambio de información a través de un medio de comunicación inseguro, de forma que, si la información es interceptada por un intruso, sea imposible su descifrado.

Esta ciencia está dividida en dos grandes ramas:

- La criptografía: ocupada del cifrado de mensajes en clave y del diseño de criptosistemas.
- El criptoanálisis: que trata de descifrar los mensajes en clave, rompiendo así el criptosistema.

Para este trabajo, se requirió programar una aplicación que pueda mostrar el resultado de un texto plano tras ser cifrado con el algoritmo de César, y se realiza un auto ataque al resultado cifrado para descubrir el texto plano original, mediante un análisis de frecuencias, de acuerdo con un idioma seleccionado que puedes ser español, inglés y portugués (de los 3 idiomas, hace el software el análisis de frecuencia) [5,6,7].

El análisis de frecuencia es una lista de letras con su correspondiente frecuencia en un determinado idioma, para este trabajo solo se seleccionaron 3 idiomas español, inglés y portugués. El usuario debe especificar qué idioma se está ingresando para realizar correctamente el cifrado, descifrado y que el programa utilice adecuadamente el análisis de frecuencias del idioma seleccionado.

Este texto de entrada puede ser de formato libre, puede incluir espacios, saltos de línea, signos, números, símbolos y puntuaciones, ya que se programó para 256 caracteres.

Este trabajo acepta como medio de entrada tanto texto ingresado por teclado, como texto obtenido de algún archivo o documento con formato de texto plano, que se podrá seleccionar desde dentro de la misma aplicación [8,9,10].

2 CIFRADO DE CÉSAR

El cifrado de César es un cifrado de textos planos muy antiguo, se cree que Julio César lo utilizó para dirigir mensajes confidenciales a sus generales en campañas militares en el primer siglo antes de Cristo. Se basa en un cifrado mono alfabético.

En criptografía, un cifrado César se clasifica como un cifrado por sustitución en el que el alfabeto en el texto plano se desplaza por un número fijo en el alfabeto [11].

2.1 LAS VENTAJAS DE USAR UN CIFRADO CÉSAR INCLUYEN:

- ❖ Uno de los métodos más fáciles de usar en criptografía y puede proporcionar una seguridad mínima a la información.
- ❖ Uso de solo una tecla breve en todo el proceso
 - ❖ Uno de los mejores métodos para usar si el sistema no puede usar ninguna técnica de codificación complicada
- ❖ Requiere pocos recursos informáticos

2.2 LAS DESVENTAJAS DE USAR UN CIFRADO CÉSAR INCLUYEN:

- ❖ Uso de estructura simple
- ❖ Solo puede proporcionar seguridad mínima a la información
- ❖ La frecuencia del patrón de letras proporciona una gran pista para descifrar el mensaje completo.

Dado un **abecedario definido** para este trabajo se escogieron 256 caracteres con todos los caracteres que se desea cifrar, y número **N** entero positivo no mayor que el tamaño de elementos del abecedario en uso, este cifrado consiste en ingresar un texto plano y recorrerlo carácter por carácter, cambiando cada uno por el símbolo que se ubique **N** posiciones después, circularmente dentro del abecedario. Para explicar el algoritmo tomaremos para ilustra un abecedario estándar de 26 letras, y un desplazamiento **N=13**, al encontrar una letra “a”, ésta se cambiará por “n”, y una “z” por una “m”, pues al alcanzar el final, se continúa el desplazamiento desde el principio del abecedario [12] ver figura 1.

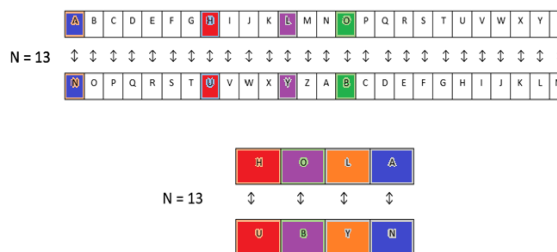


Figura 1. Representa el desplazamiento en 13

Si este algoritmo se representa en forma matemática queda $y = x+N$

Posteriormente, para realizar el descifrado, se debe realizar el mismo proceso de recorrido y cambio de caracteres del texto, pero en lugar de desplazar **N** posiciones después del símbolo actual (a la derecha), se desplazarán **N** posiciones antes de éste (a la izquierda). Dicho matemáticamente, al cifrar se deben sumar posiciones al símbolo actual, y al descifrar, se deben restar, para así obtener el resultado original **Y** si formo matemática quedan $y=x-N$ [7].

