Les métaheuristiques en cryptanalyse

Didier Müller

didiermuller.ch

Champéry

11 septembre 2024

Sommaire

Première partie

- Les métaheuristiques
- Hill climbing
- Le recuit simulé
- La recherche avec tabous

Seconde partie

- Application en cryptanalyse
- Exemples

Première partie

Deux métaheuristiques

Les méta heu… quoi ?

Une **métaheuristique** est un algorithme d'**optimisation** visant à résoudre des problèmes difficiles (souvent issus de la recherche opérationnelle) pour lesquels on ne connaît pas de méthode classique plus efficace.

Les métaheuristiques sont généralement des algorithmes **stochastiques itératifs**, qui progressent vers un optimum global, en passant d'une solution à une solution **voisine** (si possible meilleure).

Elles sont souvent inspirées par des systèmes naturels, en physique, en biologie de l’évolution ou encore en éthologie.

Les métaheuristiques les plus connues sont :

- *les algorithmes génétiques* ;
- le recuit simulé ;
- la recherche avec tabous.

Mais il y en a beaucoup d'autres…

Pour commencer,
un peu d'escalade…

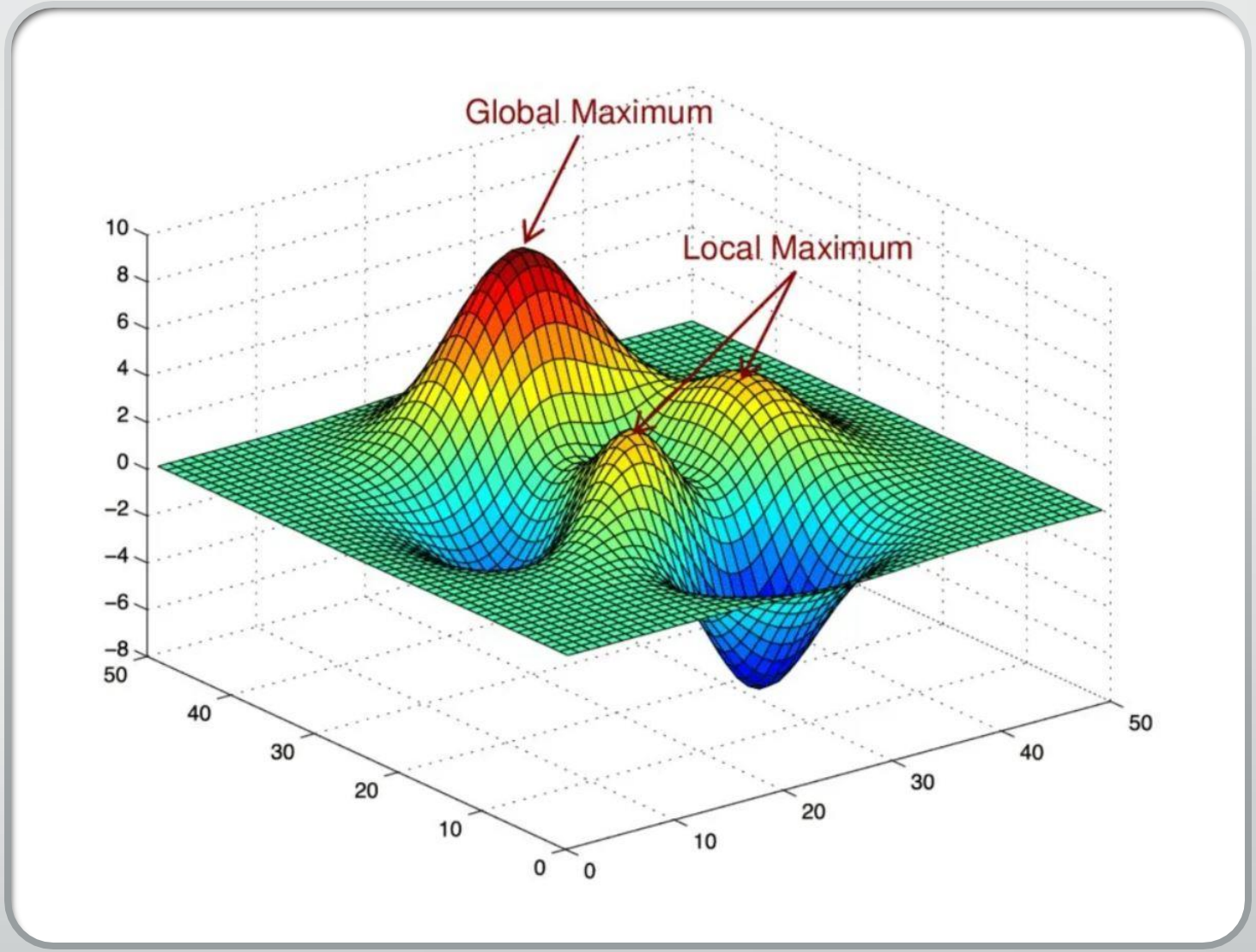
Hill climbing

Principe de cette méthode d'optimisation :

- partir d'une solution initiale aléatoire
- évaluer les solutions voisines et choisir la meilleure
- recommencer l'opération jusqu'à arriver à une solution meilleure que toutes les solutions voisines : c'est un **optimum local**

