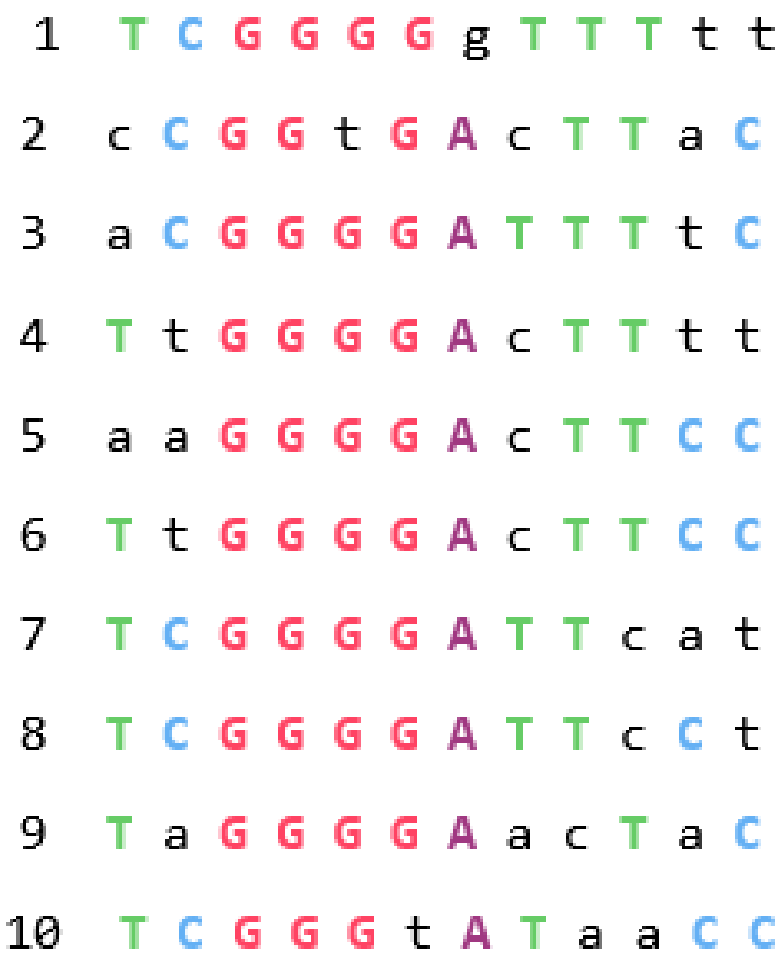
Algoritmos Aleatorios y Motivos Regulatorios

Bioinformática 2025-2

Universidad de Sonora

15 de octubre de 2025

Distintos motivos. TCGGGGATTTCC



Distintos motivos

```
 1 "atgaccgggatactgataaaaaaaaggggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg"
 2 "acccctatttttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataaaaaaaaaggggggga"
 3 "tgagtatccctgggatgacttaaaaaaaaggggggggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga"
 4 "gctgagaattggatgaaaaaaaaggggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga"
 5 "tccctttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaaggggggggcttatag"
 6 "gtcaatcatgttcttgtgaatggatttaaaaaaaaggggggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa"
 7 "cggttttggcccttgttagaggcccccgtaaaaaaaaggggggggcaattatgagagagctaatctatcgcgtgcgtgttcat"
 8 "aacttgagttaaaaaaaaggggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta"
 9 "ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaaggggggggaccgaaagggaag"
10 "ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaaggggggga"
```

Distintos motivos

```
 1 "atgaccgggatactgatAAAAAAAAGGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg"
 2 "acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataAAAAAAAAGGGGGGGa"
 3 "tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga"
 4 "gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga"
 5 "tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag"
 6 "gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa"
 7 "cggttttggcccttgttagaggcccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat"
 8 "aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta"
 9 "ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag"
10 "ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa"
```

Distintos motivos

```
 1 "atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg"
 2 "acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa"
 3 "tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga"
 4 "gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga"
 5 "tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag"
 6 "gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa"
 7 "cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat"
 8 "aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta"
 9 "ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag"
10 "ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa"
```

Nueva definición

Definición. Dada una colección de cadenas ADN y $d, k \in \mathbb{N}$, un k-mero es un (k, d)-motivo si aparece en cada cadena de ADN con a lo más d diferencias.

