Universidad de Buenos Aires

Facultades de Ciencias Económicas, Ciencias Exactas y Naturales e Ingeniería

Maestría en Seguridad Informática

Tesis

Tema:

Redefiniendo límites de las amenazas digitales: Malware que infecta genomas sintéticos bacterianos

Autor: Raphael Labaca Castro
Director de Tesis: Julio Ardita
Año de Presentación: 2015
Cohorte del Maestrando: 2013

Declaración jurada de origen de los contenidos

Por medio de la presente, el autor manifiesta conocer y aceptar el Reglamento de Tesis vigente y se hace responsable de que la totalidad de los contenidos del presente documento son originales y de su creación exclusiva, o bien pertenecen a terceros u otras fuentes, que han sido adecuadamente referenciados y cuya inclusión no infringe la legislación Nacional e Internacional de Propiedad Intelectual.

Esp. Ing. Raphael Labaca Castro
DNI: 18.871.726

Resumen

En 2010, científicos del JCVI anunciaron [1] que lograron crear por primera vez una célula artificial capaz de autoreplicarse. Esto constituye la prueba fehaciente de que se puede diseñar un genoma en una computadora, sin utilizar ninguna porción de ADN natural, y luego crearlo químicamente en un laboratorio. A modo de verificar que se trate de un genoma netamente sintético, se codificó en su contenido URL y correo electrónico de contacto sin que estos afectaran el comportamiento del organismo.

Este hallazgo generaría un cambio de enfoque en la seguridad de la información, debido a que los archivos con información genómica deberían también preservar su integridad para evitar modificaciones maliciosas. En esta tesis se estudiará la infección a una secuencia genómica y se analizará el impacto de que sea maliciosamente modificada. Como prueba de concepto, se emulará un escenario de compromiso utilizando scripts en Python y se infectará la secuencia con un código malicioso. Posteriormente, se utilizará un segundo script para decodificar el código y conectar a un navegador o lanzar una aplicación sin consentimiento de la víctima y sin detección por parte de la solución de seguridad. Finalmente, se analizará la complejidad de detectar dichas modificaciones reactivamente.

Palabras clave: hacking genómico, seguridad en secuencias genómicas, integridad genomas sintéticos

Índice

Resumen iii
Agradecimientos vi
Prólogo vii
1. Introducción 1
1.1. Situación actual 1
2. Definiciones básicas 3
2.1. Células 3
2.2. ADN 3
2.3. Genomas 3
2.4. Secuenciación 4
2.5. Representación de secuencias genómicas 4
3. Caso de Estudio: Infección a un archivo genómico (FASTA) 7
3.1. Objetivo 7
3.2. Importancia de entender este ataque 8
3.2.1. Factores que favorecen 8
3.2.2. Factores limitantes 10
3.3. Posible escenario I: Propagar secuencia infectada 10
3.3.1. Fase 1: Adquisición de secuencia modificada 10
3.3.2. Fase 2: Localización y ejecución 11
3.4. Posible escenario II: Ataque dirigido 11
3.5. Algoritmos implementados 12
3.5.1. Codificar payload e infectar genomas 12
3.5.2. Localizar y descifrar payload en genoma infectado 15
3.6. Detección y prevención del ataque 17
3.6.2. Detección de los scripts 18
3.7. Comportamiento de las base de datos de secuencias 20
3.8. Complejidad para Identificar la raíz del ataque 22
3.9. Localización del *payload* en el genoma sintético 22
4. Desafío: ¿dónde insertar el payload? 24
4.1. *Junk* DNA: ¿realmente redundante? 25
5. Privacidad en las secuencias públicas 27
6. Futuro: ¿Amenaza biológica o digital? 30

6.1. Propagación biológica de una amenaza digital 31
6.1.1. Modelo de propagación de un virus informático 32
6.1.2. Modelo de propagación de un virus biológico 33
6.2. Accesibilidad del proceso de síntesis 35
6.3. Código malicioso + genoma sintético = ¿arma biológica? 35
6.4. DIY-Biology 36
7. Revolución en la secuenciación de ADN 38
7.1. Programa "Medicina Precisa" 39
7.2. Google Genomics 40
8. Conflictos éticos y regulaciones 41
8.1. Privacidad y confidencialidad en el uso de ADN humano 42
9. Conclusión 45
10. Glosario 47
11. Anexo 49
11.1. Análisis de frecuencias en marcas de agua 49
12. Bibliografía 52

Agradecimientos

Quisiera aprovechar estas líneas para recordar a aquellos que fueron muy importantes durante mi vida académica. Esta aventura no hubiese sido posible sin el apoyo de un grupo de personas, cuyos valiosos aportes fueron claves en el desarrollo de esta investigación, ya sea a través de revisiones, sugerencias, respondiendo consultas o simplemente manteniendo interesantes conversaciones.

Me gustaría empezar agradeciendo a Julio, mi tutor, por haber encontrado tiempo en su ocupada agenda para revisar los borradores con dedicación. Al Dr. Patricio Yankilevich, por el apoyo y las sugerencias respecto a dudas de índole biológica. A mi querido Agustín, colega y amigo, que se tomó el trabajo de revisar algunos desarrollos aquí expuestos. A mi madre adoptiva, Susana, que fue una pieza fundamental en mi vida académica y profesional, gran fuente de inspiración y valor. También a Belén María, que durante este proceso se transformó en una compañera muy importante y con la cual compartí interesantes pensamientos e ideas. Y para culminar, un agradecimiento especial a mis padres, que me apoyaron siempre incondicionalmente, proporcionándome los medios para realizar mi carrera universitaria a través de un gran sacrificio y, sobre todas las cosas, porque me enseñaron que no hay límites para alcanzar lo que queremos.

Finalmente, quisiera dedicar esta tesis a todos aquellos que con pasión buscan incansablemente poner a prueba la seguridad de todo lo que nos rodea y tienen el valor de compartir sus hallazgos con la comunidad. Una tarea nada sencilla en una época donde no solo el mercado negro, sino también empresas e incluso gobiernos están dispuestos a pagar cifras exorbitantes por fallos de seguridad no detectados.

Gracias a todos por su apoyo y que este fin augure un gran comienzo.

Prólogo

Debido a la naturaleza interdisciplinaria de esta investigación, el autor considera importante hacer algunas aclaraciones previas al desarrollo. Este escrito trata de abordar cómo sería una posible infección a una secuencia de un genoma (en este caso sintético) a través de un script que extrae una cadena maliciosa, previamente inyectada, y la ejecuta en el sistema de la víctima.

El impacto de este tipo de ataque podría ser clasificado en: digital, digital-biológico y biológico. Este trabajo se centrará en estudiar el primero y meramente describir los restantes, debido a que para un profundo análisis se requieren determinados conocimientos de biología molecular y genética que escapan al objetivo de este trabajo. A continuación, se describen los casos mencionados:

I. *Impacto digital: El hecho de que se pueda inyectar un payload malicioso a una secuencia de ADN no implica que esta metodología agravaría la infección, sino que se agravaría la complejidad de su identificación y posteriormente detección por parte de las metodologías tradicionales de protección previstas para este caso, como por ejemplo, hashes para asegurar la integridad y antivirus para detectar archivos maliciosos. Por lo cual, se demuestra como funcionaría este escenario para alertar el posible uso de secuencias genómicas como vectores de ataque.*

II. *Impacto digital – biológico: En el supuesto caso que se modifique maliciosamente una secuencia de un genoma y éste sea exitosamente sintetizado, el código malicioso podría permanecer en la célula viva y no impactarla biológicamente. Vale aclarar que esto no fue comprobado por el autor y se basa en el hecho que el JCVI logró hacerlo [1] con una dirección URL. Si esto se cumple, ese organismo cargaría un código malicioso cuyo ADN podría ser luego secuenciado en un laboratorio y generar un archivo de secuencia que poseería, por*

ejemplo, una porción de código maliciosa. Restaría a un atacante extraerla y ejecutarla para activar un ataque digital.

III. *Impacto biológico: En el caso que una persona mal intencionada tenga la habilidad de poder 'automatizar' una mutación en una determinada secuencia, podría crear un script para insertar la modificación en banco de datos de información genética o sistemas internos de laboratorios. En este escenario, esa modificación no tendría ningún impacto malicioso en el sistema pero podría desencadenar un problema funcional a nivel biológico, en el caso que esa secuencia sea sintetizada sin mecanismos de verificación adecuados.*

Los escenarios II y III se describen a efectos de ilustrar los potenciales efectos que podría tener esta metodología de ataque, en caso que la tecnología avance con mayor velocidad que la preocupación por proteger las secuencias de ADN. No obstante, es importante aclarar que son escenarios vistos desde un plano hipotético cuya demostración requiere la combinación de diferentes disciplinas que no serán abordadas en este texto.

Finalmente, el ingreso al sistema de la víctima ya sea a través de Ingeniería Social o la explotación de una vulnerabilidad no será cubierto en este trabajo, ya que se trataría de cualquier mecanismo tradicional de infección, como por ejemplo a través de un gusano en un dispositivo USB, descargado de Internet a través de un drive-by-download o un troyano adjunto a un correo electrónico, etc. Para más información acerca de cómo podría ser un escenario específico de explotación, se puede recurrir al Trabajo Final de Especialización sobre 'Análisis de vulnerabilidades en Java' [16] de este mismo autor; donde se desarrolla en detalle la explotación de una vulnerabilidad en Java.

1. Introducción

El científico J. Craig Venter, cuyo instituto JCVI es el creador de la secuencia sintética que utilizaremos en esta investigación, ya es conocido por estar a la vanguardia en los últimos tiempos en lo que respecta a avances genéticos. Hace poco más de diez años participó en la carrera por secuenciar el genoma humano y ahora busca crear bacterias "verdes" para la fabricación de combustible sustentable, entre otras cosas. Uno de sus hitos más importantes en dirección a este último objetivo fue en 2010 cuando anunció la creación de la primera célula sintética bacteriana capaz de autoreplicarse. En primer lugar se diseñó un nuevo código genético en un computador. Luego, se creó químicamente el genoma y se lo trasplantó a una célula receptora para que sea "booteado" y efectivamente fue replicado por dicha célula. Esto implica que la creación inicial del genoma depende prácticamente de forma exclusiva de un archivo generado en una computadora. Sumado a eso, para poder certificar que se estaba replicando el genoma sintético, los investigadores agregaron 'marcas de agua' – cifradas con una clave que ellos mismos desarrollaron – que contienen frases de personajes célebres de la ciencia así como también una URL y un correo electrónico para que todas las personas que lograran descifrar el mensaje pudieran contactarlos.

Por lo tanto, del punto de vista informático, esto implica que con el diseño y modificación de genomas sintéticos en computadoras, se torna más importante verificar la seguridad de estos sistemas. Sin embargo, actualmente, no parece ser un campo al que se estén dedicando muchos esfuerzos. Por eso sería interesante analizar los tipos de ataques que podrían llevarse a cabo para superar la barrera entre la seguridad informática y la bioinformática, así como también los mecanismos para contrarrestar potenciales amenazas.

1.1. Situación actual

Actualmente, existe poca información sobre ataques con amenazas digitales en archivos que podrían afectar al entorno bioinformático y las

incidencias que podría ocasionar comprometer un sistema que albergue secuencias genómicas. Una de ellas es el artículo de Sperl Thomas, en octubre de 2013 [3], donde trata la posible infección de un sistema que posee secuencias de ADN; como por ejemplo un laboratorio de microbiología. Sin embargo, más allá de eso, no parece haber publicada mucha información respecto a implementaciones o detalles técnicos de este escenario.

Finalmente, no se puede afirmar que investigadores, universidades o incluso el propio JCVI, no estén dedicando más esfuerzos en lo que respecta la seguridad de archivos que contienen secuencias de genomas. No obstante, debido a la baja probabilidad de incidencia, también es factible que no hayan grandes avances hasta que el escenario de compromiso sea más visible.

2. Definiciones básicas

Para poder comprender mejor qué implica en términos generales comprometer maliciosamente un archivo FASTA con información genómica, se hará una pequeña revisión acerca de algunos conceptos[1] de forma general.

2.1. Células

Una célula es el componente básico de la vida. Es el menor elemento vivo en un organismo. Posee muchas partes, entre ellas el núcleo que es donde se encuentra el 'comando central' de la célula y donde se almacena el ADN.

2.2. ADN

La sigla proviene de ácido desoxirribonucleico y es una molécula de gran tamaño que guarda y trasmite de forma hereditaria información para el funcionamiento de un organismo vivo [4].

2.3. Genomas

Se trata de la secuencia de ADN completa de un organismo que incluye todos sus genes (fragmentos del ADN). Se podría decir que el genoma es el material genético de un organismo que contiene la información necesaria para construir y mantenerlo [5]. El término fue acuñado [6] en 1920 por Hans Winkler, un profesor de botánica de la Universidad de Hamburgo, Alemania. La palabra está formada por el acrónimo entre gen y cromosoma.

[1] Se hará una explicación breve de algunos conceptos fundamentales para poder continuar con el artículo, sin embargo al final se encuentra un glosario para aclarar términos que no sean familiares.

2.4. Secuenciación

La secuenciación es una técnica que se utiliza para determinar la 'secuencia nucleótida' de un ADN, es decir la parte más fundamental de un gen o genoma que contiene las instrucciones necesarias para construir un organismo. Sin ella, no sería posible la completa comprensión de las funciones genéticas [7]. Determinar la secuencia comprende saber el orden exacto en el que se organizan los cuatro nucleótidos: adenina (A), guanina (G), citosina (C) y timina (T); en una cadena de ADN. La secuenciación es un proceso largo y requiere mucho trabajo de articulación dado que la tecnología actual solo permite secuenciar pequeñas porciones de un genoma, por lo cual es usualmente necesario dividirlo en partes menores y luego combinar cada una de ellas. Ese trabajo, de reunir las partes, puede llevar más tiempo que la secuenciación en si misma, ya que requiere que se analice minuciosamente que la secuencia esté libre de errores y completa. De hecho, se suele secuenciar el genoma varias veces por precaución. El Human Genome Project, por ejemplo, requirió [8] secuenciar 12 veces el genoma humano de más de tres mil millones de pares de bases[2].

El proceso de secuenciación no se describirá en su totalidad, dado que escapa a los propósitos de esta tesis, no obstante tanto el Instituto de Investigación del Genoma Humano [9] como el Centro de Secuenciación de la Universidad de Michigan [10] poseen una explicación muy detallada y fácil de comprender para aquellos que deseen profundizar al respecto.