Problème :

Comment sortir d'un extremum local ?



Le recuit simulé

(Simulated annealing)

Cette méthode a été mise au point par trois chercheurs de la société IBM, **S. Kirkpatrick**, **C.D. Gelatt** et **M.P. Vecchi** en 1983.

Le recuit, en sciences des matériaux, permet de modifier les caractéristiques physiques du métal en le chauffant puis en le refroidissant de manière contrôlée.

Idée générale du recuit simulé

1. A partir de la solution courante, prendre au hasard une solution voisine.
2. Si le résultat est meilleur, garder cette solution et aller en 1.
3. Sinon, garder quand même cette nouvelle solution avec une certaine probabilité, assez grande au début mais qui va diminuer avec le temps, et aller en 1.

Recuit simulé

1. Choisir une « température » de départ T.
2. Générer une solution aléatoire. Appelons-la S_1.
3. Parmi toutes les solutions voisines de S_1, en prendre une au hasard et la nommer S_2.
4. Calculer $\Delta = f(S_2) - f(S_1)$
5. Si $\Delta > 0$, $S_1 \leftarrow S_2$; le cas échéant, mettre à jour f_{max} et la meilleure solution. Aller à 8.
6. Générer un nombre réel aléatoire r dans l'intervalle [0 ; 1].
7. Si $r < \exp(-\Delta/T)$, $S_1 \leftarrow S_2$.
8. Diminuer T toutes les n itérations.
9. Retourner à 3, tant qu'on n'a pas décidé de s'arrêter.

La recherche avec tabous

(tabu search)

La recherche avec tabous a été présentée par **Fred W. Glover** en 1986.

