Ingeniería Matemática
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

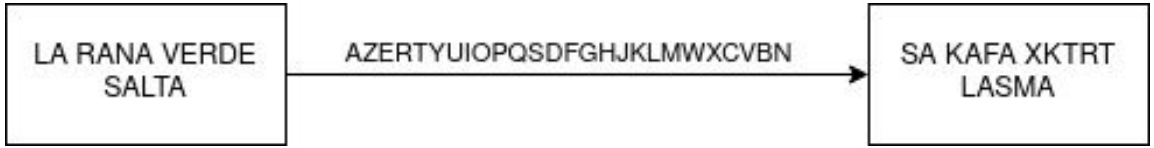
Descifrando mensajes codificados usando MCMC

Sebastian Flores y Matías Neto

MA4402-1: Simulación estocástica: T. y L.
Profesor: Joaquin Fontbona
Auxiliares: Catalina Lizana, Álvaro Márquez y Matías Ortiz.

Contexto: Criptografía clásica

- Claves de sustitución simple: reemplazo de un símbolo por otro.



- Problema: obtener un método sistemático para obtener el texto original a partir del texto codificado (“romper” el cifrado)

Formulación del problema

- Definimos un grafo para aplicar el método de Simmulated Annealing

  V = {Permutaciones del alfabeto latino estándar}

  P ↔ Q si están a una permutación de distancia

- Función a maximizar: Plausibilidad

$$\mathrm{Pl}(f) = \prod_i M_{f(s_i)f(s_{i+1})}$$

  - Ventaja: No depende de la clave que se está intentando romper
  - Computacionalmente, maximizaremos la log-plausibilidad

Primeros test

- Creación de la matriz M a partir de un texto largo (~ 120.000 palabras)
- Prueba del método con un texto corto (letra ñ removida)

“MUCHOS ANOS DESPUES FRENTE AL PELOTON DE FUSILAMIENTO EL CORONEL AURELIANO BUENDIA RECORDO AQUELLA TARDE REMOTA EN QUE SU PADRE LO LLEVO A CONOCER EL HIELO”

- Fijamos los pasos de la cadena a 10.000

Primeros test

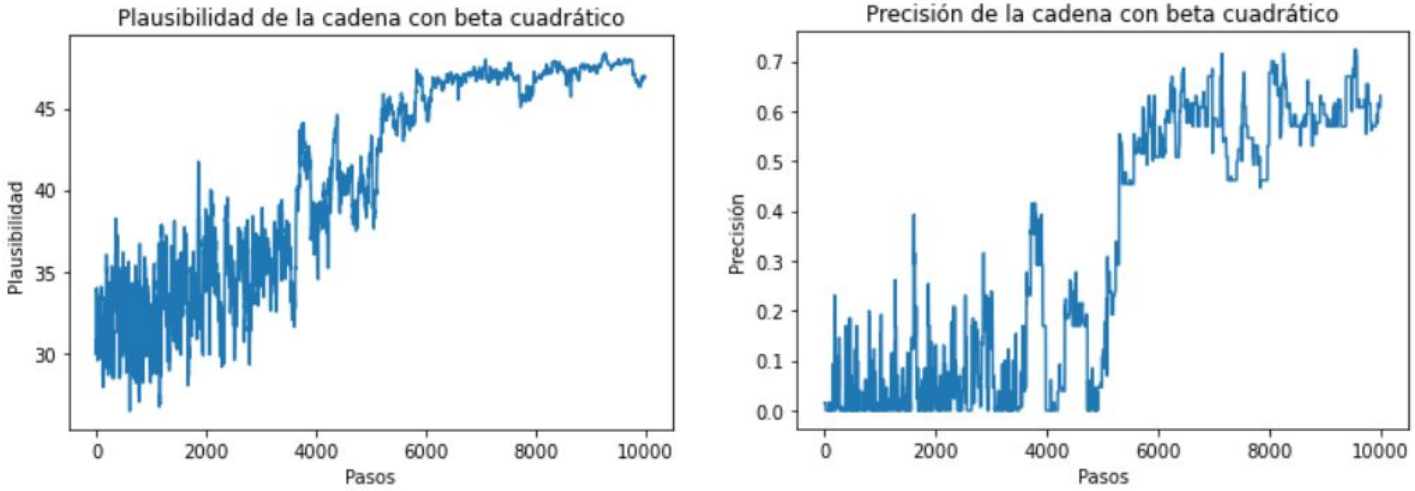


Figura 1: Resultados para beta(n) = $10^{-7}n^2$

“CUJROK ANOK VEKHUEK GDENTE AS HESOTON VE GUKISACIENTO ES JODONES AUDESIANO PUENVIA DEJODVO AQUESSA TADVE DECOTA EN QUE KU HAVDE SO SSEYO A JONOJED ES RIESO” - máximo de plausibilidad