3 ANALISIS DE FRECUENCIA

En que consiste el análisis de frecuencia, en cualquier idioma tenemos unas letras más comunes que otras. Y para realizar este análisis nos valdremos de 5 pasos [6].

- 1) Se analiza la frecuencia de cada letra del idioma español (inglés o portugués) y a cada letra se le da un valor determinado en función de su frecuencia de uso.
- 2) Se analizan diferentes textos en español (inglés o portugués) para validar los valores anteriores
- 3) Una vez obtenida esta tabla se lee el texto cifrado
- 4) Se reconoce el texto cifrado y se hace un análisis de la frecuencia de aparición de los diferentes símbolos. Se crea una nueva tabla con estos resultados.
- 5) Los valores de ambas tablas se comparan luego por su frecuencia y los símbolos del texto cifrado se sustituyen por las letras del alfabeto correspondientes.

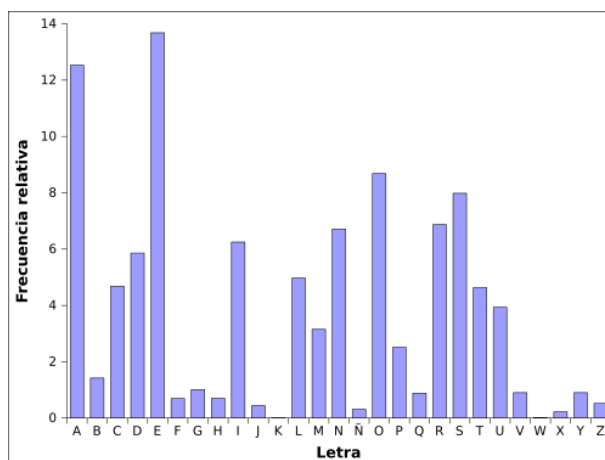


Figura 2. Frecuencias en el idioma español

Por orden de frecuencia, las 10 letras más usadas son E-A-O. Los tres primeros lugares de lista son vocales y es que por lo general la mitad de cualquier texto que leemos está formado por vocales.

En el diccionario de la Real Academia Española (RAE) la letra más frecuente es la A, pero en cualquier texto en castellano, la frecuencia de con la que se utilizan los monosílabos «que», «el», «se», «me», hace que la «e» sea más frecuente.

3.1 ANALISIS DE FRECUENCIA EN ESPAÑOL

En la figura 2 podemos observar las frecuencias de letras del idioma español

3.2 FRECUENCIA DE LETRAS EN EL IDIOMA INGLÉS

El siguiente es el resultado de un análisis de las letras que aparecen en las palabras enumeradas en las entradas principales del Concise Oxford Dictionary (novena edición, 1995) y se obtuvo la tabla 1:

Tabla 1. Frecuencias de letras en el idioma inglés

E	11.1607%	56.88	M	3.0129%	15.36
A	8.4966%	43.31	H	3.0034%	15.31
R	7.5809%	38.64	G	2.4705%	12.59
I	7.5448%	38.45	B	2.0720%	10.56
O	7.1635%	36.51	F	1.8121%	9.24
T	6.9509%	35.43	Y	1.7779%	9.06
N	6.6544%	33.92	W	1.2899%	6.57
S	5.7351%	29.23	K	1.1016%	5.61
L	5.4893%	27.98	V	1.0074%	5.13
C	4.5388%	23.13	X	0.2902%	1.48
U	3.6308%	18.51	Z	0.2722%	1.39
D	3.3844%	17.25	J	0.1965%	1.00
P	3.1671%	16.14	Q	0.1962%	(1)

La tercera columna representa proporciones, tomando la letra menos común (q) como igual a 1. La letra E es más de 56 veces más común que la Q en la formación de palabras individuales en inglés. La frecuencia de las letras al comienzo de las palabras vuelve a ser diferente. Hay más palabras en inglés que comienzan con la letra 's' que con cualquier otra letra. (Esto se debe principalmente a que grupos como 'sc', 'sh', 'sp' y 'st' actúan casi como letras independientes). La letra 'e' solo aparece en la mitad del orden y la letra 'x' Como era de esperar, viene en último lugar. Todo esto lo podemos apreciar en la gráfica de la figura 3.