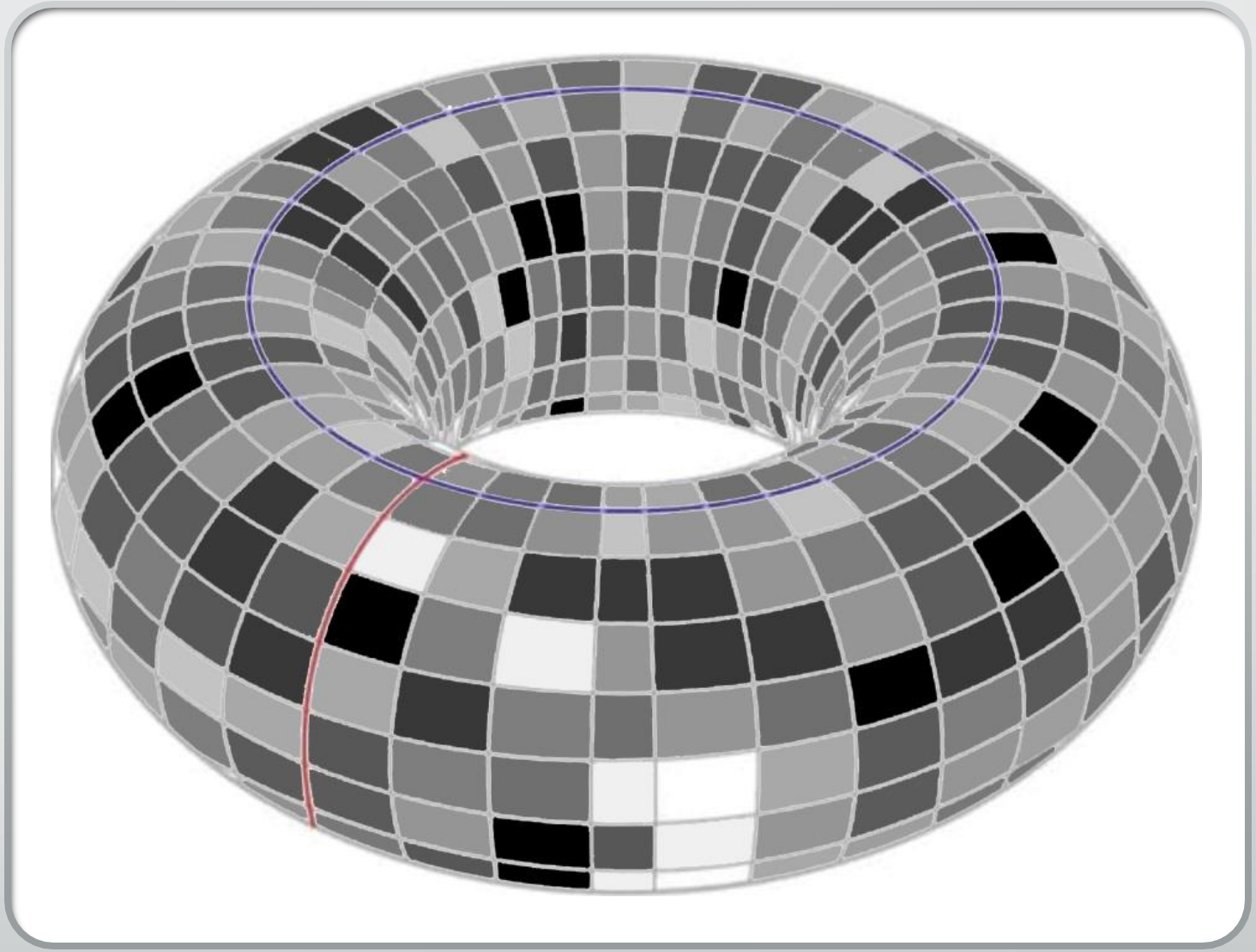
Recherche avec tabous

1. Générer une solution aléatoire.
2. Parmi toutes les solutions voisines, prendre celle qui donne le meilleur résultat et qui n'est pas dans la liste des tabous.
3. Si la liste des tabous est pleine, retirer de cette file la solution la plus ancienne.
4. Mettre la nouvelle solution en queue de la liste des tabous.
5. Retourner à 2, tant qu'on n'a pas décidé de s'arrêter.