Ataque de trigrama

- Reemplazo de la función de plausibilidad por

$$\text{Tri-Pl}(f) = \prod_i M_{f(s_i)f(s_{i+1})f(s_{i+2})}$$

Resultados del ataque de trigrama

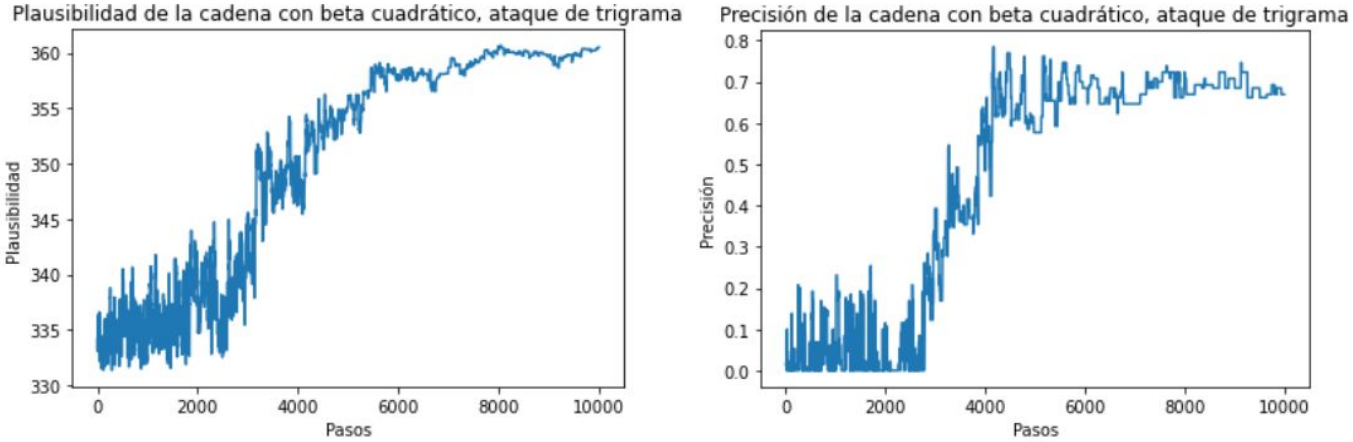


Figura 2: Resultados del ataque de trigrama para beta cuadrático

“CUPKOS AZOS DESVUES LREZME AN VENOMOZ DE LUSINACIEZMO EN POROZEN AURENIAZO BUEZDIA REPORDO AQUENNA MARDE RECOMA EZ QUE SU VADRE NO NNEYO A POZOPER EN KIENO” - máximo de plausibilidad

Influencia de las condiciones iniciales

- Iteramos con 20 condiciones iniciales para cada método
- Agregamos un texto de prueba de mayor extensión (~1800 palabras)
  - Sacados de la página de Wikipedia de Roberto Bolaño
- Consideramos la clave que maximiza la plausibilidad
- Usamos un beta cuadrático para la optimización

Resultados

Texto corto

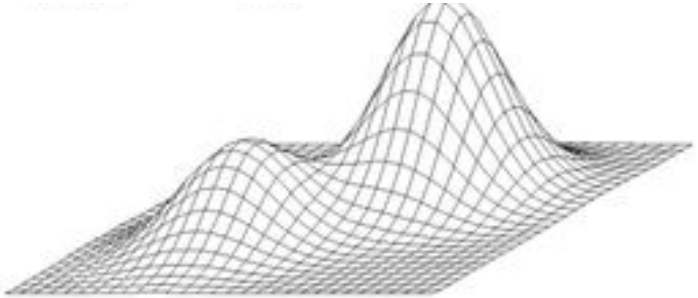
	Log-plausibilidad	Precisión
Bigrama	48.43 (47.448)	56.9%
Trigrama	360.91 (360.26)	90.8%

Texto largo

	Log-plausibilidad	Precisión
Bigrama	286.88 (283.58)	89.4%
Trigrama	649.33 (649.33)	100%

Benchmark: Stochastic Hill Climbing

- Hill Climbing: Algoritmo que encuentra óptimos locales
- Stochastic Hill Climbing: Aumenta las probabilidades de encontrar óptimos globales, al introducir aleatoriedad en escoger vecinos del estado actual.
- Diferencias entre SA y Stochastic Hill Climbing
  Ventaja: Reducción del costo computacional
  Desventaja: Sensibilidad a las condiciones iniciales