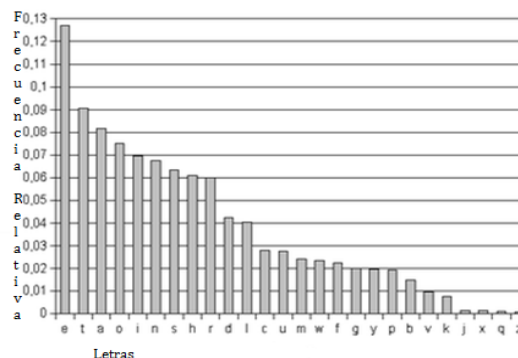


Figura 3. Grafica de las letras más frecuentes en inglés.

3.3 FRECUENCIA DE LETRAS EN EL IDIOMA PORTUGUÉS

El alfabeto portugués es de 38 letras, ver tabla 2. Contiene de todas las 26 letras básicas del alfabeto latino y se completa con otras letras. Pero esto no afecta en el software que se realizó ya que este cuenta con 256 caracteres, números, letras, letras acentuadas, caracteres especiales etc.

Tabla 2. Alfabeto Portugués

a	b	c	d	e	f	g	h	i	j	k	l	M	n	o
p	q	r	s	t	u	v	w	x	y	z	ã	á	â	ã
ç	è	é	ê	ë	í	î	ó	ô	õ	ù	ú	ü		

A continuación, se usa una gráfica de frecuencias relativa del idioma portugués ver figura 4

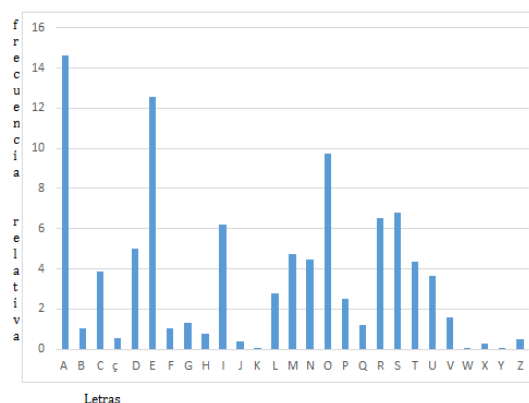


Figura 4. Grafica de frecuencias del portugués

4 DESARROLLO

Se realizó un programa ejecutable en lenguaje Java mediante el IDE NetBeans, consistiendo en dos clases, la **primera clase** contiene varios métodos, dos de ellos son cifrar() y descifrar(), ambas reciben una cadena como parámetro que representa al texto que se transformará, otro es la función getInd() de apoyo de ejecución para resolver el índice (dentro del abecedario) de un carácter dado, y otras tres funciones muy similares entre ellas, que son los analizadores de frecuencias de cada idioma llamados frecuencias_esp(), frecuencias_en() y frecuencias_pt().

La **segunda clase** comprende la interfaz de usuario con sus diferentes elementos; aquí se cuenta con un botón para cada acción, etiquetas y los tres campos de texto para contener las diferentes etapas del texto original y el de salida de tipo consola.

Dentro de la clase **Funciones**, se conservan cuatro variables globales ya predefinidas únicamente: el valor N que es el desplazamiento del cifrado, y tres arreglos de caracteres M, Mi, y Mp, que son los abecedarios en los idiomas que funcionará el programa: español, inglés y portugués, respectivamente.

El método **Cifrar** recibe la cadena de texto plano y un ID de idioma establecido: para todos los idiomas se realiza un recorrido por todos los elementos de esta cadena y se realiza el desplazamiento definido, pero según el idioma seleccionado, se trabaja sobre el abecedario que le corresponde. Cada carácter cifrado se va almacenando en otra cadena, la cual es enviada como salida al final del ciclo.