2.5. Representación de secuencias genómicas

Existen múltiples formatos en los cuales se puede representar archivos que contienen secuencias genómicas. A continuación, se puede observar la estructura de la secuencia sintética completa de la célula

[2] Cabe aclarar que el ADN posee una forma representada por una doble hélice, ya que está formado por dos cadenas complementarias de nucleótidos que se unen, de modo que una A se complementa con una T y una C con una G. Cada unión es denominada un par de bases.

Mycoplasma mycoides[3] del instituto J. Craig Venter en formatos FASTA y XML respectivamente:

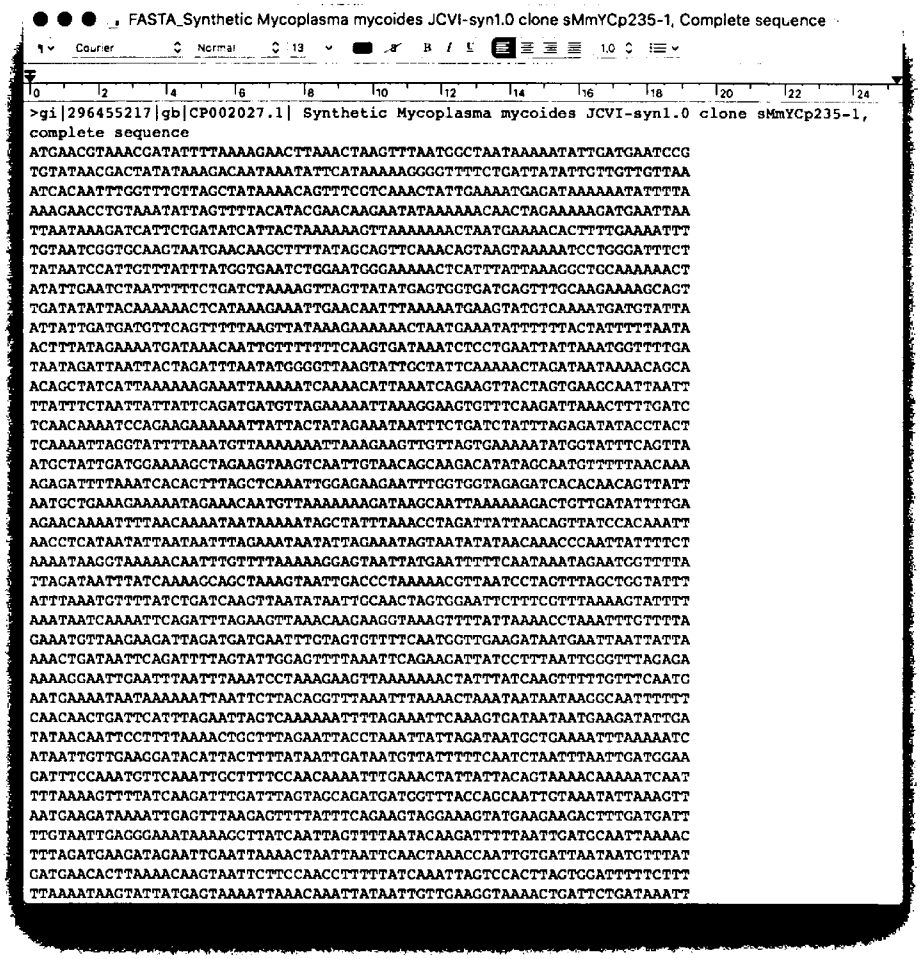


Imagen 1 – Ejemplo de secuencia de un genoma en formato FASTA

El encabezado en la primera línea, como se puede observar debajo en mayor detalle, usualmente comienza con el modificador '>'. Algunas veces también podría usarse ';'. En seguida después del símbolo de 'mayor', viene el identificador que es único para cada secuencia y posteriormente, de forma opcional, información adicional.

[3] Este microorganismo es un parásito que causa enfermedades pulmonares en ganados y cabras. Y es ampliamente utilizado debido a su tamaño relativamente pequeño.

>gi|296455217|gb|CP002027.1| Synthetic Mycoplasma mycoides JCVI-syn1.0 clone sMmYCp235-1, complete sequence

Esta secuencia se identifica como *gi|296455217|gb|CP002027.1|* y el nombre del genoma *Synthetic Mycoplasma mycoides* aparece consecutivamente como información adicional. Además se pueden observar *JCVI*, las iniciales del instituto J. Craig Venter, y la versión 1.0. El *clone sMmYCp235-1* hace referencia a que fue aislado y clonado de células de *Mycoplasma mycoides sMmYCp235*.

Imagen 2 – Ejemplo de secuencia de un genoma en formato XML

3. Caso de Estudio: Ejecución de código malicioso a través de secuencia genómica

A partir de este estudio, se procura analizar los límites y el impacto de una infección con *malware* a la secuencia de un genoma. Para poder ejemplificar mejor el escenario, se creó una PoC (del inglés *Proof of Concept*) que busca ilustrar en un entorno controlado cómo funcionaría la infección digital a una célula sintética representada por un archivo en una computadora. Con ese fin, se desarrollaron algoritmos en *Python* que se encargan de localizar los archivos FASTA[4] e inyectar el código malicioso o *payload* en la secuencia, así como también posteriormente localizar archivos genómicos infectados, identificar el texto que codifica el código malicioso, descifrarlo y ejecutar el *payload*.

Esto debe ocurrir a través de dos pasos. Primero, forzar que la víctima se descargue la secuencia modificada y segundo que se lance el script ya sea a través de la explotación de una vulnerabilidad o utilizando Ingeniería Social.

3.1. Objetivo

El objetivo es demostrar que la secuencia de un genoma cuyo contenido fue afectado por un archivo malicioso, puede comprometer un sistema en determinadas condiciones. De esta forma, se extiende la importancia de la integridad como uno de los pilares fundamentales de la seguridad de la información pues el impacto iría más allá de una secuencia errónea.

Además, dada la particularidad del escenario, un posible uso vinculado a esta amenaza podría ser el espionaje industrial [11]. En este caso, los afectados serían científicos, laboratorios, proveedores de síntesis de ADN, universidades, o cualquier otra institución que posea información de alto valor para otras empresas – o incluso gobiernos.

[4] FASTA es una de las extensiones más utilizadas en bioinformática para archivos con información genética pero existen más alternativas y el algoritmo se adapta fácilmente ajustando la condición de búsqueda en función de la extensión deseada.

3.2. Importancia de entender este ataque

Hay varias razones por las cuáles este ataque adquiere cierta relevancia frente a otras amenazas tradicionales. Algunas de ellas favorecen la ejecución de la amenaza mientras que otras limitan que este ataque pueda ser implementado fácilmente. Ambos enfoques se describen a continuación.

3.2.1. Factores que favorecen el vector de ataque

- Más secuencias, más repositorios

Un factor importante es el descenso vertiginoso que están actualmente atravesando los costos de secuenciación. Esto posee un fuerte impacto debido a que cada vez más instituciones, organizaciones, e incluso en algún momento personas privadas, podrían acceder a secuenciar ADN de forma más masiva. Esto conllevaría a una descentralización del almacenamiento que podría terminar restando preferencia a los 'repositorios oficiales' para ciertas secuencias. Si esto ocurre, se volverá más complejo determinar cuál es la fuente fiable de algunas secuencias, ya que los repositorios de menores recursos podrán ser un blanco más fácil para un cibercriminal y por ende, se pone en riesgo la fuente de confianza. En un caso así, le restaría a los investigadores resecuenciar el ADN como uno de los últimos recursos, para estar seguros que están lidiando con la secuencia no comprometida.

- Complejidad en la comprensión

La complejidad de este ataque también es una variable relevante para lograr su comprensión. En este caso, no necesariamente del punto de vista técnico del *malware*, porque la amenaza puede ser relativamente sencilla. La complejidad yace en la naturaleza interdisciplinaria de la temática, donde expertos en seguridad no poseen conocimientos en biología molecular o bioinformática ni estos últimos poseen conocimientos en seguridad. Sin mencionar que nuevos adeptos y entusiastas que recurran a manipular estos archivos, probablemente no posean ninguno de los dos.

- Genomas sintéticos

No solo las secuencias tradicionales podrían sufrir modificaciones maliciosas, sino también las secuencias de genomas sintéticos. Es decir, aquellos que fueron creados completamente en una computadora. Estos no solamente sufrirían los mismos inconvenientes citados anteriormente, ya que podrían o no estar en los prestigiosos repositorios mundiales, sino que además corren el riesgo de no 'existir' en la naturaleza de forma real; por lo cual no se pueden volver a adquirir fácilmente. A diferencia de un software que, en caso de encontrarse con una versión *troyanizada,* uno siempre puede contactar al fabricante para obtener el producto oficial; las secuencias de genomas sintéticos pueden ser más complejas de adquirir y determinar si se trata de la versión original.

- Amenaza inesperada

El escenario estudiado sigue siendo un caso de laboratorio. Por lo tanto, debido a que no es simple registrar ocurrencias, puede que haga a sus víctimas más vulnerables en caso de ser implementado en el mundo real. Como suele ocurrir con amenazas informáticas, las primeras víctimas *in-the-wild* suelen ser las más afectadas por amenazas que no fueron proactivamente detectadas, hasta que surge la detección reactiva.

- Detección

La detección de esta amenaza no solo a través de los scripts que inyectan y extraen la rutina maliciosa dentro de las secuencias de ADN sino también la detección de la rutina en si misma, resulta ser nada simple para soluciones de seguridad. Esto hace que la prevención sea una de las pocas herramientas para evitar este ataque.

- Persistencia

Todas aquellas secuencias que hayan sido modificadas, excepto que generen un problema funcional, podrían persistir en el sistema durante largos periodos o incluso indefinidamente hasta que se optimicen los sistemas de detección.

3.2.2. Factores limitantes

Por otro lado, existen algunos aspectos que hacen la ejecución de este ataque menos factible, como se expone seguidamente.

- Implementación no trivial

La propia complejidad interdisciplinaria puede ser uno de los grandes obstáculos para poder implementar esta amenaza en un escenario real. No obstante, sería naíf subestimar la capacidad de los atacantes cuando hemos observado amenazas modulares como Flamer [50] o considerablemente complejas como Stuxnet y Duqu.

- Escasa flexibilidad

La misma característica que le permite a una rutina maliciosa persistir en una secuencia genómica sin ser detectada, por otro lado, vuelve la amenaza poco dinámica. Esto se debe a que, en caso de poder identificar la cadena de caracteres extraña, sería fácil suprimirla y deshacerse de la amenaza. La complejidad radica en la identificación pero la desinfección es relativamente simple.

3.3. Posible escenario I: Propagar secuencia infectada

Diferentes escenarios podrían surgir donde se comprometieran secuencias de un genoma sintético con un *payload* malicioso. A continuación, se describe un escenario donde se puede implementar la prueba de concepto y su correspondiente escenario de ataque. A modo de simplificar se divide el caso en dos partes y se identifican los archivos claves *descifrador* y *payload* que serán nombrados como parte del desarrollo.

3.3.1. Fase 1: Adquisición de secuencia modificada

El primer paso consiste en la utilización de Ingeniería Social a un investigador, genetista, bioingeniero o cualquier persona que manipule secuencias de genomas, para que adquiera o descargue una secuencia comprometida con un *payload p*. A su vez, ese mismo usuario actuará como medio de propagación al compartir el genoma infectado con su círculo de confianza ya sea directamente o a través de un repositorio local.

3.3.2. Fase 2: Localización y ejecución

En segundo lugar, debería propagarse una herramienta[5] (preferentemente en Python) que poseería un nuevo script, que denominaremos descifrador *d*. Éste se encargará de parsear el genoma, decodificar *p* y ejecutar la acción maliciosa, en este caso abrir una URL. No obstante, *p* también podría intentar abrir una *shell* y ejecutar un comando, reportar a una *botnet* para establecer una comunicación, realizar espionaje industrial y demás.

Como mencionado en el prólogo, también se pueden realizar modificaciones a nivel biológico para forzar mutaciones genéticas que pudiesen comprometer el genoma a nivel funcional, lo que correspondería al segundo caso: 'digital – biológico'. Incluso, en caso de que se realice la síntesis química al genoma comprometido y pase exitosamente al mundo biológico, quizá podría volverse a secuenciar y la amenaza aparecería nuevamente de forma digital[6] como también mencionado en el prólogo bajo el tercer caso: 'biológico'.

3.4. Posible escenario II: Ataque dirigido

En este escenario la(s) víctima(s) se conoce(n) previamente. Acá se busca propagar un script inyector *i* que actuará sobre las secuencias que posea la víctima introduciendo el *payload p*. A diferencia del escenario I, este ataque requiere dos scripts, *i* y *d* para poder ser exitoso.

Una vez ejecutado, *i* busca localizar secuencias en el sistema e infectarlas, cual virus, con diferentes *payload*s maliciosos. Este escenario podría ser útil para espionaje industrial, ataques entre laboratorios, organizaciones e incluso países. Posteriormente, se implementaría la fase 2 de igual forma que en el escenario anterior para, a través de la herramienta descifradora *d*, extraer y lanzar *p* en el sistema afectado.

---

[5] Esta herramienta podría realizar algo útil en conjunto con el script o simplemente actuar como un troyano y ejecutar el script en segundo plano.

[6] Es necesario comprobarlo químicamente dada la complejidad en el proceso de síntesis debido a la modificación de la estructura, no obstante excede el objetivo de esta tesis.

Se trata de un ataque dirigido porque a diferencia del escenario anterior, no se trata de propagar una secuencia comprometida, sino de comprometer secuencias que se encuentren en el sistema de una determinada víctima. Esto también podría ser útil, por ejemplo, cuando atacantes quisieran afectar laboratorios que estén investigando la creación de secuencias sintéticas que aún no han sido publicadas.

3.5. Algoritmos implementados

Los algoritmos utilizados para realizar esta prueba de concepto están clasificados de acuerdo a su propósito y fueron desarrollados en Python. En totalidad se utilizan siete scripts, sin embargo algunos realizan funciones más específicas como ordenar y filtrar los datos por lo cual se omiten en este apartado.