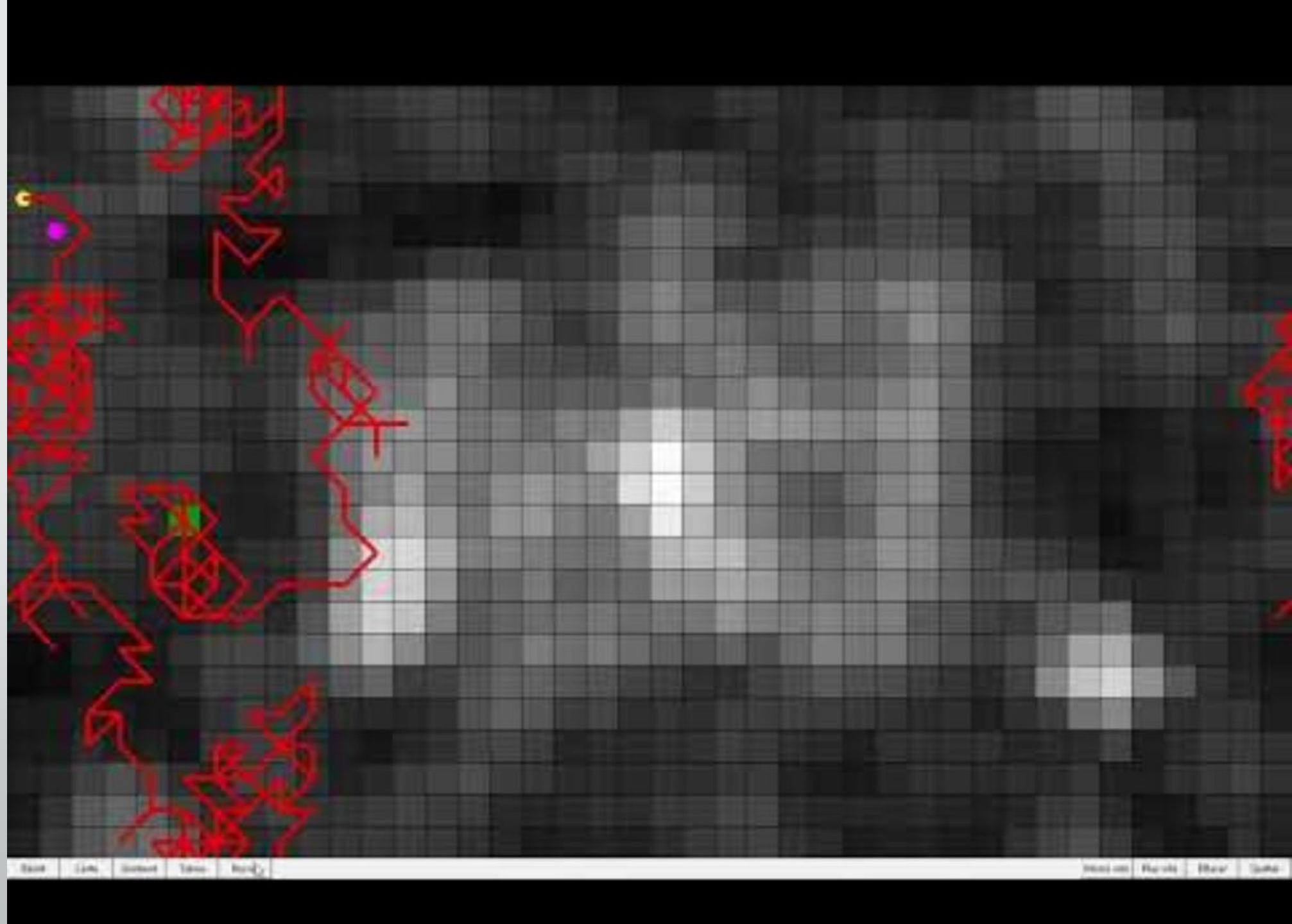
Illustration

Flatland

- Monde en deux dimensions collé sur un tore.
- Chaque zone a un niveau de gris.
- Chaque zone a 8 zones voisines.
- On cherche la zone la plus claire.