Motif Enumeration

```
MotifEnumeration(Dna, k, d)
    Patterns ← an empty set
    for each k-mer Pattern in the first string in Dna
        for each k-mer Pattern' differing from Pattern by at most d mismatches
            if Pattern' appears in each string from Dna with at most d mismatches
                add Pattern' to Patterns
    remove duplicates from Patterns
    return Patterns
```

Matriz de motivos

Motifs

T	C	G	G	G	G	g	T	T	T	t	t
c	C	G	G	t	G	A	c	T	T	a	C
a	C	G	G	G	G	A	T	T	T	t	C
T	t	G	G	G	G	A	c	T	T	t	t
a	a	G	G	G	G	A	c	T	T	C	C
T	t	G	G	G	G	A	c	T	T	C	C
T	C	G	G	G	G	A	T	T	c	a	t
T	C	G	G	G	G	A	T	T	c	C	t
T	a	G	G	G	G	A	a	c	T	a	C
T	C	G	G	G	t	A	T	a	a	C	C

Matriz de motivos

Motifs	T	C	G	G	G	G	g	T	T	T	t	t
	c	C	G	G	t	G	A	c	T	T	a	C
	a	C	G	G	G	G	A	T	T	T	t	C
	T	t	G	G	G	G	A	c	T	T	t	t
	a	a	G	G	G	G	A	c	T	T	C	C
	T	t	G	G	G	G	A	c	T	T	C	C
	T	C	G	G	G	G	A	T	T	c	a	t
	T	C	G	G	G	G	A	T	T	c	C	t
	T	a	G	G	G	G	A	a	c	T	a	C
	T	C	G	G	G	t	A	T	a	a	C	C

Score(Motifs) $3 + 4 + 0 + 0 + 1 + 1 + 1 + 5 + 2 + 3 + 6 + 4 = 30$

Perfil de motivos

Motifs		T	C	G	G	G	G	g	T	T	T	t	t
		c	C	G	G	t	G	A	c	T	T	a	C
		a	C	G	G	G	G	A	T	T	T	t	C
		T	t	G	G	G	G	A	c	T	T	t	t
		a	a	G	G	G	G	A	c	T	T	C	C
		T	t	G	G	G	G	A	c	T	T	C	C
		T	C	G	G	G	G	A	T	T	c	a	t
		T	C	G	G	G	G	A	T	T	c	C	t
		T	a	G	G	G	G	A	a	c	T	a	C
		T	C	G	G	G	t	A	T	a	a	C	C
Score(Motifs)		3	+ 4	+ 0	+ 0	+ 1	+ 1	+ 1	+ 5	+ 2	+ 3	+ 6	+ 4 = 30
Count(Motifs)	A:	2	2	0	0	0	0	9	1	1	1	3	0
	C:	1	6	0	0	0	0	0	4	1	2	4	6
	G:	0	0	10	10	9	9	1	0	0	0	0	0
	T:	7	2	0	0	1	1	0	5	8	7	3	4
Profile(Motifs)	A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
	C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
	G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
	T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4

Perfil de motivos

Motifs		T	C	G	G	G	G	g	T	T	T	t	t
		c	C	G	G	t	G	A	c	T	T	a	C
		a	C	G	G	G	G	A	T	T	T	t	C
		T	t	G	G	G	G	A	c	T	T	t	t
		a	a	G	G	G	G	A	c	T	T	C	C
		T	t	G	G	G	G	A	c	T	T	C	C
		T	C	G	G	G	G	A	T	T	c	a	t
		T	C	G	G	G	G	A	T	T	c	C	t
		T	a	G	G	G	G	A	a	c	T	a	C
		T	C	G	G	G	t	A	T	a	a	C	C
Score(Motifs)		3 + 4 + 0 + 0 + 1 + 1 + 1 + 5 + 2 + 3 + 6 + 4 = 30											
Count(Motifs)	A:	2	2	0	0	0	0	9	1	1	1	3	0
	C:	1	6	0	0	0	0	0	4	1	2	4	6
	G:	0	0	10	10	9	9	1	0	0	0	0	0
	T:	7	2	0	0	1	1	0	5	8	7	3	4
Profile(Motifs)	A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
	C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
	G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
	T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4
Consensus(Motifs)		T	C	G	G	G	G	A	T	T	T	C	C

Entropía y logo de motivos

A cada columna de la matriz de motivos le corresponde una distribución de probabilidad, esto es, una colección de números no negativos cuya suma es 1. Una forma de medir la dispersión o incertidumbre de una distribución de probabilidad $(p_1, ..., p_N)$ es con la entropía.

$$H(p_1, ..., p_n) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Entropía y logo de motivos