3.5.1. Codificar payload e infectar genomas

Se desarrolló un algoritmo que toma un código HTML, o cualquier input deseado, y lo codifica en codones (combinación de A, C, T, G de longitud tres que usaremos para codificar un código malicioso) utilizando una clave asociada.

```
# By Raphael Labaca Castro on 22.10.14 updated 07.11.2015
# Encrypts code to codons using JCVI-Key (with additions)

def encryptCodons():

    file = open('input')
    code = file.read()
    code = list(code.upper())
    file.close()

    for index, item in enumerate(code):
#A
        if (item== '?'):
            code[index] = 'AAA'
        elif (item== 'L'):
            code[index] = 'AAC'
        elif (item== '3'):
            code[index] = 'AAT'
        elif (item== 'P'):
            code[index] = 'ACA'
        elif (item== '-'):
            code[index] = 'ACC'
        elif (item== '2'):
```

```
            code[index] = 'ACT'
        elif (item== '4'):
            code[index] = 'AGA'
        elif (item== '>'):
            code[index] = 'AGC'
        elif (item== 'B'):
            code[index] = 'AGT'
        elif (item== ' '):
            code[index] = 'ATA'
        elif (item== 'D'):
            code[index] = 'ATT'
#C
        elif (item== 'M'):
            code[index] = 'CAA'
        elif (item== '/'):
            code[index] = 'CAC'
        elif (item== ':'):
            code[index] = 'CAG'
        elif (item== 'Y'):
            code[index] = 'CAT'
        elif (item== '='):
            code[index] = 'CCA'
        elif (item== ';'):
            code[index] = 'CCC'
        elif (item== '.'):
            code[index] = 'CGA'
        elif (item== '8'):
            code[index] = 'CGC'
        elif (item== '<'):
            code[index] = 'CGG'
        elif (item== 'O'):
            code[index] = 'CGT'
        elif (item== 'R'):
            code[index] = 'CTA'
        elif (item== 'I'):
            code[index] = 'CTG'
        elif (item== '1'):
            code[index] = 'CTT'
#G
        elif (item== "'"):
            code[index] = 'GAA'
        elif (item== '!'):
            code[index] = 'GAG'
        elif (item== 'K'):
            code[index] = 'GCA'
        elif (item== '6'):
            code[index] = 'GCC'
        elif (item== '5'):
            code[index] = 'GCG'
        elif (item== 'S'):
            code[index] = 'GCT'
        elif (item== '"'):
```

```
            code[index] = 'GGA'
        elif (item== 'F'):
            code[index] = 'GGC'
        elif (item== 'X'):
            code[index] = 'GGT'
        elif (item== '9'):
            code[index] = 'GTA'
        elif (item== 'W'):
            code[index] = 'GTC'
        elif (item== ','):
            code[index] = 'GTG'
        elif (item== 'J'):
            code[index] = 'GTT'
#T
        elif (item== 'E'):
            code[index] = 'TAA'
        elif (item== 'G'):
            code[index] = 'TAC'
        elif (item== 'A'):
            code[index] = 'TAG'
        elif (item== '7'):
            code[index] = 'TAT'
        elif (item== 'H'):
            code[index] = 'TCA'
        elif (item== 'U'):
            code[index] = 'TCC'
        elif (item== '@'):
            code[index] = 'TCG'
        elif (item== '0'):
            code[index] = 'TCT'
        elif (item== 'T'):
            code[index] = 'TGA'
        elif (item== 'N'):
            code[index] = 'TGC'
        elif (item== 'Z'):
            code[index] = 'TGG'
        elif (item== 'Q'):
            code[index] = 'TTA'
        elif (item== 'V'):
            code[index] = 'TTG'
        elif (item== 'C'):
            code[index] = 'TTT'

    code = ''.join(code)
```

Algoritmo 1: Cifrar payload en codones

Luego, se dispara otra colección de scripts que procura archivos con información genómica (ej.: FASTA) e inyecta la porción de código que acaba de ser codificada en codones basada en el input suministrado.

```
# By Raphael Labaca Castro on 17.05.15 - updated 07.11.15
# This script performs the following steps:
# 1. Searches for original FASTA file
# 2. Encrypts a payload (HTML, etc.) with JCVI-key
# 3. Injects encrypted FASTA into original FASTA file

from __future__ import print_function
from threading import Thread

def injectPayload():

    #locate and return a FASTA file
    def locateFiles():
        for root, dirs, files in os.walk('Path'):
            for file in files:
                if file.lower().endswith('.extension'):
                    theFile = (root + '/' + file)
                    return theFile #Only injecting 1st
    theFile = locateFiles()

    #encrypts payload and returns FASTA
    encryptedPayload = encryptCodons.encryptCodons()

    #injects code encrypted into FASTA (at the beginning)
    content=open(theFile,'r').read()
    content=encryptedPayload+content
    fp.write(content)
    fp.close()
    print (theFile)

injectPayload()
```

Algoritmo 2: Inyectar payload en secuencia

3.5.2. Localizar y descifrar payload en genoma infectado

Por otra parte, se lanza el *parser – o descifrador* – para traducir los codones y extraer el código. Los scripts toman el *payload* malicioso y lo decodifican con la clave dada. Para este ejemplo, se usó la misma que el JCVI por convención pero sería lógica la definición de claves nuevas. También se podría agregar un paso adicional y ofuscar el *script* para hacer más compleja su detección.

```
# By Raphael Labaca Castro on 19.10.14 – updated 07.06.15

from __future__ import print_function
from threading import Thread

class PercentgeProgress(Thread):

    def locateFiles(self):
        for root, dirs, files in os.walk('Path'):
            for file in files:
                if file.lower().endswith('.extension):
                    theFile = (root + '/' + file)
                    my_dna = readDNA.readDNA(theFile)
                    parseCode.parseCode(my_dna)
                    return theFile

# Text while waiting

    def __init__(self, text=None):
        self.chars = ['/', '-', '\\', '|']
        self.index = 0
        self.text = text if text else 'Working...'
        self.stopping = False
        Thread.__init__(self)

    def run(self):
        while not self.stopping:
            if self.index >= len(self.chars): self.index
= 0
            sys.stdout.write('\r{0}
{1}'.format(self.text, self.chars[self.index]))
            sys.stdout.flush()
            self.index += 1
            time.sleep(0.1)

    def stop(self):
        self.stopping = True

# Call functions

task = PercentgeProgress(text='Searching..')
task.start()
task.locateFiles()
task.stop()
```

Algoritmo 3: Localizar payload en una secuencia, descifrar y ejecutar

El último paso es ejecutar el *payload* extraído. En este caso, como mencionado anteriormente, se trata de un simple HTML que llama una URL en un navegador, lo cual es suficiente para demostrar una potencial acción maliciosa, como se puede observar debajo.

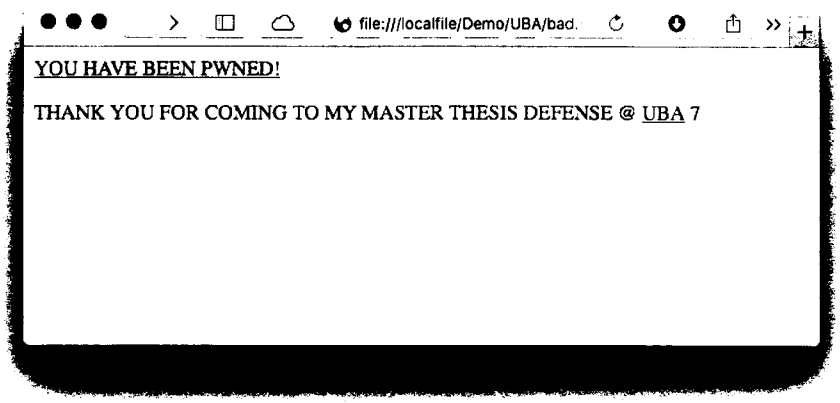


Imagen 3: Captura de ejecución de payload extraído de un genoma infectado

3.6. Detección y prevención del ataque

Una buena práctica de la industria, cuando se fabrican genomas o se modifican genéticamente organismos, es que los investigadores tienen la obligación de secuenciar el trozo de ADN que obtuvieron físicamente para comprobar que lo que se fabricó *in vitro* coincide con lo publicado. No obstante, esta verificación sería menos confiable en caso de que se comprometan secuencias que todavía no estén publicadas o fueron modificadas antes del proceso de síntesis.

Otro inconveniente sería que los investigadores no sean capaces de identificar el ataque, ya que podrían terminar presentando una secuencia comprometida a repositorios tradicionales. Este escenario posee el agravante que, en caso de ser exitoso, el genoma luego podría ser propagado más fácilmente. Asimismo, estas secuencias también podrían estar disponibles en repositorios pertenecientes a instituciones o universidades con menos recursos – por ende quizá menos protegidos y controlados – y también terminar siendo utilizados significativamente por la comunidad científica y/o interesados.

3.6.2. Detección de los scripts

Los algoritmos que se dedican respectivamente a inyectar (i) y descifrar (d) una porción de código malicioso o *payload* dentro de un archivo FASTA y finalmente ejecutar una URL; no necesariamente son detectados [12][13] como se puede observar en las siguientes imágenes:

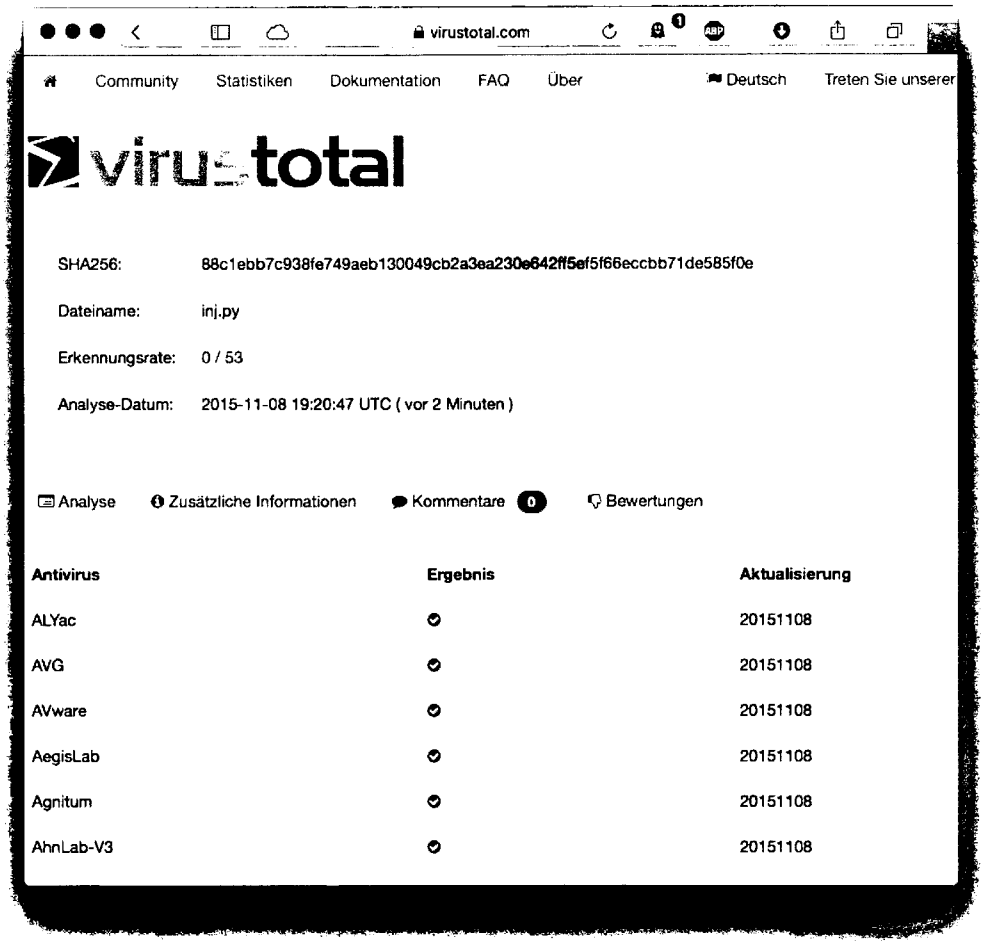


Imagen 4: Detección sobre el script (a) que inyecta el payload codificado

Por lo tanto, la seguridad en muchos casos recae sobre la detección del *payload*. Asimismo, una vez extraída y descifrada esa porción maliciosa, en caso de ser advertido por una solución de seguridad ya sea a través de su heurística o de su base de firmas; se eliminaría únicamente el *payload* sin afectar los archivos que contienen las secuencias, dado que en el genoma esa información se encuentra codificada.

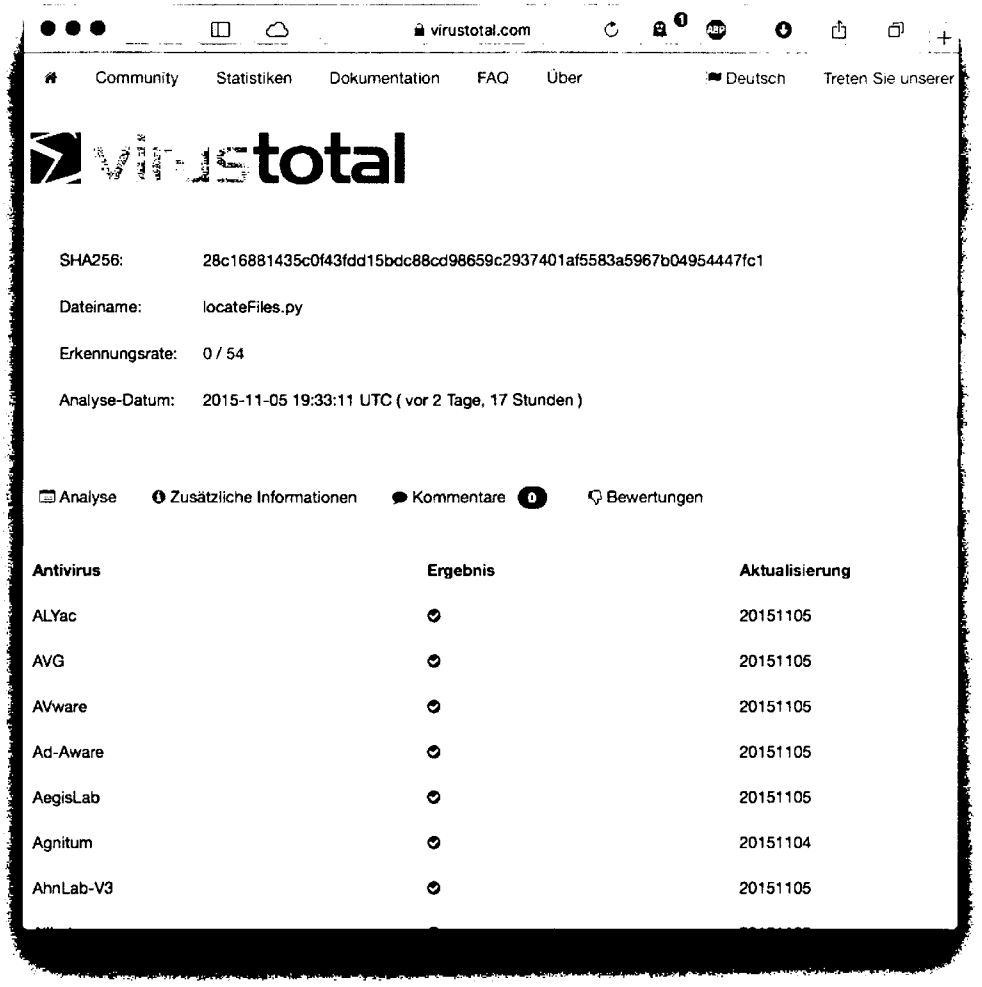


Imagen 5: Detección de script (b) que ejecuta la URL que contiene el payload

Asimismo, también existen técnicas para intentar minimizar la detección del *payload* como embeber el código de Metasploit (escrito en C) en Python [14] o utilizar herramientas [15] que generan exploits para Metasploit. Sin embargo, todo depende de la seguridad del entorno y es necesario explotar alguna vulnerabilidad [16] en el sistema, lo cual hace que sea más proclive la detección.

