



Processamento de Linguagem Natural

Distância Mínima de Edição

Prof.: Hansenclever Bassani (Hans) hfb@cin.ufpe.br

Site da disciplina: www.cin.ufpe.br/~hfb/pln/

Baseado nos slides do [curso de Stanford no Coursera](#)
por Daniel Jurafsky e Christopher Manning.

Tradução: Ygor César Sousa
Revisão: Hansenclever Bassani



UNIVERSIDADE
FEDERAL
DE PERNAMBUCO



Distância Mínima de Edição

Definição de Distância Mínima de Edição



O quão similar duas coisas são?

- Correção Ortográfica

- O usuário digitou “graffe”

Qual é o mais próximo?

- graf
 - graft
 - grail
 - giraffe

- Computação Biológica

- Alinhar duas sequências de nucleotídeos

AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC

- Alinhamento resultante:

– AGGCTATCACCTGACCTCCAGGCCGA – – TGCCC – – –
TAG – CTATCAC – – GACCGC – – GGTCGATTTGCCCGAC

- Também para Tradução Automática, Extração de Informação, Reconhecimento de Discurso



Distância de Edição

- Distância mínima de edição entre duas cadeias de caracteres
- É o número mínimo de operações de edição
 - Inserção (i)
 - Deleção (d)
 - Substituição (s)
- Necessário para transformar um no outro



Distância Mínima de Edição

- Duas cadeias de caracteres e seu alinhamento:

I	N	T	E	*	N	T	I	O	N
*	E	X	E	C	U	T	I	O	N



Distância Mínima de Edição

I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s i s

- Se cada operação tem o custo de 1
 - A distância entre eles é 5
- Se substituições custam 2 (Levenshtein)
 - A distância entre eles é 8



Alinhamento em Computação Biológica

- Dadas duas sequencias de bases:

AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTGCCCCGAC

- Alinhar cada letra a uma letra ou lacuna:

–AGGCTATCACCTGACCTCCAGGCCGA–TGCCC– – –
TAG–CTATCAC–GACCGC–GGTCGATTGCCCCGAC



Outros Usos de Distância de Edição em PLN

- Avaliar Tradução Automática e Reconhecimento de Discurso

R Spokesman confirms senior government adviser was shot

H Spokesman said the senior adviser was shot dead

S I D I

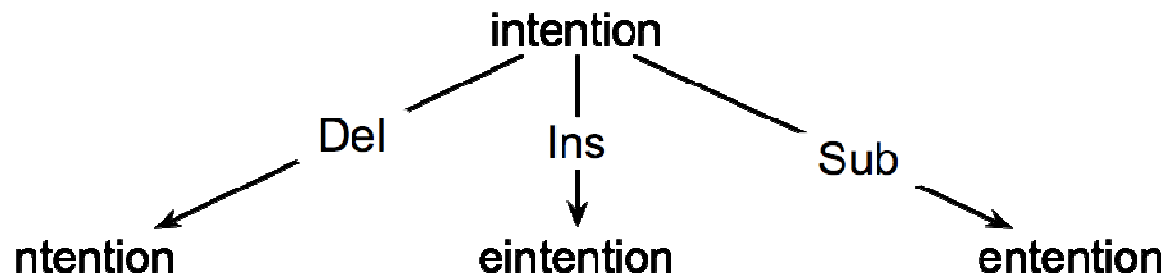
- Extração de Entidade Nomeada e Correferência de Entidade

- IBM Inc. announced today
- IBM profits
- Stanford President John Hennessy announced yesterday
- for Stanford University President John Hennessy



Como Encontrar Distância Mínima de Edição?

- Procurar por um caminho (sequência de edição) da cadeia de caracteres inicial até a final:
 - **Estado Inicial:** a sequência que está sendo transformada
 - **Operadores:** inserção deleção ou substituição
 - **Estado Objetivo:** a sequência que estamos tentando obter
 - **Custo do caminho:** o que nós queremos minimizar: o número de edições





Edição Mínima como Busca

- Mas o espaço de todas as sequências de edição é enorme!
 - Nós não podemos nos dar o luxo de navegar ingenuamente
 - Muitos caminhos distintos acabam no mesmo estado.
 - Nós não precisamos manter todos os caminhos
 - Apenas o caminho mais curto para cada um dos estados visitados.