En el método **Descifrar** se recibe la cadena de texto cifrado y la ID de idioma, y se realiza el mismo proceso de recorrido, desplazamiento negativo, caso por idioma, y salida de una cadena con los caracteres resultantes al terminar el recorrido.

Los métodos de **Análisis de Frecuencias** realizan el mismo procedimiento, pero cada uno trabaja con un idioma específico: se recibe una cadena de texto cifrado para analizar, después se definen las frecuencias oficiales de cada letra del abecedario del idioma correspondiente (como porcentaje) en un arreglo de flotantes del mismo tamaño del abecedario; se define también un arreglo del mismo tamaño para llevar los contadores de cada letra (el cual es llenado con ceros al inicio), los cuales se irán incrementando en seguida al realizar el conteo de apariciones de los caracteres en la cadena, posteriormente, estos resultados de frecuencia parciales se almacenarán en una cadena de salida hacia la consola de la interfaz, almacenando hasta este punto el número de apariciones de cada letra, su porcentaje de aparición dentro del texto completo, y su porcentaje de frecuencia oficial en el idioma, para comparar fácilmente.

Después, se define una matriz del tamaño del abecedario usado, con dos columnas, una para guardar y evaluar cada desplazamiento posible y la otra para guardar una “calificación” de su evaluación.

Esta evaluación consiste en un **análisis de similitud de dos curvas gráficas**: si colocamos cada letra del idioma elegido en una gráfica de línea con su frecuencia de aparición oficial correspondiente, obtendremos una forma o un conjunto de picos, y podremos observar que cada pico es diferente del siguiente o el anterior, pues se encuentra a alturas distintas; esto es como una “huella digital” para el idioma, y para cualquier otro grupo de texto, ya que ésta es muy particular y es bastante improbable replicar o asemejar una forma así sin que corresponda con otro texto del mismo idioma. Aprovechando esta característica, se realiza una comparación para cada uno de los desplazamientos posibles dentro del abecedario, comparando las frecuencias obtenidas de cada letra con la siguiente, y si entre ellas forman el pico esperado de acuerdo a la gráfica oficial (pico **positivo** o **negativo**) entonces la calificación del desplazamiento que se esté evaluando se incrementa en 1, y así se realiza para cada par de frecuencias de letras del abecedario, y posteriormente se rota una posición el inicio de las comparaciones para así evaluar el siguiente desplazamiento posible.

Al finalizar este ciclo de evaluaciones, se ordenan descendientemente de acuerdo con su evaluación obtenida, esto debido a que la calificación que obtiene cada desplazamiento evaluado representa el número de picos de la gráfica en los que el texto analizado se asemeja con el idioma de cifrado en ese desplazamiento.

Finalmente, se eligen de esa lista los 5 lugares más altos para ser mostrados en consola como los desplazamientos más probables, junto con su evaluación obtenida y su porcentaje de similitud.

En la otra clase, llamada **Ventana**, se realiza la interfaz visual de usuario, se ajustan los componentes anteriormente mencionados, y en cada botón, se realizan los procedimientos correspondientes.

Existe una caja de opciones para elegir un idioma para trabajar el texto correctamente, esto es necesario para realizar cualquier otro proceso; las opciones definidas son los tres idiomas que se han mencionado anteriormente. El botón Atacar realiza el análisis de frecuencias correspondiente según el **ver** idioma seleccionado. El botón Limpiar simplemente vacía las tres áreas de texto.

El botón Cargar inicializa una GUI de selección de archivo para cargar un archivo de texto en el área de texto. De texto plano, u opcionalmente, se puede escribir en el área mencionada para cifrar un texto ingresado. Los botones de Cifrar y Descifrar toman el texto de su izquierda y lo envían a los métodos Cifrar y Descifrar, respectivamente, para después colocar el texto de resultado en su área de texto correspondiente.

A continuación, se muestran capturas de pantalla de la interfaz de usuario ver Figura 5.