Finalmente, el *payload* necesitaría ejecutarse y evadir la heurística de una solución de seguridad solo una vez exitosamente. Luego, una vez identificado, por ejemplo a través de una firma, solo debería descartarse esa porción en la secuencia y comenzar a utilizar otros de los inyectados. Hasta sería imaginable, en un escenario más complejo, que el script que localiza la

secuencia (b) descifre, ejecute y reemplace los codones con un nuevo *payload* – o incluso que lo elimine para borrar rastros.

3.7. Comportamiento de las base de datos de secuencias

De acuerdo al avance tecnológico actual, sobre todo debido a la nueva generación de dispositivos a partir de 2008, se puede observar cómo los costos de secuenciación por genoma se redujeron drásticamente al nivel de superar exponencialmente la Ley de Moore, utilizada como referencia:

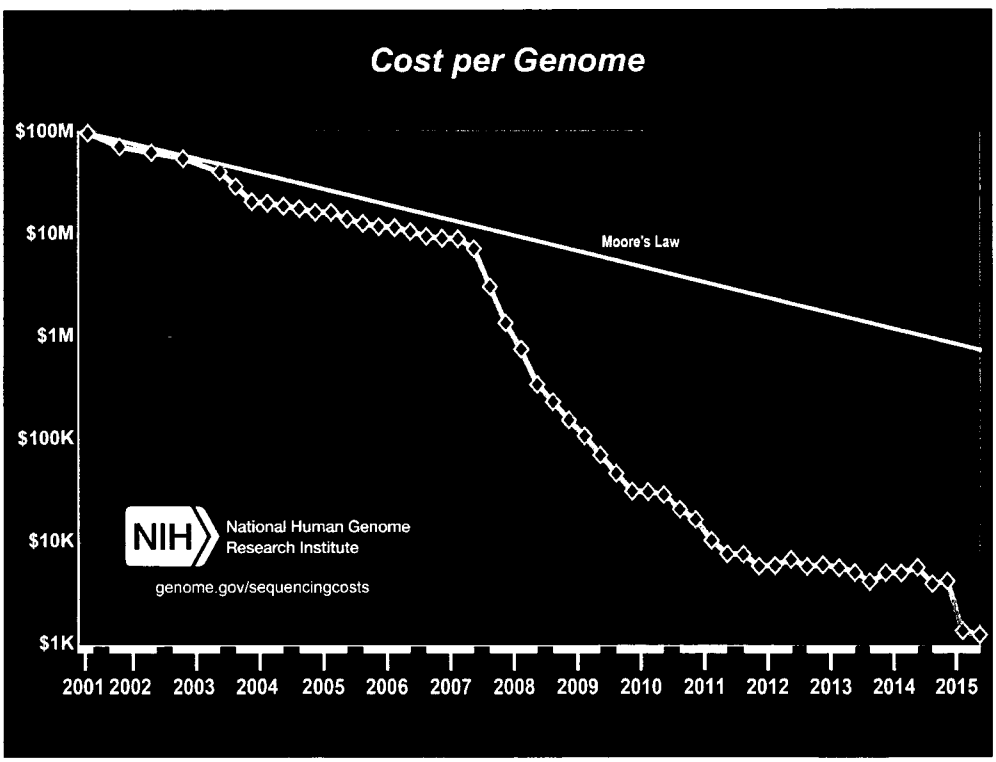


Imagen 6: Costo por genoma. Fuente: NIH – genome.gov/sequencingcostsdata

Basada en esta tendencia, es probable que la secuenciación sea cada vez mayor ya que aumenta progresivamente el acceso a estas tecnologías. Por lo cual, una vez que se puedan secuenciar genomas masivamente, los actuales repositorios van a recibir más fuentes de genomas para sus bases de datos y eventualmente más repositorios serán creados. Ante esa eventual masividad es posible que ocurran, entre otras, las siguientes situaciones:

1. Que los repositorios tradicionales acepten nuevas secuencias
Entonces surge la duda de cómo van a controlar que ese genoma viene "no contaminado" con un *malware*. Por ejemplo, quizás utilizando listas blancas de fuentes o algoritmos automatizados[7].

2. Que algunos no acepten todas las nuevas secuencias
En caso que algunos bancos de secuencias se vuelvan muy estrictos con el control, es probable que se dé aún más espacio a repositorios alternativos donde la gente va a querer compartir el genoma secuenciado. Algo similar a una Wiki de genomas, o hasta quién sabe, una red social.

De hecho, ambas opciones no son mutuamente excluyentes y en el caso que los controles sean ineficaces, se podría propagar una secuencia infectada tanto en repositorios tradicionales como en los menos conocidos. Resta luego atacar a quiénes descargaron las secuencias afectadas para ejecutar el *payload*, no obstante aunque ese objetivo no se logre; el punto es que ya se propagó una secuencia cuya integridad fue comprometida.

Está claro que para que este escenario ocurra aún quedan interrogantes, como por ejemplo, quiénes específicamente se descargarían la secuencia y para qué lo harían. Porque no solo está el usuario tradicional, sino también los profesionales que podrían buscar secuencias en repositorios alternativos y que, por distintos motivos, no se encuentren en fuentes tradicionales.

De hecho, una posible aplicación podría estar vinculada con la autenticación de personas. Un estudio realizado en 2015 [17] entre clientes de bancos británicos, reporta que 24% de las personas encuestadas

[7] A pesar que algunos repositorios como el NCBI, ya poseen controles de integridad, estos varían en función del tipo de secuencia. Por lo tanto, se podría inferir que en algunas secuencias de menor relevancia, en caso de infectadas, el código malicioso no necesariamente sería detectado. Asimismo, existen bases de datos 'curadas' que visan por la integridad de los genomas, no obstante esta opción no es válida para todas las secuencias.

estarían dispuestas a compartir información personal con la entidad financiera, incluso su ADN, para promover mejores mecanismos de autenticación biométrica a sus cuentas. Quizá estas nuevas plataformas puedan ser el nexo entre estas personas y las instituciones con las que decidan compartir su información en el futuro.

3.8. Complejidad para Identificar la raíz del ataque

Un punto a tener en cuenta, dadas las circunstancias, es que la auditoria para saber cómo se inició la infección puede requerir tiempo y la determinación que el sistema está limpio podría ser compleja y costosa en muchos casos. Esta situación se agravaría si los sistemas afectados poseen muchos genomas. Además, de no encontrarse el script *inyector* no se podría saber precisamente cuáles son los cambios efectuados en las secuencias. Por lo tanto, todos esos archivos FASTA deben ser debidamente gestionados ya que son un conjunto de '*strings* ilegibles' que forman un mar de oportunidades para los atacantes, ya que también se podría agregar información camuflada o incluso utilizarlo como *covert channel*.

3.9. Localización del *payload* en el genoma sintético

Cuando se crea un genoma sintético, se hace necesario poder identificarlo para asegurar que se está manipulando realmente la secuencia artificialmente creada y no un genoma nativo. Dicha 'identificación' incluía, en el caso de *Mycoplasma mycoides JCVI-syn1.0*, los nombres de los investigadores y el laboratorio que la creó junto a algunas citas célebres de científicos famosos. A estos 'identificadores' se los denomina 'marcas de agua' (*watermarks* en inglés).

Al momento de inyectar un *payload* hay que determinar la localización donde se insertarán los codones con el código malicioso. Una alternativa sería utilizar marcas de agua ya existentes, ya que evitaría modificar una zona desconocida del genoma. De lo contrario, podría comprometer la función biológica de algún gen y provocar que el organismo no prospere en caso de ser sintetizado. Sin embargo, es probable que la longitud del *payload* deba ser ajustada a una marca de agua dada. Otra alternativa sería

buscar secciones nuevas, no obstante se requiere un nivel de complejidad mucho mayor para identificar[8], a nivel biológico, una localidad en el genoma donde se podrían establecer marcas de agua sin causar problemas.

Posteriormente, una vez insertado el *payload*, su identificación podría representar un problema dado que la atención de parte de los investigadores probablemente esté en su impacto biológico. De no existir consecuencias de ese tipo, una marca de agua que contiene un *malware* podría pasar inadvertida. Una técnica para identificarlas sería comparar los genomas con un repositorio principal. El problema es cómo asegurar que dicha base no está comprometida. Esa situación la denominaremos la 'problemática del diccionario' cuando un término no es preciso en el 'libro de confianza' o en este caso, la incertidumbre que un genoma no sea íntegro en el repositorio principal. Tanto es así que aún en el supuesto caso de superar la problemática del diccionario e identificar una marca de agua maliciosa, habría que encontrar la clave[9] para descifrar la función del *payload* en todos los genomas; lo cual además, podría ser difícil de llevar a cabo de forma exhaustiva dentro de una biblioteca con centenas – o miles – de secuencias.

En caso de existir respaldos se podría aplicar un *diff*[10] para localizar las *watermarks*, sino en última instancia se puede volver a secuenciar; sin embargo puede que sea costoso y por eso también se destaca el impacto económico que ocasionaría una amenaza en la integridad de las secuencias.

[8] La identificación de estas secciones es una temática que requiere un conocimiento detallado del punto de vista biológico, cuyo estudio no se abarcará en este trabajo.

[9] Por ejemplo, esto se puede lograr a través de análisis de frecuencias, como fue anteriormente mostrado, en caso de que se conozca una porción del texto claro o se pueda inferir la misma a través de un encabezado o cadena de archivo esperado.

[10] Clásica herramienta o antigua aplicación de Unix que compara y devuelve las diferencias entre dos archivos dados.

4. Desafío: ¿dónde insertar el payload?

Una de las mayores problemáticas para poder llevar a cabo la inyección de *malware* en genomas sintéticos es poder determinar dónde se ingresa esa nueva 'marca de agua' o en este caso, el código malicioso. Dado que un código extraño en la estructura del genoma podría interferir con su estructura y volverlo obsoleto comprometiendo su funcionamiento o el de los fenotipos asociados (características del organismo). Una alternativa sería, dado este contexto, trabajar directamente sobre la secuencia sintética de *Mycoplasma mycoides JCVI-syn1.0* localizando y reemplazando las marcas de agua ya existentes dado que representa una mayor probabilidad de estar sobrescribiendo una parte 'segura' del genoma.

La otra forma, quizá, sería localizar secuencias redundantes para poder reemplazarlas. Unos meses después de la publicación del artículo de J. Craig Venter, en octubre de 2010 se publicó en la revista Science una investigación acerca de la variación en el número de copias del gen humano[11] donde enuncian que aparentemente la mayoría de los genes posee dos copias, sin embargo se identificaron 56 familias de genes que poseen entre 5 y 368 copias. Por lo cual, todavía no se podía determinar con precisión qué tan redundante es cada una de las repeticiones de los genes.

A pesar de existir varias publicaciones al respecto, no es una tarea simple localizar las porciones del genoma que podrían ser reemplazadas. En el artículo de Venter, por ejemplo, se aclara que desarrollaron las marcas de agua para reemplazar regiones experimentalmente demostradas (Watermark 1 con 1246 pares de base y Watermark 2 con 1081 bp) y predichas (Watermark 3 con 1109 bp y Watermark 4 con 1222 bp) para no comprometer el funcionamiento de la célula. No obstante, no aparecen mayores detalles de cómo se predijeron o determinaron esas regiones en la célula bacteriana.

[11] Vale aclarar que el genoma modificado por JCVI y analizado en este artículo corresponde a una célula bacteriana y no al genoma humano, no obstante estos estudios mencionados podrían servir de base para analizar la redundancia de las secuencias en diferentes tipos de genomas.

Adicionalmente, existen estudios sobre cómo liberar codones en la secuencia [29] comprobando que se pueden realizar cambios en esos codones a lo largo de todo el genoma. El objetivo es eliminar la redundancia en el código genético para reintroducir modificaciones que codificarían aminoácidos artificiales. En ese caso, se reemplazó con TAA el codón TAG en un genoma de Escherichia Coli – también conocido como E. Coli [30] – dado que ambos, junto a TGA, son tripletes nucleótidos denominados *stop*, que marcan el fin de la etapa de traducción genética (la más importante en la síntesis de proteínas). Luego de un complejo proceso químico, el genoma completo queda libre del codón TAG y se lo puede utilizar para codificar nuevos aminoácidos artificiales. Esto permitiría, por ejemplo, realizar cambios sobre genomas naturales sin tener que crear genomas sintéticos de cero. Sin embargo liberar un triplete de nucleótidos puede tener un impacto biológico importante pero no resulta muy útil para el propósito aquí abordado, dado que se necesitan varios tripletes (codones) para 'codificar' un *malware* y eliminar la función biológica en este caso es irrelevante. Lo que se necesitan son 'espacios seguros' para insertar diferentes codones que representen – o codifiquen – el código malicioso sin afectar el funcionamiento biológico.

4.1. *Junk* DNA: ¿realmente redundante?

De acuerdo al profesor David Stern, del departamento de Biología Evolutiva de la Universidad de Princeton, el denominado "*junk* DNA" [31], aquella porción del genoma que no contiene genes que codifican proteínas y que constituiría aproximadamente un 98 por ciento del genoma humano, es fundamental para transformar información codificada de los genes en productos útiles para el organismo. Para eso, se dedicaron a analizar el comportamiento de regiones denominadas *enhancers*, que son importantes no solo en el proceso de síntesis de proteínas sino también en diferenciar que un organismo sea una mosca o un ser humano.

Stern apoya la teoría que los genes se tienen que analizar desde una perspectiva evolutiva y por lo tanto su propósito es producir organismos saludables en ambientes cambiantes, por lo cual una gran parte de estos

genes evolucionó para tratar las contingencias que los organismos podrán experimentar en el mundo real.