Resultados Stochastic Hill Climbing

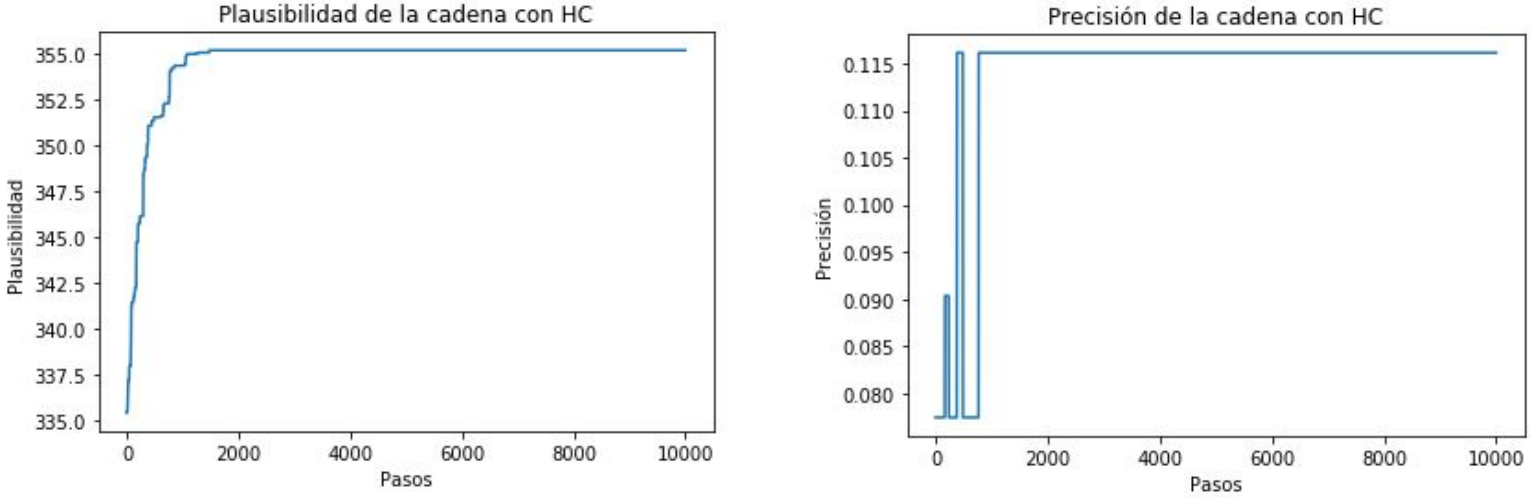
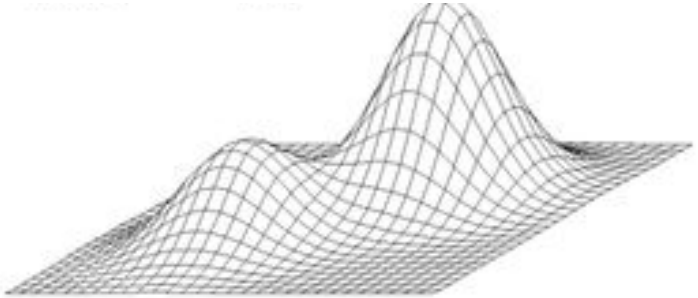


Figura 5: Resultados del ataque de trigrama para HC con texto corto

"FRIGUE AQUE DOEMROE ZNOQHO AS MOSUHUQ DO ZRELSAFLOQHU OS IUNUQOS ARNOSLAQU TROQDLA NOIUNDU APROSSA HANDO NOFUHA OQ PRO ER MADNO SU SSOCU A IUQUION OS GLOSU - máximo de plausibilidad

Benchmark: Stochastic Hill Climbing Backtracking

- Backtracking: Permite encontrar múltiples candidatos a solución
- Stochastic Hill Climbing Backtracking: Aumenta las probabilidades de encontrar óptimos globales, pues encuentra múltiples óptimos locales y así propone un óptimo global.
- Stochastic Hill Climbing vs Stochastic Hill Climbing Backtracking
  Ventaja: Pierde la sensibilidad a su condición inicial
  Desventaja: Aumenta levemente el costo computacional de HC