Flatland : la démo

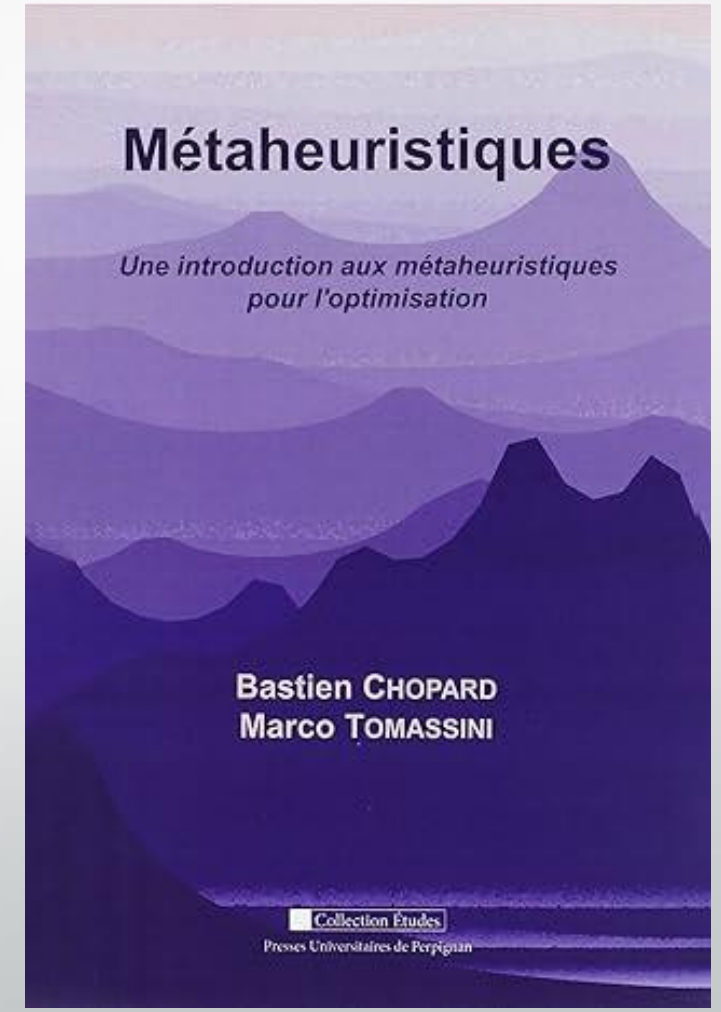


Articles

- D. Müller, "Ruzzle : à la recherche de la plus belle grille", Bulletin no 124 de la SSPMP, janvier 2014
- D. Müller, "Le problème des *n* dames pour illustrer les métaheuristiques", Bulletin no 113 de la SSPMP, juin 2010

Aussi disponibles sur didiermuller.ch

Bibliographie



Seconde partie

Application en cryptanalyse

Questions à se poser

- Comment évaluer une suite de lettres sans signification ?
- Comment lui donner un "score" ?
- Comment définir le voisinage ?

Quelle est la meilleure suite de 42 lettres ?

- ETDJFGIEVLPOPMUIRPVONIMORUENPIEELLOFRERELS
- JUPOERTUBESFRSAERTERSGLNMSQUARLIFAMINESIRI
- URTPOURLILSENMCVTSAPILUSTRELNMSPOEILFSPOIX

Et d'abord, que signifie "meilleure" ?

Comment évaluer une suite de lettres ?

"Score" d'une suite de *n* lettres :

$$\text{score} = -\sum_{i=0}^{n-4} \log\left(\textit{fréquence}\left(c_i c_{i+1} c_{i+2} c_{i+3}\right)\right)$$

$c_i c_{i+1} c_{i+2} c_{i+3}$: 4 lettres consécutives.

Cette formule a été proposée par **Abraham Sinkov** en 1966.

Tétragrammes en français

- Statistiques faites sur 26 textes écrits entre 1831 et 2017
- Sans espace, sans ponctuation
- 10 millions de lettres non accentuées (majuscules)
- 68’183 tétragrammes différents

Tétragrammes	Apparitions
MENT	30207
ELLE	28622
QUEL	23983
EMEN	22824
TION	20768
DANS	20276
IENT	19220
ESDE	18171
DELA	17248
OMME	16947

Application de la formule de Sinkov

Plus le score est petit, plus le texte est français :

- score("KFQTDEGRNWKNIQ") = 94.79783559030287
- score("CESTDUFRBNCBIS") = 77.57382632660952
- score("CESTDURFANCAIS") = 50.39139100671383
- score("CESTDUFRANCAIS") = 45.97301718417951

D'après Sinkov

Quelle est la meilleure suite de lettres ?

ETDJFGIEVLPOPMUIRPVONIMORUENPIEELLOFRERELS

Score = 281.94

JUPOERTUBESFRSAERTERSGLNMSQUARLIFAMINESIRI

Score = 235.63

URTPOURLILSENMCVTSAPILUSTRELNMSPOEILFSPOIX

Score = 232.87

Premier cas :
substitution monoalphabétique

Chiffré																							
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
q	w	e	r	t	z	u	i	o	p	a	s	d	f	g	h	j	k	l	y	x	c	v	b	n	m
Clair																							

Clef : QWERTZUIOPASDFGHJKLYXCVBNM

C'est la clef de déchiffrement.