Definição de Distância Mínima de Edição

- Para duas cadeias de caractere
 - X de tamanho n
 - Y de tamanho m
- Nós definimos $D(i,j)$
 - A distância de edição entre $X[1..i]$ e $Y[1..j]$
 - i.e., os primeiros i caracteres de X e os primeiros j caracteres de Y
 - A distância de edição entre X e Y é, portanto: $D(n,m)$



Distância Mínima de Edição

Definição de Distância Mínima de Edição



Distância Mínima de Edição

Computando Distância
Mínima de Edição



Programação Dinâmica para Distância Mínima de Edição

- **Programação Dinâmica:** Um cálculo tabular de $D(n,m)$
- Resolver problemas combinando soluções para subproblemas.
- Bottom-up
 - Calculamos $D(i,j)$ para i,j pequenos
 - E calculamos $D(i,j)$ maiores baseados nos valores menores computados anteriormente
 - i.e., calcular $D(i,j)$ para todo i ($0 < i < n$) e j ($0 < j < m$)



De onde veio o nome Programação Dinâmica?

...Os anos 50 não foram bons anos para pesquisa matemática. O Secretário de Defesa ...tinha um medo patológico e odiava a palavra pesquisa...

Eu portanto decidi utilizar a palavra, “**Programação**”.

Eu queria passar a ideia de que era dinâmico, multiestágio... eu pensei, vamos ... escolher uma palavra que tem um significado absolutamente preciso, ou seja, **dinâmico**... é impossível usar a palavra, **dinâmico**, em um sentido pejorativo. Tente pensar em alguma combinação que vai dar um sentido pejorativo. É impossível.

Assim, eu pensei que programação dinâmica era um bom nome. Era algo que nem mesmo um congressista poderia opor-se.”

Richard Bellman, “Eye of the Hurricane: an autobiography” 1984.



Definindo Distância Mínima de Edição (Levenshtein)

- Inicialização

$$D(i, 0) = i$$

$$D(0, j) = j$$

- Relação de Recorrência:

For each $i = 1..M$

For each $j = 1..N$

$$D(i, j) = \min \begin{cases} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + \begin{cases} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases} \end{cases}$$

- Término:

$D(N, M)$ é a distância mínima de edição

Tabela de Distância de Edição

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Tabela de Distância de Edição

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
i/j	#	E	X	E	C	U	T	I	O	N

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$




Tabela de Distância de Edição

N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
i/j	#	E	X	E	C	U	T	I	O	N



Distância Mínima de Edição

Computando Distância
Mínima de Edição



Distância Mínima de Edição

Backtrace para Cálculo
de Alinhamentos



Calculo de Alinhamentos

- Distância de Edição não é suficiente
 - Nós frequentemente precisamos **alinhar** cada caractere das duas cadeias, um com o outro
- Nós fazemos isso mantendo um “backtrace”
- Toda vez que entramos em uma célula, lembramos de onde viemos
- Quando chegarmos ao fim,
 - Seguimos o caminho de volta a partir do canto superior direito para ler o **alinhamento**

Distância de Edição

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

MinEdit com Backtrace

n	9	↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↙←↓ 12	↓ 11	↓ 10	↓ 9	↙ 8	
o	8	↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↓ 10	↓ 9	↙ 8	← 9	
i	7	↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↓ 9	↙ 8	← 9	← 10	
t	6	↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙ 8	← 9	← 10	←↓ 11	
n	5	↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↙↓ 10	
e	4	↙ 3	← 4	↙← 5	← 6	← 7	←↓ 8	↙←↓ 9	↙←↓ 10	↓ 9	
t	3	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙ 7	←↓ 8	↙←↓ 9	↓ 8	
n	2	↙←↓ 3	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↓ 7	↙←↓ 8	↙ 7	
i	1	↙←↓ 2	↙←↓ 3	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙ 6	← 7	← 8	
#	0	1	2	3	4	5	6	7	8	9	
	#	e	x	e	c	u	t	i	o	n	



Adicionando Backtrace à Distância Mínima de Edição

- Condições iniciais:

$$D(i, 0) = i$$

$$D(0, j) = j$$

- Término:

$$D(N, M) \text{ is distance}$$

- Relação de Recorrência:

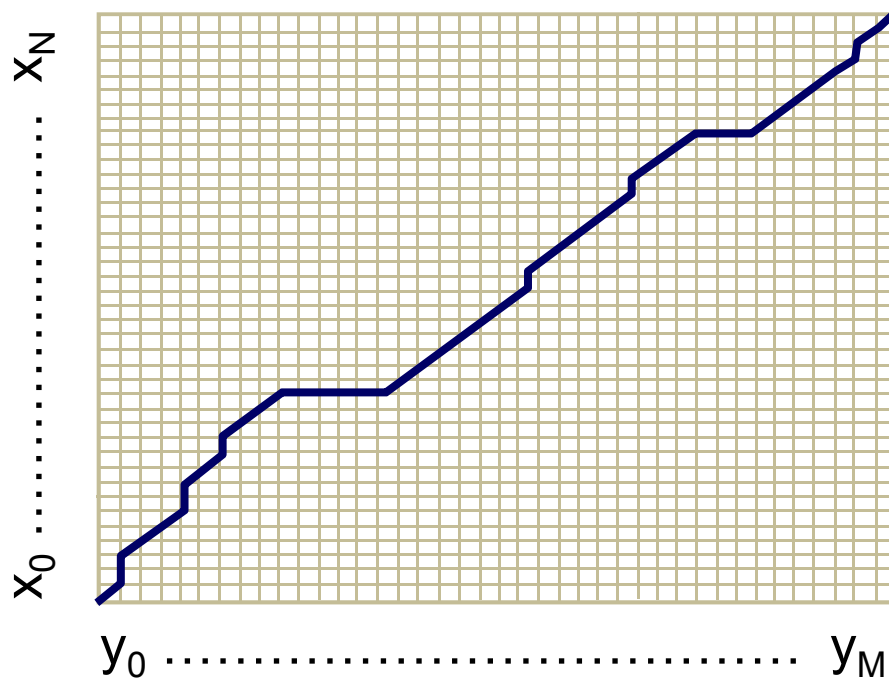
For each $i = 1 \dots M$

For each $j = 1 \dots N$

$$D(i, j) = \min \begin{cases} D(i-1, j) + 1 & \text{deleção} \\ D(i, j-1) + 1 & \text{inserção} \\ D(i-1, j-1) + \begin{cases} 2; & \text{if } X(i) \neq Y(j) & \text{substituição} \\ 0; & \text{if } X(i) = Y(j) & \text{casamento} \end{cases} \end{cases}$$
$$\text{ptr}(i, j) = \begin{cases} \text{LEFT} & \text{insertion} \\ \text{DOWN} & \text{deletion} \\ \text{DIAG} & \text{substitution} \end{cases}$$



Matriz de Distância



Todo caminho não-decrescente

de $(0,0)$ para (M, N)

corresponde a um alinhamento
de duas sequencias

Um alinhamento ótimo é composto
de subalinhamentos ótimos



Resultado do Backtrace

- Duas cadeias de caracteres e seu **alinhamento**:

I	N	T	E	*	N	T	I	O	N
*	E	X	E	C	U	T	I	O	N



Performance

- Time:

$O(nm)$

- Space:

$O(nm)$

- Backtrace

$O(n+m)$



Distância Mínima de Edição

Backtrace para Cálculo
de Alinhamentos



Distância Mínima de Edição

Distância Mínima de Edição com Pesos



Distância de Edição com Pesos

- Por que adicionaríamos pesos ao cálculo?
 - Correção Ortográfica: algumas letras são mais prováveis de serem digitadas incorretamente do que outras
 - Biologia: certos tipos de remoção e inserção são mais prováveis que outros

Matriz de Confusão para Erros de Ortografia

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X	Y (correct)																									
	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
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	5	0	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0





Distância Mínima de Edição com Pesos

- Inicialização:

$$D(0,0) = 0$$

$$D(i,0) = D(i-1,0) + \text{del}[x(i)]; \quad 1 < i \leq N$$

$$D(0,j) = D(0,j-1) + \text{ins}[y(j)]; \quad 1 < j \leq M$$

- Recorrência:

$$D(i,j) = \min \begin{array}{l} D(i-1,j) + \text{del}[x(i)] \\ D(i,j-1) + \text{ins}[y(j)] \\ D(i-1,j-1) + \text{sub}[x(i),y(j)] \end{array}$$