Benchmark: Stochastic Hill Climbing Backtracking

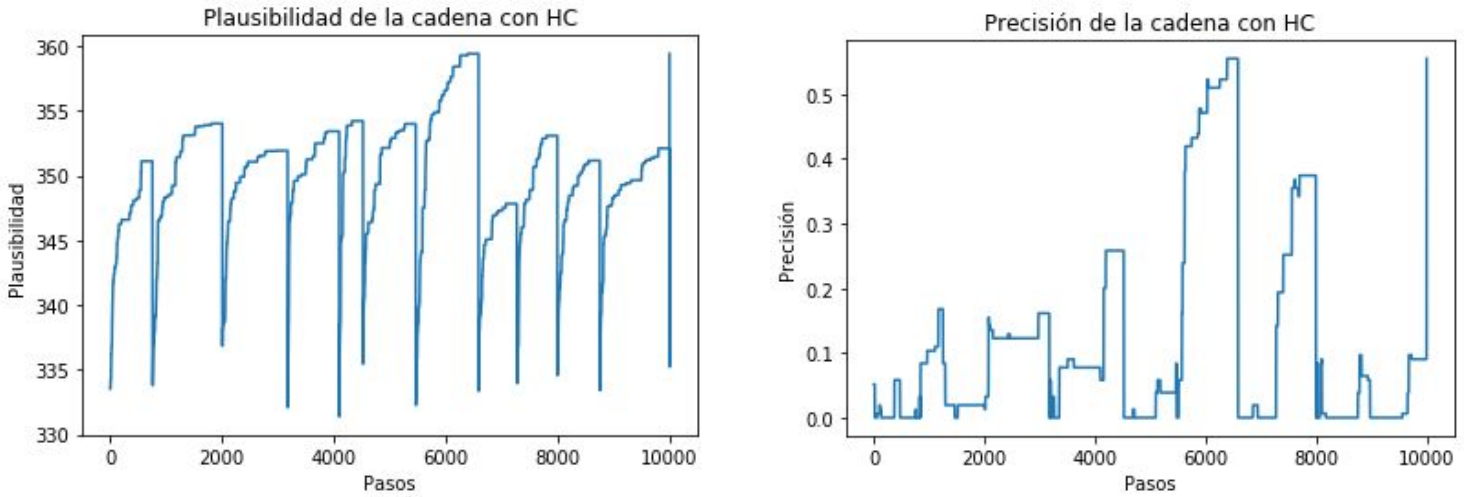


Figura 6: Resultados del ataque de trigrama para HC Backtracking con texto largo

"CUPKOS AZOS DESVUES BREZLE AN VENOLOZ DE BUSINACIEZLO EN POROZEN AURENIAZO FUEZDIA REPORDO AQUENNA LARDE RECOLA EZ QUE SU VADRE NO NNEYO A POZOPER EN KIENO"- máximo de plausibilidad

Resultados comparación entre SA y Benchmark

Texto corto

	Log-plausibilidad	Precisión
SA	360.91 (360.26)	90.8%
HC Backtracking	360.66 (360.26)	88.7%

Texto largo

	Log-plausibilidad	Precisión
SA	649.33 (649.33)	100%
HC Backtracking	649.33 (649.33)	100%

Resultado mensaje cifrado texto largo

“AVMKABV MVSIEV ILISVF FIEBCIJV ZK QXCSK LKCEBCVQXV ZK IMACS ZK GCS EVLKQCKEBVF QCEQDKEBI T BAKF MIAQKSVEI WDCEQK ZK NDSCV ZK ZVF GCS BAKF ODK DE KFQACBVA T UVKBI QXCSKEV IDBVA ZK GIF ZK ZVF ZKQKEIF ZK SCMAVF KEBAK SVF QDISKF ZKFBIQIE FDF EVLKSIF SVF ZKBKQBCLKF FISLINKF JIEIZVAI ZKS UAKGCV XKAAISZK KE GCS EVLKQCKEBVF EVLKEBI…”

Extracto mensaje cifrado

“ROBERTO BOLANO AVALOS SANTIAGO DE CHILE VEINTIOCHO DE ABRIL DE MIL NOVECIENTOS CINCUENTA Y TRES BARCELONA QUINCE DE JULIO DE DOS MIL TRES FUE UN ESCRITOR Y POETA CHILENO AUTOR DE MAS DE DOS DECENAS DE LIBROS ENTRE LOS CUALES DESTACAN SUS NOVELAS LOS DETECTIVES SALVAJES GANADORA DEL PREMIO HERRALDE EN MIL NOVECIENTOS NOVENTA…”

Extracto resultado mensaje descifrado con SA y HC Climbing

Conclusiones

- Ambos métodos logran descifrar mensajes cifrados con sustitución monoalfabética en español
- Ataque con trigrama descifra mejor los mensajes cifrados con ambos métodos.
- Benchmark: Hill Climbing es muy sensible a las condiciones iniciales
- SA tiene mejores resultados que Hill Climbing
- SA tiene resultados cercanos a los obtenidos con Hill Climbing Backtracking

Referencias

- Persi Diaconis. “The Markov Chain Monte Carlo revolution”. En: Bulletin of the American Mathematical Society 46.2 (2009), págs. 179-205.
- Jian Chen y Jeffrey S Rosenthal. “Decrypting classical cipher text using Markov chain Monte Carlo”. En: Statistics and Computing 22.2 (2012), págs. 397-413.
- Luka Bulatovic et al. “Automated cryptanalysis of substitution cipher using hill climbing with well designed heuristic function”. En: Mathematica Montisnigri 44 (2019), págs. 135-143.