Clefs voisines

Echanger 2 lettres :

QWERTZUIO**P**ASDFGHJKLY**X**CVBNM
QWERTZUIO**X**ASDFGHJKLY**P**CVBNM

Déplacer 1 lettre :

QWERTZUIO**P**ASDFGHJKLYXCVBNM
QWERTZUIOASDFGHJKLY**P**XCVBNM

Recuit simulé

1. Choisir une « température » de départ T.
2. Générer une clef aléatoire. Appelons-la C_1.
3. Parmi toutes les clefs voisines de C_1, en prendre une au hasard et la nommer C_2.
4. Calculer $\Delta = \text{score}(C_2) - \text{score}(C_1)$
5. Si $\Delta \leq 0$, $C_1 \leftarrow C_2$; le cas échéant, mettre à jour le meilleur score et la meilleure clef. Aller à 8.
6. Générer un nombre réel aléatoire r dans l'intervalle [0 ; 1].
7. Si $r < \exp(-\Delta/T)$, $C_1 \leftarrow C_2$.
8. Diminuer T toutes les n itérations.
9. Retourner à 3, tant qu'on n'a pas décidé de s'arrêter.

Recherche avec tabous

1. Générer une clef aléatoire.
2. Parmi toutes les clefs voisines, prendre celle qui donne le meilleur score et qui n'est pas dans la liste des tabous.
3. Si la liste des tabous est pleine, retirer de cette file la clef la plus ancienne.
4. Mettre la nouvelle clef en queue de la liste des tabous.
5. Retourner à 2, tant qu'on n'a pas décidé de s'arrêter.

Textes pour les tests

- Un extrait du texte sur **Leonhard Euler** tiré de Wikipédia (681 lettres).
- Le poème « **L'albatros** » de Baudelaire (565 lettres).
- Le début de « **1984** ». Traduction de l'anglais (4331 lettres).
- Un extrait de « **La disparition** » de Georges Perec (834 lettres). Pas de E.
- Le début de « **Salammbo** » de Gustave Flaubert (255 lettres). Court et contient des noms propres comme Mégara, Carthage et Hamilcar.
- La fable « **Le cheval et l'âne** » de Jean de la Fontaine (496 lettres). Vieux français.

Un résultat (Euler sur Wikipédia, recuit)

```
Score: 6130.599465914166
TDATLFCGYCUEILGVOGTQYTJIDZTLGTQYVOQYTQYIUVCOTQVDQQCZVLCTQMDTATJVAJDA
------
Score: 5997.105715931042
ADTALFCGYCUEILGVOGAQYAJIDZALGAQYVOQYAQYIUVCOAQVDQQCZVLCAQMDATAJVTJDT
------
Score: 5537.621471325572
EYMELRUOJUXBALOKVOENJEDAYTELOENJKVNJENJAXKUVENKYNNUTKLUENSYEMEDKMDYM

[…]

Score: 2729.173193215509
EULERFITDIMPORTANTESDECOUVERTESDANSDESDOMAINESAUSSIVARIESQUELECALCUL
------
alphabet déchiffrant (0.92 minutes): PBEVDHKMNQTAWZGJCSFIYLORUX
```

Résultats

Les durées de décryptement sont données en minutes.

	Leonhard Euler	L'albatros	1984	La disparition	Salammbo	Le cheval et l'âne
Recuit	1.5	1.3	11.2	2.0	0.6	1.15
Tabous	0.7	0.7	5.2	0.9	0.25	0.5

Commentaires

- La durée de décryptement dépend surtout de la longueur du texte. On peut aller (bien) plus vite en ne prenant que les 300 premières lettres du cryptogramme.
- Le texte décrypté comporte parfois des erreurs mineures, qui n'empêchent pas la compréhension.
- **Fun fact** : dans ce cas, le score du texte décrypté peut être plus petit (et donc "meilleur") que le texte original !
- Cela nous indique que le score de Sinkov avec des tétragrammes est une bonne mesure, mais pas parfaite.