- Término:

$D(N,M)$ é a distância



Distância Mínima de Edição

Distância Mínima de Edição com Pesos



Distância Mínima de Edição

Distância Mínima de
Edição em Computação
Biológica

Alinhamento de Sequências

AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTGCCCCGAC

–AGGCTATCACCTGACCTCCAGGCCGA–TGCCC– –
TAG–CTATCAC–GACCGC–GGTCGATTGCCCCGAC



Por que alinhamento de sequências?

- Comparar genes ou regiões de diferentes espécies
 - Para encontrar regiões importantes
 - Determinar função
 - Encontrar forças evolucionárias
- Montar fragmentos para sequenciar DNA
- Comparar indivíduos à procura de mutações



Alinhamento em Dois Campos

- Em Processamento de Linguagem Natural
 - Nós geralmente falamos de distância (minimizada)
 - E pesos
- Em Computação Biológica
 - Nós geralmente falamos de similaridade (maximizada)
 - E scores



The Needleman-Wunsch Algorithm

A matriz contém scores (match, mismatch, gap)

Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

- Inicialização:

$$D(i, 0) = -i * d \quad (d: \text{penalização})$$

$$D(0, j) = -j * d$$

- Relação de Recorrência:

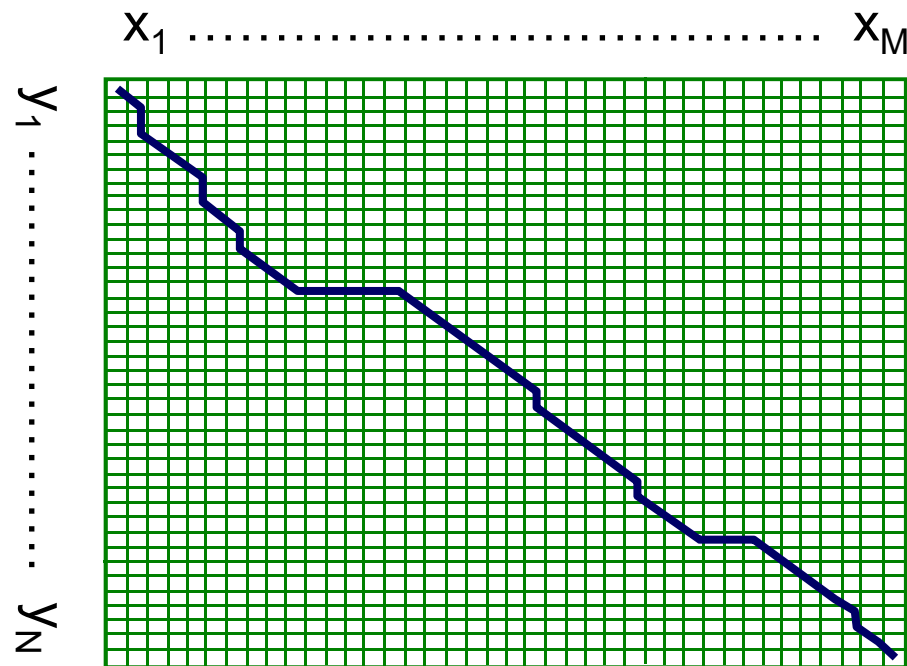
$$D(i, j) = \max \begin{cases} D(i-1, j) & - d \\ D(i, j-1) & - d \\ D(i-1, j-1) + s[x(i), y(j)] & (s: \text{score}) \end{cases}$$

- Término:

$D(N, M)$ is distance



A Matriz Needleman-Wunsch



(Note que a origem é no canto superior esquerdo.)