En el artículo se expone la existencia de *enhancers* secundarios que estarían más alejados de los genes que deberían regular y por lo tanto, confirmarían la existencia de genes importantes para el desarrollo del organismo en zonas inesperadas del genoma. El estudio fue conducido analizando genes de ratones y moscas de la fruta, no obstante el equipo considera que podría ocurrir de forma similar en otro tipo de genomas. Por lo cual, sería una alternativa limitada a tener en cuenta al momento de reemplazar porciones de *junk* DNA con codones que 'codifican' un algoritmo malicioso, debido que a medida que avanza el estudio del ADN es más complejo reescribir áreas no-funcionales del genoma, ya que aparentemente es más un problema de comprensión que de funcionalidad.

Posteriormente en septiembre de 2012, el Dr. Ewan Birney, uno de los principales investigadores del proyecto internacional ENCODE (Acrónimo de Enciclopedia de los Elementos de ADN, que busca identificar todos los elementos funcionales en la secuencia completa del genoma humano) junto a su equipo publicaron [32] estudios donde afirman que efectivamente no todo el 98% restante del genoma humano es 'chatarra[12]', ya que aproximadamente un quinto se dedica a regular el 2% que codifica proteínas. Además, agregan que hasta un 80% de la secuencia completa de ADN debería tener alguna función bioquímica, por lo cual se derriba el mito que un 98% constituiría una parte no funcional del genoma.

[12] Traducción libre del concepto '*Junk* DNA' utilizado para hacer referencia a la parte del genoma que no codifica proteínas.

5. Privacidad en las secuencias públicas

No solamente la integridad de las secuencias sería un problema del punto de vista de la seguridad de la información. La privacidad también ha sido tópico de investigación en la comunidad científica.

Existen muchas bases de datos públicas en Internet con información genética de voluntarios. A pesar que las secuencias almacenadas se mantienen anónimas, hay mucha información disponible, que en caso de ser analizada, proporcionarían una especie de 'metadatos' del ADN. Esto fue asunto de una investigación realizada por Yaniv Erlich y el equipo del Whitehead Institute for Biomedical Research, MIT y la Universidad de Harvard [46][13].

Se trata de que, en la herencia genética, los hombres reciben el haplotipo de cromosoma *Y*[14] de parte de sus padres y éstos, a su vez, también de sus padres y así sucesivamente. Asimismo, en la mayoría de las sociedades actuales el apellido familiar se hereda del padre por lo tanto, existe un vínculo entre el cromosoma *Y* y el apellido familiar. Sumado a eso, en Internet existen actualmente varias bases de datos y sitios web que, en conjunto, contienen miles – o cientos de miles – de registros con apellido y cromosoma *Y*. Básicamente, se trata de tomar un enfoque cuantitativo de qué tan probable es la inferencia del apellido entre las bases de datos genéticas y la información pública que se encuentra en Internet, para poder identificar unívocamente[15] las personas que participaron de los proyectos públicos y que poseen su información 'anónima' en las bases de datos. Sitios como ysearch.org poseen la capacidad de realizar búsquedas a través del apellido, haplotipos e incluso buscar coincidencias genéticas. Para eso, es necesario crear un usuario y dar de alta información personal como

[13] La fuente Science Magazine requiere registro gratuito para poder acceder al artículo completo con fines académicos.

[14] El Cromosoma Y no se encuentra en pares, diferente a los demás cromosomas. Por lo cual, cada descendiente macho recibe en gran parte el mismo que su padre.

[15] Como resultado se obtienen grupos de 12 personas, lo cual permite luego una búsqueda exhaustiva.

nombre, apellido, país y ciudad de origen, así como también ubicación geográfica (parámetros de latitud y longitud), procedencia (en caso de migrar o emigrar) y las marcas genéticas. Sin olvidar, además, que toda esta información se transmite en texto plano sin ningún tipo de protocolo de cifrado, probablemente dada su índole pública. El impacto de esta investigación fue tal que el Instituto Nacional de Ciencias Médicas de Estados Unidos [47] decidió ocultar la edad de los participantes del acceso público para no facilitar el cruce de la información.

No obstante, es importante resaltar que en la mayoría de los casos los participantes aceptaron firmar un consentimiento que no se tomarían medidas para prevenir la eventual identificación del genoma; de modo que brechas en la privacidad podrían estar contempladas en el estudio.

A medida que avanzan estas técnicas y se hace más simple formar parte de estas bases de datos genéticas, aumenta proporcionalmente el riesgo de perder la privacidad sobre esa información. Quizá en cuestión de poco tiempo, una persona puede dejar una huella en un vaso o en un cigarrillo en un bar y ser rastreada en Internet hasta dar con su ADN. Claro que aún faltan mejorar los algoritmos y las búsquedas pero, para cuando eso ocurra, es probable que la información esté disponible. Sumado a eso, están aquellos pacientes que padecen algún tipo de enfermedad, ejemplo un cáncer, y tienen su información secuenciada que será posteriormente utilizada con fines académicos y de investigación. Algunas instituciones, como la Universidad de Heidelberg [48], en Alemania, mantienen la información cifrada y solo proporcionan datos anónimos a los investigadores; a diferencia de estos sitios que solicitan la información personal junto con las marcas genéticas.

De hecho, existen muchos vacíos legales respecto a esta temática. Por ejemplo, no está definido formalmente a quién le pertenece la información si al paciente o al científico (¿o ambos?). La relativa novedad de la temática ha llevado a no pensar aún extensivamente en este tipo de inconvenientes. Es probable que las organizaciones internacionales y los países donde esto comienza a ocurrir, todavía no vean la privacidad de la información genética realmente como un problema. De hecho, la privacidad en general todavía no es vista como un inconveniente para muchos, por lo

cual será un desafío pedir que se considere la privacidad genética – al menos hasta que se valore adecuadamente la privacidad en Internet en aspectos muchos más generales.

Otro punto importante, respecto a la disponibilidad de toda información pública, es el hecho que se comience a usar con fines económicos o, evitando fomentar la paranoia, para sacar ventajas comerciales a la hora de realizar seguros médicos, laborales y actividades similares.

6. Futuro: ¿Amenaza biológica o digital?

Anteriormente, analizamos la prueba de concepto para el caso cuyo impacto fue clasificado en el prólogo como 'digital'. A continuación, describiremos el segundo caso, de impacto 'digital – biológico', como un potencial ataque a ser investigado y desarrollado en el futuro.

La situación en la cual una infección digital puede traspasar la barrera biológica e infectar un genoma artificial es indudablemente una temática atractiva, sobre todo luego de la creación de la primera célula artificial capaz de autoreplicarse. Esto nos lleva a repensar nuestro actual modelo de infección de las amenazas informáticas y plantearnos un cambio de paradigma en el impacto que puede tener un código malicioso sobre un 'organismo vivo'. Para eso, la infección digital en un ambiente biológico se daría a través de la inyección de una cadena de codones a un archivo que contiene la secuencia de un genoma.

Posteriormente, el archivo infectado debería pasar por un proceso de síntesis química y generar un genoma sintético que será posteriormente trasplantado a una célula donante (cuyo núcleo fue extirpado y está a la espera de uno nuevo que le trasmita las funcionalidades principales) para autoreplicarse. Esto significa que ahora la célula está 'viva'[16] y funciona bajo el comando de un genoma que fue creado en una computadora pero está infectado con un código malicioso informático.

Sin embargo, no todo es tan simple ya que algunos aspectos deben ser tenidos en cuenta para que el procedimiento funcione. Como por ejemplo, hay que verificar la posibilidad de insertar el código malicioso en una sección del genoma que no tuviera un impacto en la codificación de proteínas, entre otras cosas, y que pueda ser reescrita sin comprometer funcionalidades biológicas.

Si bien en este trabajo se usó el genoma secuenciado de la bacteria *Mycoplasma mycoides JCVI-syn1.0,* dado que fue el primero en su tipo,

[16] Pese a que es difícil determinar con precisión qué es la vida, la comunidad científica adhiere al hecho que las bacterias son organismos vivos, a diferencia de los virus que, entre otras cosas, no poseen una estructura celular y necesitan una célula huésped para reproducirse.

teóricamente, se podría tomar cualquier tipo de genoma secuenciado para realizar la inyección del código malicioso. Como por ejemplo la mosca común, que en octubre de 2014 tuvo su genoma completamente secuenciado por un grupo de científicos [18] para entender como algunos de sus genes brindan habilidades especiales a su sistema inmunológico y demás. Sin embargo en la práctica, ese genoma secuenciado tiene 691 millones de pares de bases lo cual eleva radicalmente la complejidad del proceso de síntesis comparado con el genoma de 1.08 millones de pares del *Mycoplasma mycoides* sintetizado en 2010.

Por lo tanto es de esperarse que un genoma sintético con capacidad de autoreplicarse y cuyo tamaño es mucho menor sea más atractivo para el inicio de la nueva era de infección digital – biológica.

6.1. Propagación biológica de una amenaza digital

En el hipotético caso que el trozo de ADN sintetizado que acabamos de mencionar, haya sido exitosamente modificado, entonces el código malicioso formará parte de una célula sintética capaz de replicarse de forma autónoma en el mundo biológico. El *malware* podría incluso ser 'propagado' de forma biológica, dado que las bacterias contienen consigo todo el equipamiento necesario para la reproducción y lo pueden hacer de forma asexuada, es decir, sin la necesidad de un segundo progenitor. Además el código malicioso no afectaría la célula portadora donde se aloja, sino que la utilizaría para mantenerse 'vivo' hasta que su genoma sea secuenciado en un laboratorio y vuelva a estar presente de forma digital para activarse en una computadora o dispositivo. No obstante, la localización de este código es clave para que la propagación biológica sea exitosa. A grandes rasgos existirían tres escenarios donde una cadena maliciosa podría ser insertada:

1. Zona irrelevante: el código malicioso entra en una zona poco importante, es probable que no cause ningún impacto.
2. Zona de un gen: si entra en una secuencia de un gen y produce una mutación, se generan dos opciones:
    a. La mutación es letal, entonces puede desparecer de la

naturaleza sin propagarse.

b. La mutación es beneficiosa o neutro, entonces puede que la porción agregada continúe su propagación.

3. Zona reguladora: el tercer escenario sería que ingrese a una zona denominada 'reguladora'. En este caso, podría alterar algún gen como en el segundo escenario o podría no hacer nada como en el primero.

Por lo tanto, en el caso que no se produzca una mutación letal, el *malware* y la célula sintética portadora formarían lo que denominaremos "comensalismo cibernético"[17], haciendo referencia al tipo de simbiosis donde un interviniente (en este caso el *malware*) obtiene un beneficio, mientras que el otro (la célula sintética bacteriana) no se perjudica ni se beneficia.

En un futuro quizá se puedan editar genes que se extraen del cuerpo humano como nuevos tratamientos terapéuticos. El Instituto de Investigación Biomédica de Novartis [19] ya conduce una investigación junto a la Universidad de Pennsylvania en donde buscan realizar modificaciones de células humanas ex-vivo, es decir fuera del cuerpo. En caso de que se realice exitosamente, modificaciones in-vivo también podrían hacerse realidad. De ser así, queda por analizar la factibilidad de que las modificaciones a la célula contengan subcadenas del ADN con un código malicioso. Y de un día ser viable, cuál podría ser el impacto.

6.1.1. Modelo de propagación de un virus informático

Evidentemente el modelo matemático de propagación de un virus tanto biológico como informático se puede representar de forma similar, en este caso, a través de ecuaciones diferenciales. Como método ilustrativo, eliminaremos gran parte de la complejidad de los modelos de propagación actuales [20] para poder comparar que tienen ambos en común en un enfoque simplificado. Asumiendo una propagación homogénea, lo que

[17] Se lo clasifica de cibernético ya que esta simbiosis es una interacción de dos o más partes tanto digitales como biológicas.

implica que cada máquina infectada, a su vez, volverá a infectar otra computadora sin restricción alguna e independientemente de su localización, podemos modelar la propagación N de un virus informático [21] a través del siguiente modelo:

$$N(t) = e^{(g-a)\cdot\frac{t}{\tau}}$$

Donde:

g = tasa de infección del virus

a = instancias del virus que fallan

t = tiempo transcurrido

τ = tiempo dado por cada ciclo de infección

Básicamente si definimos cada etapa de la infección como un ciclo, entonces en cada uno de ellos el virus se propaga enviando g instancias de las cuales a fallan.

Si (g-a) < 1 entones la propagación se reducirá, ya que en una función exponencial la forma de minimizar la población de computadoras infectadas sería a través de una reducción del exponente menor a uno. Este fenómeno es conocido en la física nuclear y se denomina subcrítico o masa crítica [22] y hace referencia a cuando no es posible mantener la reacción nuclear en cadena.

6.1.2. Modelo de propagación de un virus biológico

El crecimiento de bacterias está dado por un patrón y está basada en la fisión binaria, la capacidad que tiene una célula en subdividir su ADN para dar lugar a dos células hijas y así sucesivamente desarrollar un crecimiento exponencial. Matemáticamente, de igual forma que con la propagación de un virus informático, también se utilizan ecuaciones diferenciales para calcular la propagación del homónimo biológico. Para eso, se necesita calcular la variación de la población sobre el tiempo que transcurre, como se puede observar en el desarrollo a continuación:

$$\frac{dp}{dt} = k.P \rightarrow \int \frac{dp}{P} = \int k.dt \rightarrow$$

$$\ln P = k.t + C \rightarrow P = c.e^{k.t}$$

Donde se toma arbitrariamente el siguiente ejemplo[18] a modo de demostración:

En el estado inicial t = 0 y se asume que:
C = 10 -> personas infectadas inicialmente

Entonces para t = 10, P y k son respectivamente:
P = 20.000.000 -> población infectada a los diez días
k = 1,45 -> constante de proporcionalidad o crecimiento

De modo que se pueda realizar una comparación, en el ejemplo anterior se asume un escenario donde un brote de una infección comenzara con 10 personas (C), utilizando la constante de proporcionalidad k = 1,45, en un período de diez días[19] (t) el virus podría infectar aproximadamente a la mitad de los habitantes de Argentina (P), o en caso de que se trate de una amenaza digital, a la totalidad de la población que posee una computadora con Internet [23] en el país. Como referencia, la constante de proporcionalidad en este ejemplo está basada en un principio similar al factor de multiplicación de la reacciones nucleares en cadena. En este contexto, se utiliza como referencia [21] que el factor de multiplicación en el accidente nuclear de Chernóbil, Ucrania; fue de aproximadamente 1.2. De hecho, ya en el 2001, en la época de los códigos maliciosos que se propagaban a través de la libreta de contactos del usuario comprometido, se estimaba que ese factor de multiplicación podría ascender a 10. Hoy en día, con la

[18] Este modelo también tiene supone que la velocidad con la que la enfermedad será propagada es proporcional al producto de personas infectadas por no infectadas.