Second cas : le chiffre de Playfair

On chiffre le texte par groupes de 2 lettres (bigrammes).

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
T	U	V	X	Z

NA est chiffré **SD**

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
T	U	V	X	Z

EL est chiffré **DA**

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
T	U	V	X	Z

GV est chiffré **OI**

Clef : CHIFREDPLAYBGJKMNOQSTUVXZ

Chiffrons le mot : ALLUMETTES

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
T	U	V	X	Z

Bigrammes : AL LU ME TT ES

AL LU ME TX TE SC

Cryptogramme : EA DX TY UZ CY MR

Clefs voisines (1)

Echanger 2 lettres (recuit : probabilité de 92%)

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
T	U	V	X	Z

C	H	I	F	R
E	S	P	L	A
Y	B	G	J	K
M	N	O	Q	D
T	U	V	X	Z

Clefs voisines (2)

Echanger 2 colonnes (recuit : probabilité de 2%)

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
T	U	V	X	Z

C	R	I	F	H
E	A	P	L	D
Y	K	G	J	B
M	S	O	Q	N
T	Z	V	X	U

Echanger 2 lignes (recuit : probabilité de 2%)

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
T	U	V	X	Z

T	U	V	X	Z
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
C	H	I	F	R

Clefs voisines (3)

Inverser 1 colonne (recuit : probabilité de 2%)

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
T	U	V	X	Z

C	U	I	F	R
E	N	P	L	A
Y	B	G	J	K
M	D	O	Q	S
T	H	V	X	Z

Inverser 1 ligne (recuit : probabilité de 2%)

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
T	U	V	X	Z

C	H	I	F	R
E	D	P	L	A
Y	B	G	J	K
M	N	O	Q	S
Z	X	V	U	T

Résultats

Taux de réussite (sur 10 essais) et
durées moyennes de décryptement (en minutes).

	Leonhard Euler	L'albatros	1984	La disparition	Salammbo	Le cheval et l'âne
Recuit	7/10	8/10	8/10	4/10	3/10	7/10
	17.2	19.7	20.7	27.1	15.7	20.5

La recherche avec tabous ne donne aucun résultat !