Uma Variante do Algoritmo Básico:

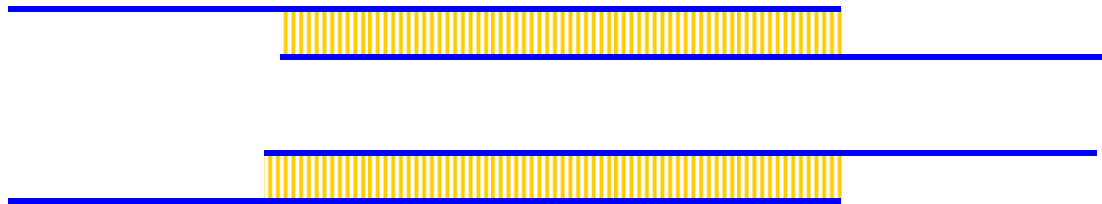
- Talvez esteja correto ter um número ilimitado de lacunas # no começo e no fim:

-----CTATCACCTGACCTCCAGGCCGATGCCCCTTCCGGC
GCGAGTTCATCTATCAC--GACCGC--GGTCG-----

- Sendo assim, nós não queremos penalizar lacunas nas extremidades

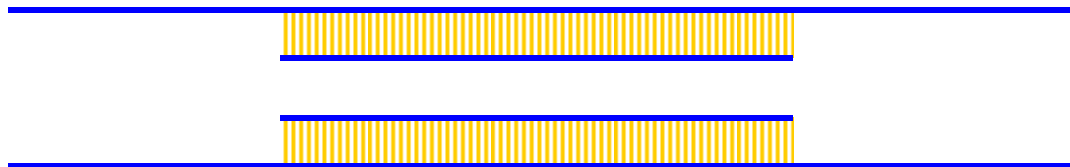


Diferentes tipos de Sobreposições (*Overlap*)



Exemplo:

2 sobreposições “lidas” de um projeto de sequenciamento

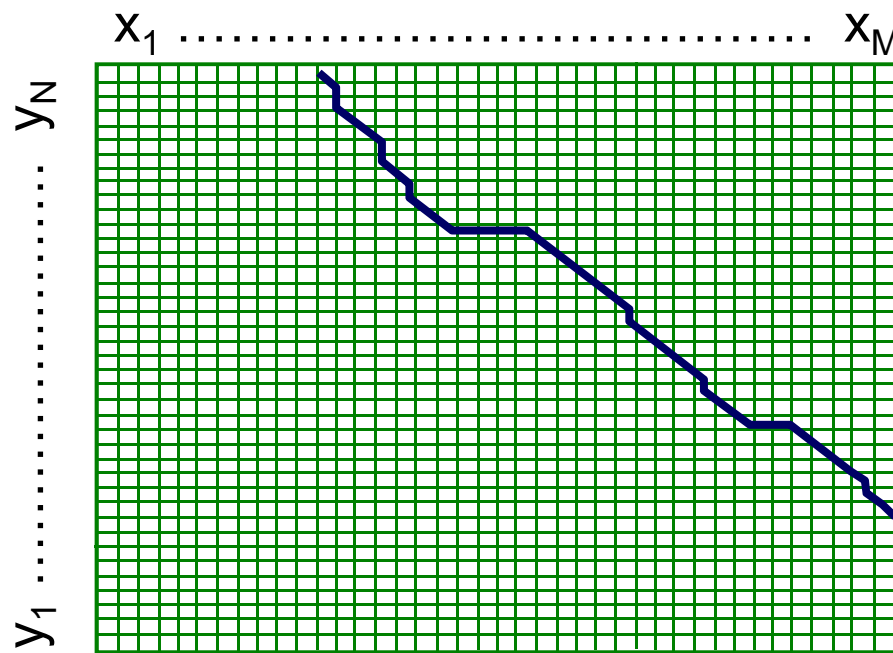


Exemplo:

Procurar um gene do rato dentro de um cromossomo humano



Variante de Detecção de Sobreposição



Mudanças:

1. Inicialização

For all i, j ,
 $F(i, 0) = 0$
 $F(0, j) = 0$

2. Término

$$F_{\text{OPT}} = \max \begin{cases} \max_i F(i, N) \\ \max_j F(M, j) \end{cases}$$