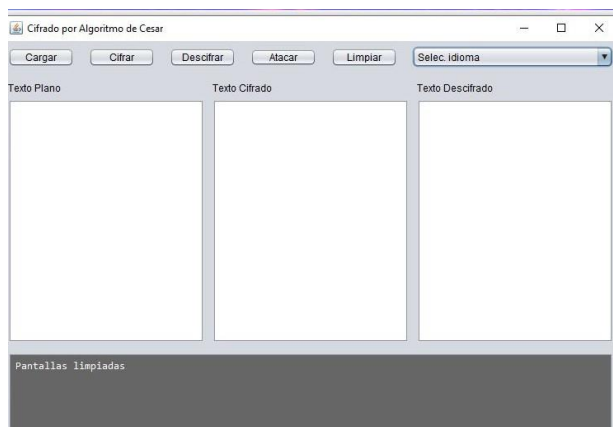


Figura 5. Interfaz de inicio del programa, se observan los componentes descritos

Inicialmente se utiliza la función de carga de texto desde archivos para cargar un texto amplio de prueba de literatura en español, se selecciona idioma español y se realiza el cifrado, el resultado se muestra en la Figura 6.

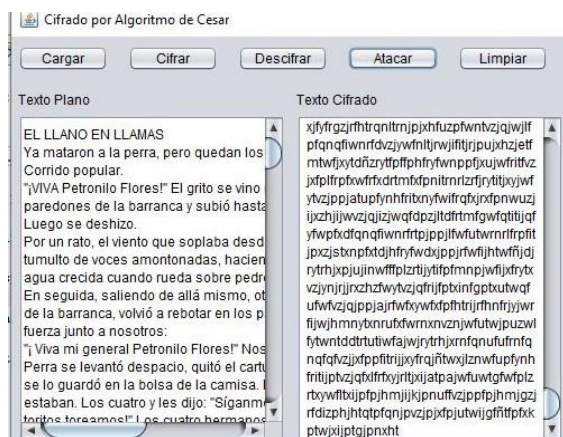


Figura 6. Cifrado del texto de entrada.

Posteriormente, se realiza el auto ataque del texto cifrado, que nos da como resultado en la Figura 7 el análisis de frecuencias con los desplazamientos evaluados mostrados en consola, y el texto descifrado con el desplazamiento más probable, descubierto correctamente.

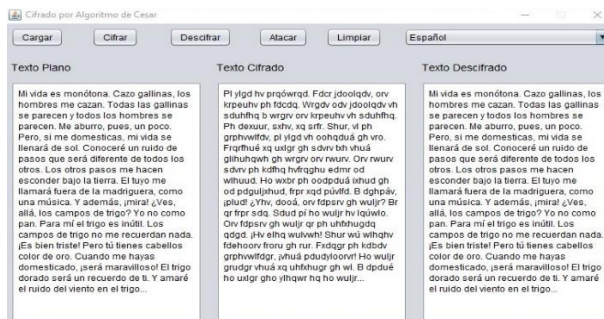


Figura 7. Ejecución del análisis de frecuencias, vista general de la interfaz.

Se puede apreciar en el texto llano mayúsculas, minúsculas, acentos, signos de interrogación, y a la hora de descifrar por análisis de frecuencia, el programa lo realiza perfectamente bien.

Ahora se realizó para el idioma inglés, ver figura 8

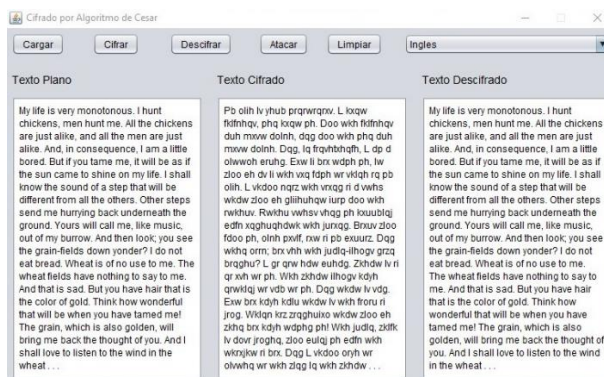


Figura 8. Ataque de por análisis de frecuencia en el idioma ingles

Lo mismo hizo el programa para el idioma inglés, el texto llano tiene mayúsculas, minúsculas, signos de puntuación etc. Y el programa lo descifro correctamente. Ahora observemos para el idioma portugués, figura 9



Figura 9. Ataque por análisis de frecuencia para el idioma portugués

Podemos observar, que el programa también descifro bien, mayúsculas, minúsculas, acentos, signos de puntuación etc. Para el idioma portugués.