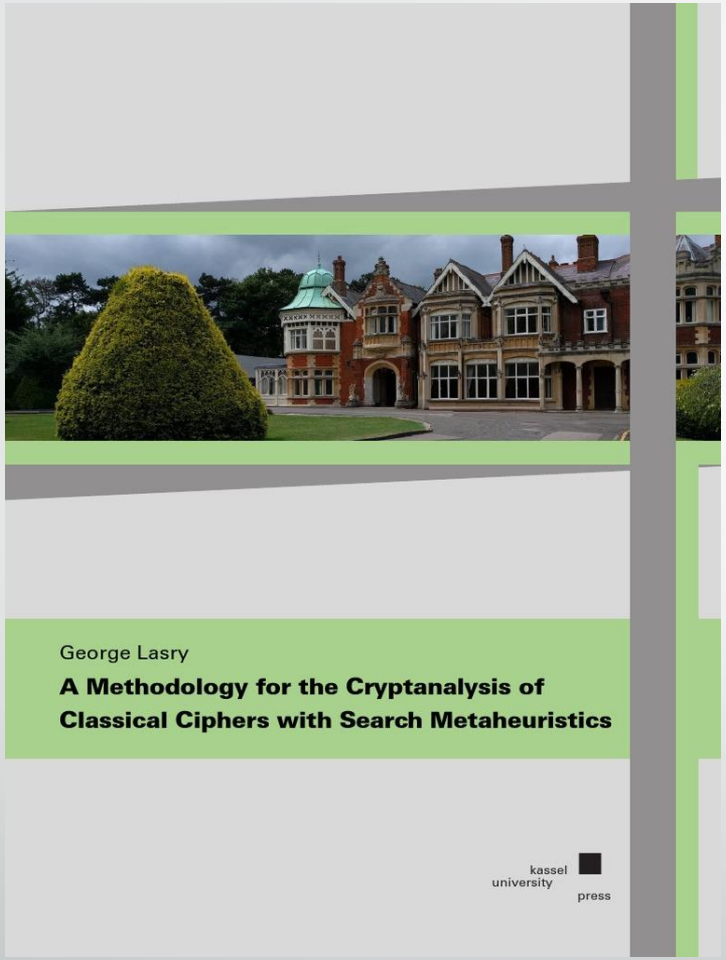
Commentaires

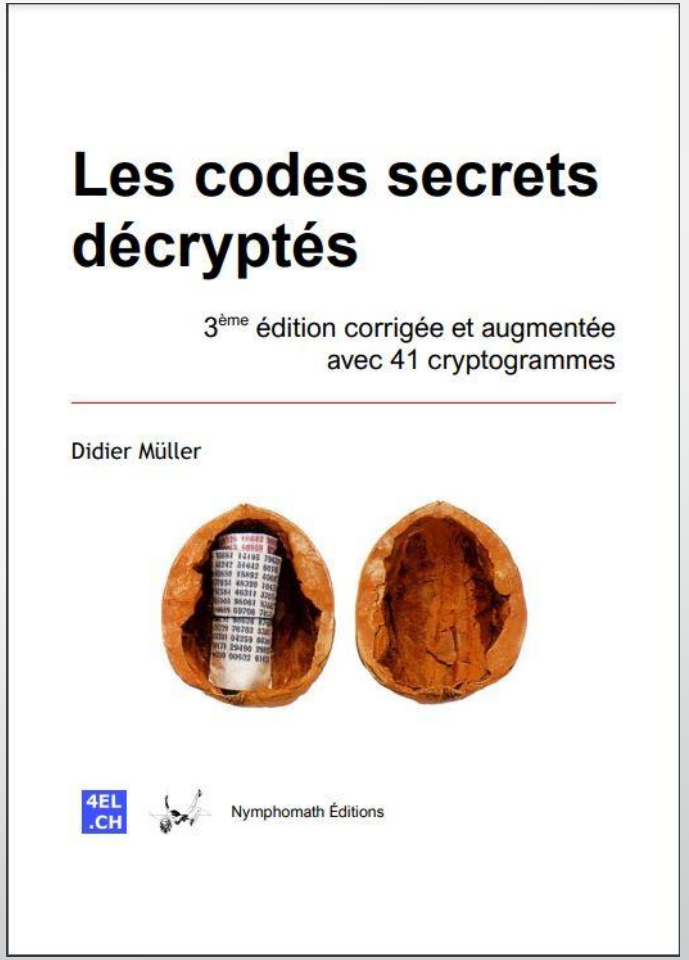
- Les temps de décryptement sont bien plus longs que pour une substitution monoalphabétique, mais cela reste raisonnable.
- Il y a des échecs.
- Certains cryptogrammes (**La disparition** et **Salammbo**) sont assez difficiles à décrypter.
- La recherche avec tabous ne fonctionne pas…

Articles

- Jean-Louis Morel (alias Rossignol), "Décrypter une substitution monoalphabétique", 2015, https://bribes.org/crypto/substitution_mono.html
- S. Reber, "Tentative de décryptement automatique du chiffre de Playfair", Quadrature no 83, Janvier-février-mars 2012, pp. 15-20
- D. Müller, "Les métaheuristiques en cryptanalyse", Bulletin no 143 de la SSPMP, mai 2020, pp. 26-30

Bibliographie





Les deux livres sont disponibles au format PDF.

Pour en savoir plus

Ars cryptographica

www.4el.ch/crypto

Questions ?


QUESTIONS IN ACADEMIC CONFERENCES
A PIE CHART
ATTEMPTS TO LOOK CLEVER
ALREADY ANSWERED ON THE SLIDES
UNCONNECTED RANDOM THOUGHTS
INTERESTING AND ADD TO THE CONFERENCE
ODDLY PERSONAL ATTACKS ON THE SPEAKER
@ERRANTSCIENCE