Motifs	T	C	G	G	G	G	g	T	T	T	t	t
	c	C	G	G	t	G	A	c	T	T	a	C
	a	C	G	G	G	G	A	T	T	T	t	C
	T	t	G	G	G	G	A	c	T	T	t	t
	a	a	G	G	G	G	A	c	T	T	C	C
	T	t	G	G	G	G	A	c	T	T	C	C
	T	C	G	G	G	G	A	T	T	c	a	t
	T	C	G	G	G	G	A	T	T	c	C	t
	T	a	G	G	G	G	A	a	c	T	a	C
	T	C	G	G	G	t	A	T	a	a	C	C

Score(Motifs) $3 + 4 + 0 + 0 + 1 + 1 + 1 + 5 + 2 + 3 + 6 + 4 = 30$

Count(Motifs)												
A:	2	2	0	0	0	0	9	1	1	1	3	0
C:	1	6	0	0	0	0	0	4	1	2	4	6
G:	0	0	10	10	9	9	1	0	0	0	0	0
T:	7	2	0	0	1	1	0	5	8	7	3	4

Profile(Motifs)												
A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4

Consensus(Motifs)	T	C	G	G	G	G	A	T	T	T	C	C

Nuevo problema

Problema. Dada una colección de cadenas, encontrar un conjunto de k-meros, uno de cada cadena, que minimice el puntaje de la matriz de motivos.

Un algoritmo de fuerza bruta requeriría checar todos los k meros posibles de cada cadena, esto es, $(n-k+1)^t$ posibilidades, y para calcular el puntaje, se necesitarían kt pasos, lo cual implica que necesitaríamos $(n-k+1)^t kt$ pasos. Esto es, un algoritmo de búsqueda bruta es de complejidad $O(n^t kt)$.

Nuevo problema

Necesitamos reformular este problema:

Motivos $\rightarrow$ *Consensus*(*Motifs*)
Consensus(*Motifs*) $\rightarrow$ *Motifs*

Dada una colección de k-meros Motivos y un k-mero *Pattern*, definimos $d(Pattern, Motifs)$ como la suma de las distancias de Hamming entre *Pattern* y cada Motivo.

Distancia de Pattern a Motifs

Motifs	T	C	G	G	G	G	g	T	T	T	t	t	3
	c	C	G	G	t	G	A	c	T	T	a	C	4
	a	C	G	G	G	G	A	T	T	T	t	C	2
	T	t	G	G	G	G	A	c	T	T	t	t	4
	a	a	G	G	G	G	A	c	T	T	C	C	3
	T	t	G	G	G	G	A	c	T	T	C	C	2
	T	C	G	G	G	G	A	T	T	c	a	t	3
	T	C	G	G	G	G	A	T	T	c	C	t	2
	T	a	G	G	G	G	A	a	c	T	a	C	4
	T	C	G	G	G	t	A	T	a	a	C	C	+ 3
SCORE(*Motifs*)	3	+ 4	+ 0	+ 0	+ 1	+ 1	+ 1	+ 5	+ 2	+ 3	+ 6	+ 4	= 30
CONSENSUS(*Motifs*)	T	C	G	G	G	G	A	T	T	T	C	C	

Distancia de Pattern a Motifs

$$d(Pattern, Motifs) = \sum_{i=0}^{t} HammingDistance(Pattern, Motifs)$$

$$Score(Motifs) = d(Consensus(Motifs), Motifs)$$

Distancia de Pattern a Motifs

	ttaccttAAC	1
	gATAtctgtc	1
Dna	ACGgcgttcg	2
	ccctAAAgag	0
	cgtcAGAggt	1

Algoritmo Median String

```
MedianString(Dna, k)
    distance ← ∞
    Median ← ""
    for each k-mer Pattern from AA…AA to TT…TT
        if distance > d(Pattern, Dna)
            distance ← d(Pattern, Dna)
            Median ← Pattern
    return Median
```