5 CONCLUSIONES

Como conclusión, se desarrolló satisfactoriamente un programa que aplica este cifrado de una forma

práctica, visible y entendible; así mismo, para diversos abecedarios e idiomas, y un autoataque mediante el uso de análisis de frecuencias, y, a partir del conocimiento obtenido de análisis de frecuencia para diferentes idiomas, basadas en evaluaciones de similitud de curvas gráficas, por lo que también se pueden conseguir diseñar algoritmos de cifrados derivados de éste, de forma autónoma. Cada recalcar que se obtuvo como un buen porcentaje del texto descifrado.

REFERENCIAS

- [1] Rubén Daniel Varela Velasco.. *Criptografía, una necesidad moderna*. Revista Digital Universitaria, (10 de julio de 2006) Número 7, Volumen 7, 1067-6079.
- [2] Gibrán Granados Paredes. *Introducción a la criptografía*. Revista Digital Universitaria, (10 de julio de 2006). Número 7, Volumen 7, 1067-6079
- [3]Manuel J. Prieto. *Historia de la criptografía*. Madrid: La Esfera de los Libros, (2020). 978-84-9164-737-9.
- [4] Gabriel Sánchez Cano. *Seguridad cibernética. Hacking ético y programación defensiva*. México,: (2018). Alfaomega, 978-607-538-294-4.
- [5] David Moisés Terán Pérez, *Administración y seguridad en redes de computadoras*. (2018) Alfaomega, 978-607-538-132-9.

- [6] Barbara E. Sanchez Rinza, Diana Alejandra Bigurra Zavala, Alonso Corona Chávez, *De-Encryption Of A Text In Spanish Using Probability And Statistics*, (2008) IEEE vol 18, ISBN 978-07695-3120-5
- [7] Christof Paar, Jan Pelzl.). *Understanding Cryptography. A textbook for student and practitioners*. (2010): Springer, 978-3-642-04100-6
- [8] Stuart McClure “*Hackers secretos y soluciones para la seguridad de redes*”, 2000. Mc Graw Hill,
- [9] Jonathan Knudsen “ *Java Cryptography*”, 1998, O`reilly,
- [10] Rolf OPplieger “*Sistemas de autenticación para seguridad en redes*, 2002 ra-ma,
- [11] Randall K. Nichols, “*Seguridad para comunicaciones inalámbricas*, 2003, Mc Graw Hill,
- [12] Eric Maiwald, *Fundamentos de seguridad de redes*, 2000 Mc Graw Hill,

Acerca de los autores



Bárbara Emma Sánchez Rinza. Licenciada en Física, Maestría en Óptica, Doctora en Óptica. Ha escrito 60 capítulos de libros, 56 artículos nacionales e internacionales, 12 memorias. Ha participado en 105 conferencias en diferentes foros. Ha dirigido 34

Tesis de Licenciatura y 10 Tesis de Maestría y 2 tesis de doctorado.



Marlon Paul Flores Sáyago. Mexicano, originario de Iguala de la Independencia, Guerrero, Estudiante de la Licenciatura en Ciencias de la Computación, de la Facultad de Ciencias de la Computación, de la Benemérita

Universidad Autónoma de Puebla. Realizó un período académico en

el Instituto de Ciencias Matemáticas de Computación (ICMC) de la Universidad de São Paulo, campus São Carlos, Brasil. Actualmente se desempeña como Associate Software Test Engineer en Ellucian (Puebla).



Tomás Rocha Rinza obtuvo su doctorado en la Universidad Nacional Autónoma de México y posteriormente hizo una estancia postdoctoral en la Universidad de Århus en Dinamarca. Luego, se

incorporó como investigador al Instituto de Química de la UNAM. Sus líneas de investigación se centran en la aplicación y el desarrollo de métodos de análisis de funciones de onda electrónicas. Su trabajo se ve plasmado en 60 artículos publicados en revistas de arbitraje internacional en los campos de la Química física y la Química general.