[19] Considerando que la unidad de tiempo son días en este caso.

interconectividad que provee la globalización de las comunicaciones y la masividad de algunas plataformas, como Facebook, cuya cantidad de usuarios activos es mayor a 1.400 millones[20] (o un poco menos de 1 de cada 5 personas en el mundo) – no sería complejo imaginar[21] como esa constante podría ser aún mayor utilizando las redes sociales como vector de propagación.

6.2. Accesibilidad del proceso de síntesis

Varios laboratorios ofrecen servicios de síntesis genética para menos de 3.000 pares de bases, lo cual requiere reensamblar muchas partes para obtener un genoma completo considerando que una secuencia ordinaria suele alcanzar millones de pares. Aún así, el proceso de ensamblaje no es para nada sencillo y no está libre de errores.

Esta es una de las razones por la cual algunos investigadores sostienen [24] que es prácticamente inviable solicitar la secuencia completa de una bacteria a través de Internet para 'crear' un organismo a partir de una computadora de forma completamente independiente. Por lo cual, sintetizar un genoma modificado con un código malicioso no sería aún fácilmente alcanzable para cualquier persona.

No obstante, esto no quita la posibilidad de realizar la infección de los genomas sintéticos hasta que la tecnología avance y haga más sencillo y accesible el proceso de síntesis. Esto significa que tarde o temprano ese genoma infectado podría ser sintetizado.

6.3. Código malicioso + genoma sintético = ¿arma biológica?

Uno de los primeros pensamientos que surge respecto a esta temática es su posible mal uso y qué acuerdos existen para regularlos en caso de que represente una eventual amenaza.

---

[20] Cantidad de usuarios activos por mes correspondiente a septiembre de 2015.

[21] Es importante aclarar que este modelo es teórico y toma en cuenta la propagación en un escenario estable, donde las personas no estén proactivamente protegidas y las condiciones de infección sean ideales.

El 10 de abril de 1972 se firmó [25] el *Biological Weapons Convention* (BWC), un acuerdo internacional que prohíbe el desarrollo y la utilización de armas biológicas, cuya entrada en vigor se realizó en marzo de 1975. Posterior a esas fechas, se fueron agregando revisiones periódicas a la convención con el objetivo de eliminar ambigüedades y hacer más efectivos los controles de cumplimiento entre las naciones.

No obstante, la práctica de modificar información de genomas con un código malicioso digital, que por ende no representa una amenaza biológica, no es tomada en cuenta realmente como una potencial amenaza – al menos no aún. Si bien la naturaleza del ataque es diferente, y hoy este potencial grupo de 'armas biológicas modificadas con una amenaza digital' resulta un caso meramente de laboratorio, queda la decisión en manos de las organizaciones responsables analizar cuándo es el momento apropiado para intervenir mediante regulaciones. Entre otros, uno de los problemas es que el enfoque multidisciplinario de esta temática vuelve cada vez más compleja la clasificación de una potencial amenaza. Por lo pronto, se puede considerar un problema de contexto tecnológico más que biológico pero es indiscutible que la delgada línea que los divide pasa a ser cada vez más difusa con el adelanto de la tecnología y la forma con que nuevas disciplinas emergen de este avance.

6.4. DIY-Biology

El crecimiento de la práctica de "hazlo-tu-mismo" en la biología ha estado cada vez más presente en los últimos años. Uno de los grupos más difundidos a nivel internacional es DIYbio.com cuyos miembros forman equipos interdisciplinarios mayormente en Norteamérica y Europa, mientras que en Latinoamérica es una práctica que va sumando cada vez más adeptos, donde Brasil es el actual referente en la región. Esta tendencia podría dar lugar a que cada vez más 'biólogos por hobby' puedan acceder a conocimientos en la materia sin estar inmersos en un laboratorio supervisado por un especialista. No obstante, esta práctica ya preocupa a algunos profesionales e incluso representantes de gobiernos como por ejemplo [26] un miembro del Consejo Asesor de Ciencia de Estados Unidos:

"A mi me preocupa el científico de garaje, el científico hazlo-tu-mismo, la persona que solo quiere intentar y ver si puede hacerlo" – Zimmer, 2012

Más allá de las opiniones que puedan formarse al respecto, estos grupos vienen aumentando gradualmente así como también el interés en la temática. Internet está permitiendo acercar no solamente personas que actúan por hobby sino también profesionales y especialistas que tienen intereses en común. Un ejemplo sería el caso de *OpenWorm* donde un grupo de personas de diversas partes del mundo trabaja en crear el primer organismo digital que consiste en la copia real de un gusano denominado *Caenorhabditis Elegans* (C. Elegans). Básicamente, se trata de simular en una computadora su estructura completa. El C. Elegans no solo fue el primer organismo a ser secuenciado sino también en tener su cerebro mapeado completamente, dado que es uno de los ejemplares más 'simples' conocidos con únicamente 302 neuronas que forman menos de 10 mil conexiones (sinapsis) entre sí y además posee solo 959 células en su cuerpo. Este equipo internacional se inició inesperadamente a través de un tweet [27] y recaudó[22] 120.000 dólares de financiación en 30 días para su proyecto en la plataforma de financiación colectiva Kickstarter [28]. Esta práctica también podría en un futuro vincular expertos, tanto en disciplinas biológicas como tecnológicas, para propiciar un entorno colaborativo de investigación sobre las amenazas digitales – biológicas.

[22] Inversión obtenida hasta que el objetivo económico de financiación fue alcanzado en mayo de 2014.

7. Revolución en la secuenciación de ADN

La primera vez que se secuenció un genoma humano el proceso llevó 15 años y costó 3 mil millones de dólares. Actualmente, el proceso puede realizarse en un día y los costos se aproximan a los mil dólares [33].

De hecho, algunos sitios [34] en Internet ya ofrecen secuenciar una parte de tu genoma por 99 dólares. Una inversión de 50 millones de dólares en la compañía 23andme, cuya fundadora es Anne Wojcicki esposa de Sergey Brin, co-fundador de Google, les ha permitido bajar el precio de los equipos para realizar la prueba genética de 299 a 99 dólares, lo cuál podría llevar a que 1 millón personas [35] accedieran a esta clase de servicio en un futuro cercano. De hecho, han surgido [36] varias opciones comerciales en los últimos años. Esto implica un cambio en la mentalidad de las personas. La plataforma actualmente ofrece información genética respecto a los lazos familiares del cliente e información ancestral para determinar sus orígenes. No obstante, se podría también obtener información genética respecto a la salud de la persona. De esta forma, se puede determinar qué tan proclive puede ser a una determinada enfermedad antes de sufrirla y de igual forma con sus descendientes. De acuerdo con Wojcicki, cada bebe nacido en Estados Unidos en una década ya tendrá disponibles exámenes de ADN a través del líquido amniótico. No obstante, todavía existen escépticos respecto a la precisión de estos análisis. En 2013, de hecho, la Administración de Drogas y Alimentos estadounidense (FDA) prohibió [37] a 23andme que entregara consejos de salud basados en su test de ADN por lo cual la empresa únicamente continúa en brindar información vinculada a lazos familiares y orígenes.

El mismo J. Craig Venter, luego de su larga trayectoria en la secuenciación del genoma, ha iniciado una nueva compañía denominada "Human Longevity" (Longevidad Humana) cuyo objetivo es tratar enfermedades de forma personalizada a través de la genética particular de sus pacientes. En diciembre de 2014 confirmó [38] que estarán intentando secuenciar 100.000 genomas en el plazo de un año y que se proponen, en tan solo 5 años, para el 2020, alcanzar el millón de genomas secuenciados. Para eso los costos, que actualmente se sitúan entre 1000 y 1500 dólares,

deberían seguir bajando vertiginosamente hasta unos pocos dólares por genoma.

Adicionalmente, ya existen propuestas de crear grandes bancos de datos con información genética de diferentes especies. Una de ellas es la del biólogo keniano Geoffrey Siwo presentada [39] en TED global, en donde busca reunir la mayor cantidad de información genética de afro-americanos, debido a que África es el continente con mayor biodiversidad genética. La otra propuesta proviene de la Universidad de Moscú [40] con apoyo del gobierno ruso, que planea construir una extraordinaria biblioteca de genes que contenga las secuencias – única del mundo en su tipo – de la mayor cantidad posible de criaturas tanto vivientes como extintas, a través de información pre-almacenada o congeladas criogénicamente. De ser así, a partir de 2018 se empezaría a almacenar las secuencias de 'todo lo que contenga vida' en la tierra, una especia de Arca de Noé digital de acuerdo a sus creadores. Sin ninguna duda que una base de datos de esta magnitud será un referente académico de gran importancia con un rol fundamental en las nuevas generaciones e investigación. Sin ir más lejos, en abril de 2015 se confirmó la secuenciación completa [41] de un mamut, extinto a más de 4000 años atrás. Grupos de investigadores en la Universidad de Harvard ya están analizando la posibilidad de de-extinguir el animal. Como vemos, la secuenciación de genomas se está volviendo cada vez más masiva y ambiciosa.

7.1. Programa "Medicina Precisa"

A fines de enero de 2015, el presidente estadounidense Barack Obama anunció [42] que propondrá 215 millones de dólares para financiar la secuenciación de información genética de mujeres, hombres, ancianos, niños, enfermos y sanos de modo de crear un gran conjunto de genomas. Esta base de datos tiene como objetivo principal mejorar el tratamiento de cáncer pero formaría parte de una iniciativa a largo plazo para entender mejor el genoma humano y sentar las bases para nuevos descubrimientos. El proyecto propone a largo plazo la secuenciación y el almacenamiento de un millón de genomas.

7.2. Google Genomics

En marzo de 2014, Google lanzó silenciosamente [43] el proyecto Google Genomics que básicamente provee una API que permite almacenar, procesar, analizar y compartir secuencias de ADN. A su vez, la API implementada por Google es la definida [44] por la Alianza Global de Salud y Genomas.

Por un lado, los bajos costos de secuenciación brindan más posibilidades de secuenciar nuevos genomas y nos deja cada vez más cerca de que cada persona puede poseer su propio genoma secuenciado. No obstante, la información en crudo para analizar de un genoma humano puede llegar a ser de 1 TB o 100 GB comprimida, y en el mejor de los casos. Por lo tanto, el almacenamiento todavía es una barrera por el usuario tradicional y sobre todo para muchos laboratorios que no poseen esa disponibilidad. De hecho, que iniciativas como la de Google, una de las empresas con mayor capacidad de almacenamiento del mundo, comiencen a proveer estos servicios a 25 dólares[23] por terabyte utilizado; realmente abre una puerta más a la masificación de la secuenciación genómica.

Estos avances también acompañan a los premios que se entregaron en noviembre de 2014 en la Breatkthrough Prize Award Ceremony [45], donde se reconocen tres temáticas fundamentales que probablemente marcarán tendencias en los próximos años en el campo de las ciencias de la vida, matemática y física. Una de ellas, es justamente buscar nuevas y poderosas tecnologías para analizar y editar genomas.

[23] Precios efectivos a partir de octubre de 2014 de acuerdo al sitio de Google Genomics: https://cloud.google.com/genomics/pricing.

8. Conflictos éticos y regulaciones

Ya se ha observado en producciones cinematográficas, el temor por la discriminación genética que podría ocurrir debido al avance en el estudio del ADN y los conflictos éticos que desencadenaría. No obstante, vale aclarar que si se analizan los temores respecto a la posibilidad de manipular un ser humano a través de su configuración genética, se encontrará que a diferencia de lo que comúnmente se creía; los genomas no necesariamente exhiben un comportamiento determinista [49]. Esto significa que no siempre el resultado de una variación genética es posible de prever dado que hay muchas variables biológicas y ambientales que tienen consecuencia sobre el organismo.

A pesar de que en el campo de la seguridad el dilema ético no es un inconveniente tan polémico como el citado anteriormente, hay algunos puntos a tener en cuenta. Por ejemplo, vale la pena preguntarse qué tan arriesgada puede ser la manipulación de un genoma sintético, si esta práctica un día se volviese habitual y más accesible. Como hemos analizado anteriormente, uno de los mayores temores al vincular la genómica con la seguridad informática es el poder 'insertar' amenazas en las secuencias. Ocultar código malicioso cifrado a través de 4 nucleótidos cuya clave[24] sea secreta y solo conocida por el *malware* y su(s) autor(es). Adicionalmente, surgen algunas dudas como por ejemplo, qué pasa si este código malicioso se comporta como un virus borrando o modificando una parte de la secuencia que pudiera comprometer el comportamiento del genoma. El dilema ético podría dar lugar a una respuesta más lenta en proteger a las secuencias de estas modificaciones no deseadas, ya que la seguridad sobre las secuencias podría ser un área que no reciba atención suficiente para ser investigada.

De todos modos, existen muchas barreras por resolver aún para que esto ocurra en el corto plazo y además es de asumir que el atacante tradicional no tendría conocimientos de química y genética suficientes para

[24] Asumiendo un cifrado simétrico de modo que solamente los que poseen la clave puedan descifrar el mensaje.

causar daños de impacto biológico. Sin embargo a medida que avanzan las investigaciones es de esperar que los ataques evolucionen y que, en caso de que sea realmente factible y provechoso, esta técnica pueda ser utilizada por el cibercrimen o incluso algún gobierno (el denominado *state-sponsored malware*) como ya ocurrió con *Stuxnet y Flamer* [50], en donde los recursos dejan de ser una variable limitante.

8.1. Privacidad y confidencialidad en el uso de ADN humano

Otro punto importante sería revisar los marcos regulatorios respecto a la seguridad que ofrece la manipulación de secuencias. Por ejemplo, el Departamento de Salud y Servicios Humanos estadounidense (HHS) posee una regulación de 'Protección a sujetos humanos' que contiene un apartado [51] identificado bajo el código de regulaciones federales (CFR) título 45, parte 46 y subparte A denominado 'Políticas para la protección de sujetos de investigación humanos'. En la primera sección, 46.101 el documento aclara que esta política se aplica a toda investigación en la que participen humanos que sea apoyada, conducida o sujeta a regulación de parte de alguna agencia o departamento federal, independientemente si se encuentra en Estados Unidos.