Calcular probabilidades con el perfil de motivos

$$\text{Profile} = \begin{array}{l} \mathbf{A}: .2\ .2\ 0\ 0\ 0\ 0\ .9\ .1\ .1\ .1\ .3\ 0 \\ \mathbf{C}: .1\ .6\ .0\ 0\ 0\ 0\ 0\ .4\ .1\ .2\ .4\ .6 \\ \mathbf{G}: 0\ 0\ 1\ 1\ .9\ .9\ .1\ 0\ 0\ 0\ 0\ 0 \\ \mathbf{T}: .7\ .2\ 0\ 0\ .1\ .1\ 0\ .5\ .8\ .7\ .3\ .4 \end{array}$$

$$\Pr(\mathbf{ACGGGGATTACC}, \text{Profile}) = .2 \cdot .6 \cdot 1 \cdot 1 \cdot .9 \cdot .9 \cdot .9 \cdot .5 \cdot .8 \cdot .1 \cdot .4 \cdot .6$$

$$= 0.000839808$$

$$\Pr(\mathbf{TCGGGGATTTCC} \mid \textit{Profile}) = 0.7 \cdot 0.6 \cdot 1.0 \cdot 1.0 \cdot 0.9 \cdot 0.9 \cdot 0.9 \cdot 0.5 \cdot 0.8 \cdot 0.7 \cdot 0.4 \cdot 0.6$$

$$= 0.0205753$$

GreedyMotifSearch

```
GreedyMotifSearch(Dna, k, t)
    BestMotifs ← motif matrix formed by first k-mers in each string from Dna
    for each k-mer Motif in the first string from Dna
        Motif_1 ← Motif
        for i = 2 to t
            form Profile from motifs Motif_1, …, Motif_i - 1
            Motif_i ← Profile-most probable k-mer in the i-th string in Dna
        Motifs ← (Motif_1, …, Motif_t)
        if Score(Motifs) < Score(BestMotifs)
            BestMotifs ← Motifs
    return BestMotifs
```

GreedyMotifSearch

```
GreedyMotifSearch(Dna, k, t)
    BestMotifs ← motif matrix formed by first k-mers in each string from Dna
    for each k-mer Motif in the first string from Dna
        Motif_1 ← Motif
        for i = 2 to t
            form Profile from motifs Motif_1, …, Motif_i - 1
            Motif_i ← Profile-most probable k-mer in the i-th string in Dna
        Motifs ← (Motif_1, …, Motif_t)
        if Score(Motifs) < Score(BestMotifs)
            BestMotifs ← Motifs
    return BestMotifs
```

Algunos problemas con GreedyMotifSearch

ttACCTtaac

gATGTctgtc

acgGCGTtag

ccctaACGAg

cgtcagAGGT

Algunos problemas con GreedyMotifSearch

```
A: 1 0 0 0
C: 0 1 1 0
G: 0 0 0 0
T: 0 0 0 1
```

Algunos problemas con GreedyMotifSearch

$$
\text{Profile} = \begin{array}{cccccccccccc}
\mathbf{A}: & .2 & .2 & .0 & .0 & .0 & .0 & .9 & .1 & .1 & .1 & .3 & .0 \\
\mathbf{C}: & .1 & .6 & .0 & 0 & .0 & .0 & .0 & .4 & .1 & .2 & .4 & .6 \\
\mathbf{G}: & .0 & .0 & 1 & 1 & .9 & .9 & .1 & .0 & .0 & .0 & .0 & .0 \\
\mathbf{T}: & .7 & .2 & .0 & .0 & .1 & .1 & .0 & .5 & .8 & .7 & .3 & .4
\end{array}
$$

$$
\text{Pr}(\text{"TCGTGGATTTCC"}, \text{Profile}) = .7 \cdot .6 \cdot 1 \cdot .0 \cdot .9 \cdot .9 \cdot .9 \cdot .5 \cdot .8 \cdot .7 \cdot .4 \cdot .6 = 0
$$

Algunos problemas con GreedyMotifSearch

Motifs

```
T A A C
G T C T
A C T A
A G G T
```

Count(Motifs)

A:	2	1	1	1
C:	0	1	1	1
G:	1	1	1	0
T:	1	1	1	2

Profile(Motifs)

2/4	1/4	1/4	1/4
0	1/4	1/4	1/4
1/4	1/4	1/4	0
1/4	1/4	1/4	2/4

Algunos problemas con GreedyMotifSearch

Count(Motifs)

A:	2+1	1+1	1+1	1+1
C:	0+1	1+1	1+1	1+1
G:	1+1	1+1	1+1	0+1
T:	1+1	1+1	1+1	2+1

Profile(Motifs)

3/8	2/8	2/8	2/8
1/8	2/8	2/8	2/8
2/8	2/8	2/8	1/8
2/8	2/8	2/8	3/8

Suavizado de Laplace

```
GreedyMotifSearch(Dna, k, t)
    form a set of k-mers BestMotifs by selecting 1st k-mers in each string from Dna
    for each k-mer Motif in the first string from Dna
        Motif_1 ← Motif
        for i = 2 to t
            apply Laplace's Rule of Succession to form Profile from motifs Motif_1, …, Motif_i-1
            Motif_i ← Profile-most probable k-mer in the i-th string in Dna
            Motifs ← (Motif_1, …, Motif_t)
            if Score(Motifs) < Score(BestMotifs)
            BestMotifs ← Motifs
    output BestMotifs
```