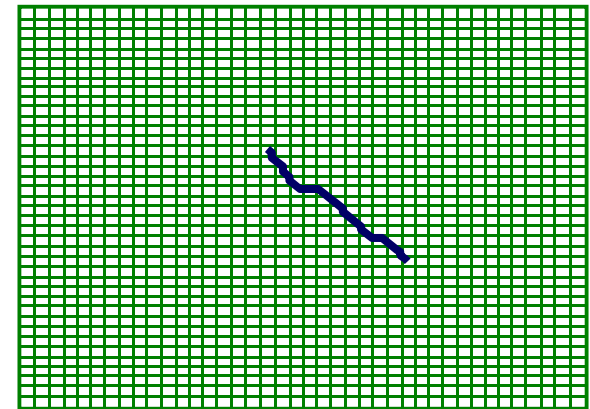
O Problema de Alinhamento Local

Dadas duas strings

$$x = x_1 \dots x_M, \quad y = y_1 \dots y_N$$

Encontrar substrings x' , y' as quais similaridade
(valor ótimo alinhamento global)
é máxima

$x = \text{aaaacc} \boxed{\text{ccgggg}} \text{tta}$
 $y = \text{tt} \boxed{\text{ccgggga}} \text{accaacc}$





O Algoritmo Smith-Waterman

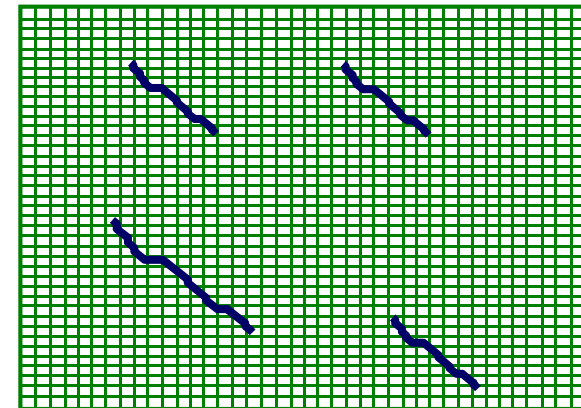
Ideia: Ignorar regiões de mal alinhamento

Modificações ao Needleman-Wunsch:

Inicialização:

$$F(0, j) = 0$$
$$F(i, 0) = 0$$

Iteração :

$$F(i, j) = \max \begin{cases} 0 \\ F(i - 1, j) - d \\ F(i, j - 1) - d \\ F(i - 1, j - 1) + s(x_i, y_j) \end{cases}$$




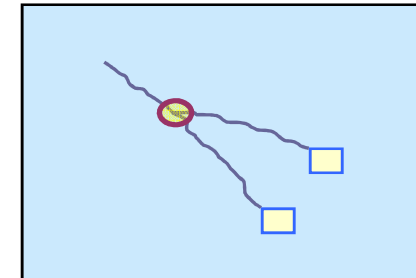
O Algoritmo Smith-Waterman

Término:

1. Se nós queremos o **melhor** alinhamento local...

$$F_{OPT} = \max_{i,j} F(i, j)$$

Encontrar F_{OPT} e fazer caminho de volta (trace back)



2. Se nós queremos **todos** os alinhamentos locais com **score > t**

Para todo i, j encontrar $F(i, j) > t$, e trace back?

Complicada pela sobreposição de alinhamentos locais



Exemplo de Alinhamento Local

X = ATCAT

Y = ATTATC

Seja:

$m = 1$ (1 ponto para combinações)

$d = 1$ (-1 ponto para del/ins/sub)

		A	T	T	A	T	C
	0	0	0	0	0	0	0
A	0						
T	0						
C	0						
A	0						
T	0						



Exemplo de Alinhamento Local

X = ATCAT
Y = ATTATC

		A	T	T	A	T	C
	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0
T	0	0	2	1	0	2	0
C	0	0	1	1	0	1	3
A	0	1	0	0	2	1	2
T	0	0	2	0	1	3	2



Exemplo de Alinhamento Local

X = **ATCAT**

Y = **ATTATC**

		A	T	T	A	T	C
	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0
T	0	0	2	1	0	2	0
C	0	0	1	1	0	1	3
A	0	1	0	0	2	1	2
T	0	0	2	0	1	3	2



Exemplo de Alinhamento Local

X = **ATC**AT

Y = ATT**ATC**

		A	T	T	A	T	C
	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0
T	0	0	2	1	0	2	0
C	0	0	1	1	0	1	3
A	0	1	0	0	2	1	2
T	0	0	2	0	1	3	2



Distância Mínima de Edição

Distância Mínima de
Edição em Computação
Biológica