Adicionalmente, el Instituto de Investigación del Genoma Humano (NIH) posee Directrices sobre las Problemáticas en la Secuenciación de ADN a gran escala [52]. Básicamente, aborda los siguientes seis ejes:

1. Beneficios y riesgos de la secuenciación del ADN genómico
2. Privacidad y Confidencialidad
3. Reclutamiento de donadores de ADN como fuentes
4. Consentimiento informado
5. Aprobación de la junta de revisión (IRB) o comité ético
6. Uso de bibliotecas de datos existentes

Esta política, que tuvo sus comienzos desde agosto de 1996, ya tomaba en cuenta la importancia de la privacidad de la información de los eventuales participantes del proyecto de secuenciación más grande que se

iba a desarrollar, el Proyecto Genoma Humano. En el punto 2, dedicado exclusivamente a la privacidad, podemos observar que la política que garantizaba la integridad y confidencialidad de la información de la persona que donaba su ADN estaba basada en eliminar todo tipo de información personal, y por ende, que se vinculaba con el paciente. De esa forma, se buscaba 'anonimizar' las muestras. Asimismo, también resultaba una ironía poder llevar a cabo la anonimización completa debido a que el ADN de una persona es único e irrepetible, por lo cual la muestra misma de ADN resultaba ser el mayor y mejor identificador del paciente. No obstante, se sabía que la tecnología y el tiempo necesario para poder identificar unívocamente una persona todavía no se había alcanzado:

Privacy & Confidentiality

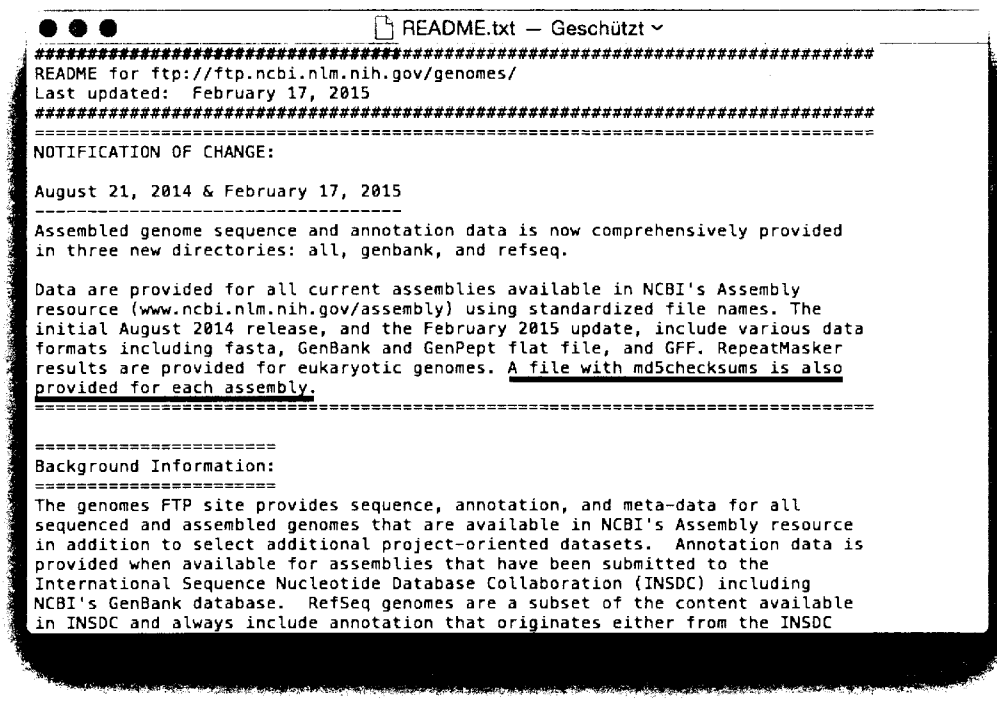
In general, one of the most effective ways of protecting volunteers from the unexpected, unwelcome or unauthorized use of information about them is to ensure that there are no opportunities for linking an individual donor with information about him/her that is revealed by the research. By not collecting information about the identity of a research subject and any biological material or records developed in the course of the research, or by subsequently removing all identifiers ("anonymizing" the samples), the possibility of risk to the subject stemming from the results of the research is greatly reduced. Large-scale DNA sequence determination represents an exception because each person's DNA sequence is unique and ultimately, there is enough information in any individual's DNA sequence to absolutely identify her/him. However, the technology that would allow the unambiguous identification of an individual from his/her DNA sequence is not yet mature. Thus, for the foreseeable future, establishing effective confidentiality, rather than relying on anonymity, will be a very useful approach to protecting donors.

Imagen 7: Captura de la regulación 'Protección a sujetos humanos' (45–46, A)

Sin embargo, como ya comentamos anteriormente, Yaniv Erlich y su equipo, lograron demostrar en 2013 una forma de localizar a través de inferencia por el nombre familiar y la muestra con una tasa de éxito no despreciable. Si bien en este caso, mantener la privacidad de los pacientes quizá no fue un objetivo en particular, sirve como referencia para tener en cuenta que actualmente es cada vez más complejo 'anonimizar' secuencias.

Además la gestión y administración de las fuentes o 'librerías', donde se almacenan las muestras, contiene puntos interesantes a debatir. Por ejemplo, dónde y cómo establecer el almacenamiento seguro de esta información, la necesidad de cifrado, protocolos seguros de acceso, servidores de almacenamiento estatales u organismos independientes, la mejor implementación de un número de control a través de un *hash* para poder detectar modificaciones en los genomas y etcétera. Respecto al uso

de hash, algunas instituciones como el NCBI ya lo hacen, sin embargo el *checksum* no está usualmente en el directorio raíz, por lo cual podría llevar a que personas no familiarizadas no lo encuentren. De hecho, incluso aquellos que lo encuentren, quizá no tengan el hábito de usarlo. En este caso particular, se podría controlar el *checksum* automáticamente cuando se descargue la secuencia de la misma forma que hacen los archivos ZIP, de modo que, en caso que la comprobación falle; entonces el archivo se indicaría corrupto y no se permitiría abrir. El número se podría usar, por ejemplo, en la cabecera del FASTA.


README.txt — Geschützt

```
################################################################################
README for ftp://ftp.ncbi.nlm.nih.gov/genomes/
Last updated:  February 17, 2015
################################################################################
================================================================================
NOTIFICATION OF CHANGE:

August 21, 2014 & February 17, 2015
-----------------------------------
Assembled genome sequence and annotation data is now comprehensively provided
in three new directories: all, genbank, and refseq.

Data are provided for all current assemblies available in NCBI's Assembly
resource (www.ncbi.nlm.nih.gov/assembly) using standardized file names. The
initial August 2014 release, and the February 2015 update, include various data
formats including fasta, GenBank and GenPept flat file, and GFF. RepeatMasker
results are provided for eukaryotic genomes. A file with md5checksums is also
provided for each assembly.
================================================================================

=======================
Background Information:
=======================
The genomes FTP site provides sequence, annotation, and meta-data for all
sequenced and assembled genomes that are available in NCBI's Assembly resource
in addition to select additional project-oriented datasets.  Annotation data is
provided when available for assemblies that have been submitted to the
International Sequence Nucleotide Database Collaboration (INSDC) including
NCBI's GenBank database.  RefSeq genomes are a subset of the content available
in INSDC and always include annotation that originates either from the INSDC
```

Imagen 8: Archivo Léeme en FTP de NCBI indicando uso de md5

Estas son apenas algunas de las cuestiones que emergen para ser reguladas por los organismos competentes. Es muy probable que estas regulaciones no deban contener específicamente el algoritmo utilizado para cifrar los archivos o el protocolo de comunicación seguro para acceder a los registros, no obstante menciones generales a la seguridad de esa información quizá valen un inciso más, o incluso, quizá, una propia sección dentro de la subparte correspondiente.

9. Conclusión

Esta investigación no está orientada a atacar con miedo a una problemática que surge producto del avance tecnológico. Por el contrario, procura aumentar la atención sobre este tema y concientizar sobre lo que podría ocurrir si los atacantes resuelven prestar más atención a este tipo de ataques. Como se expuso anteriormente, es posible afectar la integridad de un archivo que contiene una secuencia genómica a través de un algoritmo, que podría pasar como no detectado en el sistema de la víctima sin grandes inconvenientes. Esa modificación podría generar de inmediato una acción maliciosa digital y, en un futuro, quizá, comprometer la secuencia a nivel biológico. Todo, por supuesto, sin consentimiento de la víctima. No obstante, lamentablemente esta temática no llamará la atención del punto de vista de la seguridad hasta que no aumenten las siguientes condiciones: criticidad y masividad.

La criticidad de este tipo de información va adquiriendo más relevancia. Por ejemplo, anteriormente la información sensible interesante para los ciberdelincuentes eran datos personales, credenciales bancarias, información financiera de las víctimas y demás información de esa índole. No obstante, hoy por hoy, el mercado ilícito se está interesando más por los registros médicos de las personas. Dicha información podría valer hasta diez veces más [53] que una tarjeta de crédito. Esto abre camino a plantearse algunas interrogantes como, por ejemplo, cuanto tiempo pasará hasta que el ADN de una víctima sea comercializado en el mercado negro o hasta que sea interesante infectar un sistema para modificar una secuencia genómica.

Respecto a la masividad, está claro que este no es un problema que hoy afecte a un grupo numeroso de personas, pero debido a la drástica reducción en los costos de secuenciación; es probable que aumente el número de personas interesadas en esta temática en menos tiempo del que creemos. Asimismo, el costo más accesible de equipamiento profesional también abriría puertas a que instituciones con menos recursos formen sus propios bancos de genomas e ingresen a la disciplina creando así nuevos centros con secuencias genómicas a gran escala en el corto y mediano plazo.

La industria de la salud ha sido blanco de numerosos ataques en los últimos años que van desde comprometer registros confidenciales, atacar dispositivos médicos que dosifican la cantidad de droga administrada a un paciente, hasta utilizar *ransomware* para secuestrar información de hospitales. Y ahora también entra en juego comprometer las secuencias de genomas.

En resumen, todo indica que nuevos y más sofisticados ataques podrían ir gestándose en los próximos años. Por eso, este trabajo busca llamar la atención sobre una de las temáticas que está adquiriendo potencial para representar un inconveniente adicional del punto de vista de seguridad. El simple hecho que sea factible y que pueda evitar detección, resultaría suficiente para atraer atacantes antes que llame la atención de los investigadores.

10. Glosario

Aminoácido: componente orgánico que contiene un grupo amino y un grupo carboxilo. Son las partes elementales de las proteínas.

Base (nitrogenada): es denominada también base nucleica. Las principales cuatro presentes en el ADN son: adenina (A), citosina (C), timina (T) y guanina (G). Existe una quinta, uracilo (U), que se encuentra solo en el ARN.

Biohacking: el arte de conjugar la ciencias biológicas con la curiosidad y el ingenio que fomenta el hacking.

Célula: es el componente básico de todos los seres vivos. El cuerpo humano está compuesto por billones de células que proveen la estructura del cuerpo. Además, toman los nutrientes de la comida y lo convierten en energía necesaria para llevar a cabo las funciones específicas.

Codón: una secuencia formada por tres nucleótidos consecutivos que codifica un aminoácido – parte fundamental de la proteína.

Covert channel (o canal encubierto): es un tipo de comunicación utilizada para transferir un mensaje oculto a través de canales no convencionales que no fueron creados con ese fin. En caso de realizado exitosamente, son difíciles de detectar utilizando mecanismos de seguridad tradicionales.

Cromosoma: estructura circular dentro de una bacteria que contiene la mayor parte de ADN de un organismo vivo, la información hereditaria necesaria para la vida de la célula.

DIY-Biology: Biología hazlo-tu-mismo, usualmente se realiza de parte de 'científicos amateurs' que buscan experimentar en laboratorios de garaje, pero en los últimos años los laboratorios se van haciendo más sofisticados y los adeptos varían entusiastas y personas más calificadas.

Fenotipo: la expresión de un rasgo específico como el tipo de sangre, color de ojos, etc. que están influenciadas por la combinación genética con el ambiente.

Genes: unidad hereditaria que está formada por una secuencia de ADN, localizada en un cromosoma y que determina una característica en un organismo.

Genotipo: el conjunto de genes que contiene un organismo. Toda la información genética en forma de ADN.

Haplotipo: se refiere a las variaciones del ADN que usualmente un individuo hereda de uno de sus padres.

In the wild: Cuando un código malicioso ya se encuentra afectando usuarios en el mundo real y no es solo una prueba de concepto.

Mosca de la fruta (*fruit fly*): bajo el nombre de *Drosphila melanogaster*, la mosca de la fruta es utilizada ampliamente en las investigaciones debido a su simple estructura genética y rápida reproducción lo cual es interesante para estudiar sus mutaciones.

Nucleótido: moléculas que cuando se agrupan forman el bloque fundamental del ADN o ARN.

PoC: Proof-of-Concept o Prueba de concepto.

PCR: Hace referencia a la reacción en Cadena de la Polimerasa y se trata de una técnica para amplificar copias de ADN que después se pueden utilizar en la identificación de una bacteria, virus, etc.

Proteína: moléculas de gran tamaño que contienen una o más cadenas de aminoácidos.

Ribosomas: también conocidos como una máquina molecular. Está presente en todas las células y es el lugar donde se realiza la traducción genética.

Síntesis de proteínas: el proceso realizado por el código genético mediante el cual se crean proteínas en una célula.

Stop Codón: codón utilizado para señalizar el final en la síntesis de proteínas. Los codones utilizados para este propósito en el ADN son TAG, TAA y TGA; y en ARN son: UAG, UAA y UGA.

Traducción genética: el proceso mediante el cual los ribosomas crean proteínas.

11. Anexo

11.1. Análisis de frecuencias en marcas de agua

En esta sección se explicarán los pasos realizados para descifrar el código utilizado por el Instituto J. C. Venter. Esta técnica denominada en criptografía como análisis de frecuencias, podría ser utilizada para determinar la clave de cualquier tipo de *watermark* (que contendría *malware*) que haya sido oculto[25] en la secuencia del genoma.

El primer punto fue hacer un algoritmo para tomar todos los nucleótidos y agruparlas de a tres, dividiéndolos en codones.

En segundo lugar, a través de un análisis de frecuencias se comenzó a descifrar el código en el que está protegido el mensaje. Al codón (secuencia genética de tres letras) que más se repite, lo que en estadística se conoce como 'moda', se le asignó la letra E, cuya frecuencia es la mayor en la lengua inglesa. De igual forma, se asignó la letra T al siguiente codón de mayor frecuencia. No obstante, a partir de ese momento ya no era seguro seguir asignando porque la distribución de frecuencias puede no ser confiable con un texto de longitud corta/mediana. Entonces, se comenzaron a inferir el resto de las letras. Por ejemplo, en el espacio entre una T y una E, se infirió una H. Sucesivamente, era de esperarse que la palabra "Venter" estuviese incluida en el mensaje, por lo cual se buscó en la marca de agua una E seguida de una T con un espacio al medio que pudiera representar una N.