Algunos problemas con GreedyMotifSearch

ttACCTtaac

gATGTctgtc

acgGCGTtag

ccctaACGAg

cgtcagAGGT

GreedyMotifSearch con suavizado de Laplace

Motifs ACCT

Count(Motifs)

A:	1+1	0+1	0+1	0+1
C:	0+1	1+1	1+1	0+1
G:	0+1	0+1	0+1	0+1
T:	0+1	0+1	0+1	1+1

Profile(Motifs)

2/5	1/5	1/5	1/5
1/5	2/5	2/5	1/5
1/5	1/5	1/5	1/5
1/5	1/5	1/5	2/5

Suavizado de Laplace

gATG	ATGT	TGTc	GTct	Tctg	ctgt	tgtc
$1/5^4$	$4/5^4$	$1/5^4$	$4/5^4$	$2/5^4$	$2/5^4$	$1/5^4$

Suavizado de Laplace

Motifs

ACCT
ATGT

Count(Motifs)

A:	2+1	0+1	0+1	0+1
C:	0+1	1+1	1+1	0+1
G:	0+1	0+1	1+1	0+1
T:	0+1	1+1	0+1	2+1

Profile(Motifs)

3/6	1/6	1/6	1/6
1/6	2/6	2/6	1/6
1/6	1/6	2/6	1/6
1/6	2/6	1/6	3/6

Suavizado de Laplace

acg**G**	cg**GC**	g**GCG**	**GCGT**	**CGT**t	**GT**ta	**T**tag
$12/6^4$	$2/6^4$	$2/6^4$	$12/6^4$	$3/6^4$	$2/6^4$	$2/6^4$

Suavizado de Laplace

Motifs

ACCT
ATGT
acgG

Count(Motifs)

A:	3+1	0+1	0+1	1+1
C:	0+1	2+1	1+1	0+1
G:	0+1	0+1	2+1	1+1
T:	0+1	1+1	0+1	2+1

Profile(Motifs)

4/7	1/7	1/7	1/7
1/7	3/7	2/7	1/7
1/7	1/7	3/7	2/7
1/7	2/7	1/7	3/7

Suavizado de Laplace

ccct	ccta	ctaA	taAC	aACG	ACGA	CGAg
$18/7^4$	$3/7^4$	$2/7^4$	$1/7^4$	$16/7^4$	$36/7^4$	$2/7^4$

Suavizado de Laplace

Motifs

ACCT
ATGT
acgG
ACGA

Count(Motifs)

A:	4+1	0+1	0+1	0+1
C:	0+1	3+1	1+1	0+1
G:	0+1	0+1	3+1	1+1
T:	0+1	1+1	0+1	2+1

Profile(Motifs)

5/8	1/8	1/8	2/8
1/8	4/8	2/8	1/8
1/8	1/8	4/8	2/8
1/8	2/8	1/8	3/8

Suavizado de Laplace

cgtc	gtca	tcag	cag**A**	ag**AG**	g**AGG**	**AGGT**
$1/8^4$	$8/8^4$	$8/8^4$	$8/8^4$	$10/8^4$	$8/8^4$	$60/8^4$

Suavizado de Laplace

	ACCT
	ATGT
Motifs	acgG
	ACGA
	AGGT
Consensus(Motifs)	ACGT

Suavizado de Laplace

Profile

A:	4/5	0	0	1/5
C:	0	3/5	1/5	0
G:	1/5	1/5	4/5	0
T:	0	1/5	0	4/5

Dna

```
ttaccttaac
gatgtctgtc
acggcgttag
ccctaacgag
cgtcagaggt
```

Suavizado de Laplace

```
                                ttaccttaac
                                gatgtctgtc
Motifs(Profile, Dna)            acggcgttag
                                ccctaacgag
                                cgtcaggggt
```

RandomizedMotifSearch

```
RandomizedMotifSearch(Dna, k, t)
    randomly select k-mers Motifs = (Motif_1, …, Motif_t) in each string from Dna
    BestMotifs ← Motifs
    while forever
        Profile ← Profile(Motifs)
        Motifs ← Motifs(Profile, Dna)
        if Score(Motifs) < Score(BestMotifs)
            BestMotifs ← Motifs
            else
            return BestMotifs
```