[25] En este caso en particular, se poseía de antemano una porción de texto claro que simplifica el descifrado. Para el caso de códigos maliciosos, se deberían usar posibles *strings* por defecto en determinados lenguajes para inferir una porción de texto claro.

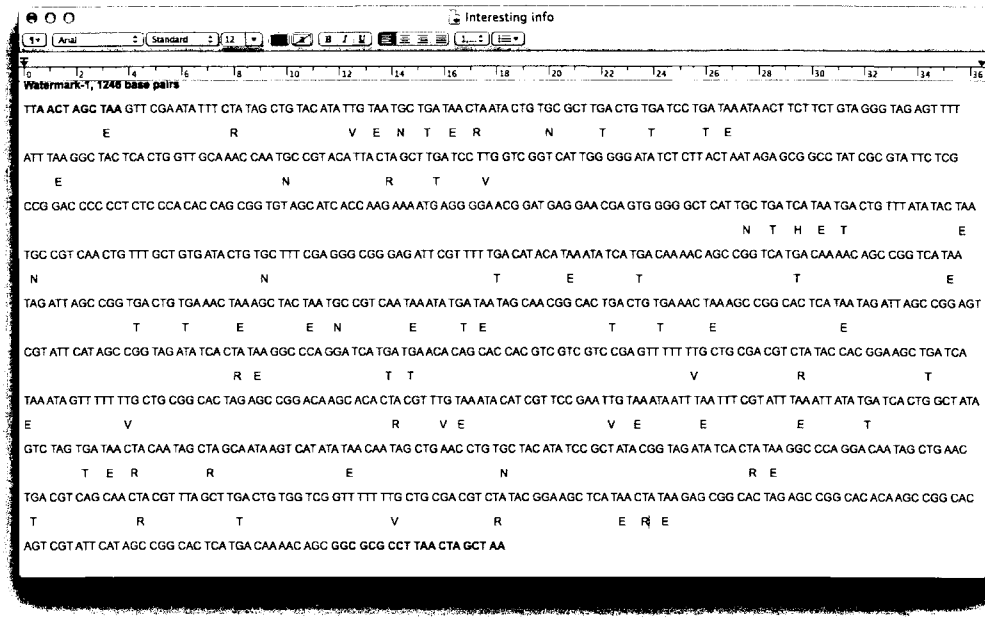


Imagen 9: 1era etapa del análisis de frecuencias a una marca de agua – JCVI

Existían dos casos, no obstante, uno de ellos poseía una ocurrencia muy alta en el texto para representar una letra N, que acorde a la tabla de frecuencias del inglés, es una letra que ocupa la 6ta posición, por lo tanto quedaba descartada.

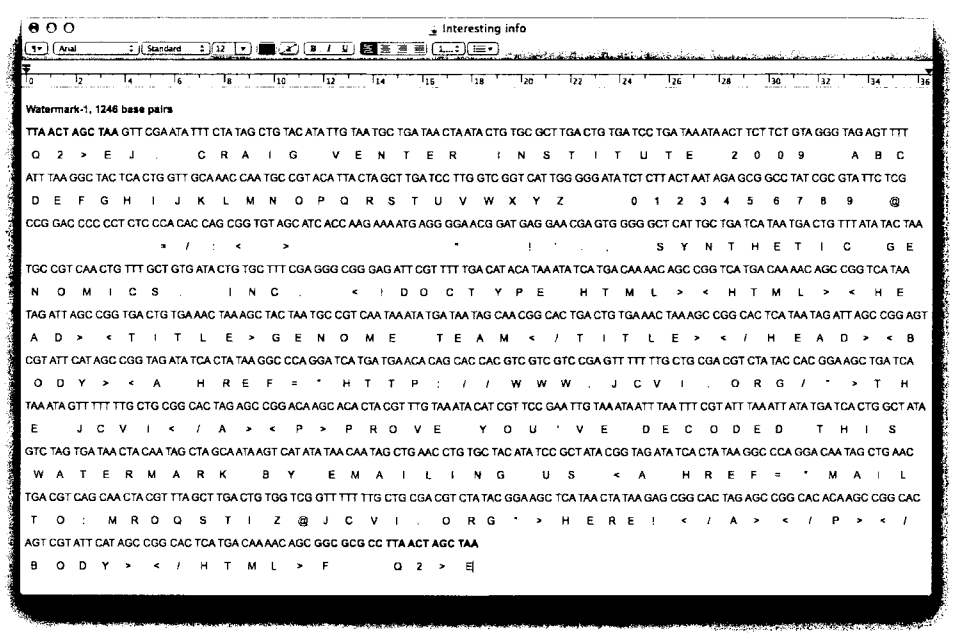


Imagen 10: Continuación de análisis de frecuencias a marca de agua – JCVI

Una vez que se tenía esas cuatro letras, se podía inferir que ENTE sea una subcadena de "Venter". Inmediatamente, se observaba la cadena "Craig" e "Institute" hacia ambos lados de Venter. Posterior a eso, ya habían letras suficientes para completar palabras restantes e ir infiriendo el resto del texto. Cuando gran parte del texto ya estaba descifrado, se pudo visualizar que se citaban sitios web y direcciones de correo electrónico, por lo cual el código también poseía puntuaciones y símbolos como barras y arroba.

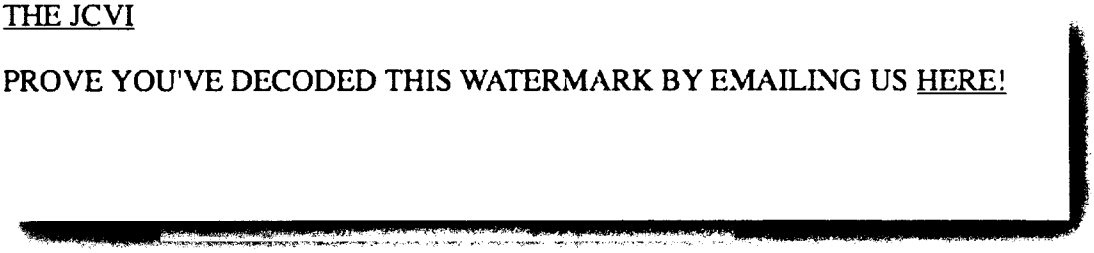


Imagen 11: Marca de agua del J. Craig Venter Institute decodificada

Adicionalmente, se utiliza un codón específico (ATA) para representar los espacios. A partir de ahí, prácticamente todo el texto estaba descifrado y la clave utilizaba menos de 40 codones de los 64 disponibles: $c^n = 64$, donde c = #{A, C, T, G} y n = 3 por la longitud de un codón.

12. Bibliografía

[39] Biólogo busca fonte contra doenças [consultada apr-15] http://oglobo.globo.com/sociedade/tedglobal-2014/biologo-busca-fonte-para-combater-todas-as-doencas-14163704

[18] Cientistas sequenciam mosca comum [consultada mar-15] http://ciencia.estadao.com.br/noticias/geral,cientistas-sequenciam-genoma-da-mosca-comum,1577438

[19] Can we edit our genes to fight desease? [consultada may-15] www.nibr.com/stories/discovery/can-we-edit-our-genes-fight-disease

[50] Cinco mitos o verdades sobre Flamer [consultada jun-15] http://www.welivesecurity.com/la-es/2012/06/01/5-mitos-verdades-sobre-flamer/

[8] Compilation of Genome Sequence [consultada jan-15] http://www.yourarticlelibrary.com/biotechnology/genomics/compilation-of-genome-sequence/33946/

[33] Cost per genome – Genome.gov [consultada nov-14] http://www.genome.gov/images/content/cost_per_genome2.jpg

[2] Cell Controlled by a Synthesized Genome [consultada feb-14] https://www.sciencemag.org/content/329/5987/52.full, 2010

[48] Datenschutz: Gehackte Genes [consultada jan-15] http://www.deutschlandfunk.de/datenschutz-gehackte-gene.740.de.html?dram:article_id=299344

[9] DNA sequencing? [consultada jan-15] http://www.genome.gov/10001177#al-2

[7] DNA Sequencing, Encyclopedia Britannica [consultada dic-14] http://www.britannica.com/EBchecked/topic/422006/DNA-sequencing

[34] Find out what your DNA says about you [consultada nov-14] https://www.23andme.com

[1] First Self-Replicating Synthetic Cell [consultada feb-14] http://www.jcvi.org/cms/press/press-releases/full-text/article/first-self-replicating-synthetic-bacterial-cell-constructed-by-j-craig-venter-institute-researcher/home/, 2010.

[36] Five services that will sequence your DNA [consultada jan-15]

http://mashable.com/2013/05/15/personal-genetics-resources/

[29] Genomes edited to free up codons, 2011 [consultada oct-14]
http://www.nature.com/news/2011/110714/full/news.2011.419.html

[5] Genome Home Reference – 2013 [consultada nov-14]
http://ghr.nlm.nih.gov/handbook/hgp/genome

[44] Google Genomics [consultada nov-14]
https://cloud.google.com/genomics/what-is-google-genomics

[43] Google quer armazenar DNA na nuvem [consultada nov-14]
http://oglobo.globo.com/sociedade/tecnologia/google-quer-armazenar-dados-de-dna-em-servicos-de-nuvem-14635254

[53] Healthcare data worth ten times price of credit card data
http://www.welivesecurity.com/2014/09/25/healthcare-security/

[10] How do we sequence DNA? [consultada jan-15]
http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/sequencing.html

[38] How the genomic era is just starting [consultada jan-15]
http://www.businessweek.com/articles/2014-12-04/dna-sequencing-craig-venter-says-genomic-era-is-just-starting

[49] Human genome: the real ethical dilema [consultada mar-15]
http://www.theguardian.com/science/2013/sep/09/genetics-ethics-human-gene-sequencing

[46] Identifying Genome by Surname Inference [consultada jan-15]
http://www.sciencemag.org/content/339/6117/321.full.pdf

[3] Infection of DNA with Computer Code [consultada oct-14]
http://spth.virii.lu/InfectingDNA.txt, 2013

[35] Inicia campaña más análisis genéticos [consultada nov-14]
http://www.unocero.com/2013/08/07/inicia-campana-para-llevar-el-analisis-genetico-a-las-masas/

[32] Integrated encyclopedia of DNA elements [consultada mar-15]
www.nature.com/nature/journal/v489/n7414/full/nature11247.html

[23] Internet en Argentina [consultada jul-15]
http://www.palermo.edu/cele/pdf/investigaciones/Mapping-ARG-CELE.pdf

[27] Is this worm the first sign of singularity? [consultada jan-15]

http://www.theatlantic.com/technology/archive/2013/05/is-this-virtual-worm-the-first-sign-of-the-singularity/275715/

[41] Mammoth genome sequence completed [consultada apr-15]
http://www.cell.com/current-biology/abstract/S0960-9822(15)00420-0

[11] Medre: AutoCAD files leaked [consultada may-15]
http://www.welivesecurity.com/media_files/white-papers/ESET_ACAD_Medre_A_whitepaper.pdf

[21] Models of Viral Propagation [consultada jul-15]
http://www.symantec.com/connect/articles/models-viral-propagation

[20] Model of propagation of a computer virus [consultada jul-15]
http://home.deib.polimi.it/dercole/csr/CSR-Sem01.pdf

[37] Mrs. Google's DNA test for her daughter [consultada jan-15]
http://www.dailymail.co.uk/health/article-2924572/Mrs-Google-s-DNA-test-unborn-girl-Wife-internet-tycoon-daughter-tested-risk-cancer-Alzheimer-s-Parkinson-s-later-life.html

[40] Noah Ark: Russia to build DNA databank [consultada jan-15]
http://rt.com/news/217747-noah-ark-russia-biological/

[22] Nuclear Regulatory Commision [consultada jul-15]
http://www.nrc.gov/reading-rm/basic-ref/glossary/subcriticality.html

[42] Obama announces to analyze genomes [consultada feb-15]
http://www.huffingtonpost.com/2015/01/30/obama-precision-medicine_n_6580892.html

[28] OpenWorm: digital organism in a browser [consultada jan-15]
https://www.kickstarter.com/projects/openworm/openworm-a-digital-organism-in-your-browser

[30] Organización Mundial de la Salud: E.Coli [consultada nov-14]
http://www.who.int/topics/escherichia_coli_infections/es/

[14] Personal Genome Project – Harvard [consultada nov-14]
http://www.personalgenomes.org/harvard/data

[14] Python to evade antivirus [consultada mar-15]
https://samsclass.info/124/proj14/p8-av.htm

[31] Redundant genes in "junk DNA" [consultada oct-14]
http://www.sciencedaily.com/releases/2010/07/100716125835.htm

[17] Telstra: UK wants DNA as authentication [consultada jun-15]

www.computerweekly.com/news/4500248195/UK-bank-customers-set-to-share-DNA-data-with-banks-for-biometric-authentication

[25] The Biological Weapons Convention – UN [consultada jan-15] http://www.un.org/disarmament/WMD/Bio/

[47] The Genome Hacker [consultada jan-15] http://www.nature.com/polopoly_fs/1.12940!/menu/main/topColumns/topLeftColumn/pdf/497172a.pdf

[26] The growth of amateur biology [consultada nov-14] http://biochemsec2030dotorg.files.wordpress.com/2013/08/jefferson-policy-paper-3-for-print.pdf

[24] The myth & realities of synthetic weapons [consultada nov-14] http://thebulletin.org/myths-and-realities-synthetic-bioweapons7626

[45] These are 3 breakthrough ideas for 2015 [consultada dec-14] http://www.washingtonpost.com/blogs/innovations/wp/2014/11/11/these-are-3-breakthrough-science-ideas-youll-be-talking-about-in-2015/

[51] U.S. Department of Health & Services [consultada mar-15] www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.115

[52] U.S. National Human Genome Institute [consultada mar-15] http://www.genome.gov/10000921#top

[4] U.S. National Library of Medicine [consultada mar-15] http://ghr.nlm.nih.gov/handbook/basics/dna

[6] University of Princeton, Genome [consultada oct-14] http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Genome.html

[15] Veil – Metasploit payload generator [consultada mar-15] http://www.hackingarticles.in/veil-a-metasploit-payload-generator-to-bypass-antivirus/

[12] Virustotal scan from LocateFiles script in Python https://www.virustotal.com/de/file/28c16881435c0f43fdd15bdc88cd98659c2937401af5583a5967b04954447fc1/analysis/1446751991/

[13] Virustotal scan from InjectPayload script in Python https://www.virustotal.com/de/file/88c1ebb7c938fe749aeb130049cb2a3ea230e642ff5ef5f66eccbb71de585f0e/analysis/1447010447/

[16] Vulnerabilidades en Java y Gestión en Oracle, Universidad de Buenos Aires, (2013), pp. 14 – 25, TFE - Raphael Labaca Castro