Algoritmos Paralelos Eficientes para Alguns Problemas em Processamento de Cadeias de Caracteres

Siang Wun Song

JAI/SBC 2007 - Rio de Janeiro
Parte 2

Problema de Subseqüência Máxima

- Dada uma seqüência de caracteres (exemplo: bases do DNA ou amino ácidos).
- A cada caractere ou a cada amino ácido é associado um valor numérico (podendo ser negativo).
- O problema consistem em determinar a subseqüência (contígua) cujos valores numéricos associados tenha soma máxima.
- Este problema aparecem em aplicações de Biologia Computacional, na identificação de domínios transmembranos em proteínas, análise de proteinas e seqüências de DNA, identificação de genes, etc.

Problema de Todas Subseqüências Maximais

- É uma versão estendida do problema básico.
- Depois de obtida a subseqüência máxima, obtém também as subseqüências maximais das partes restantes da seqüência original, assim recursivamente, até que sobrem apenas números não-positivos.

Notações: L(X), R(X)

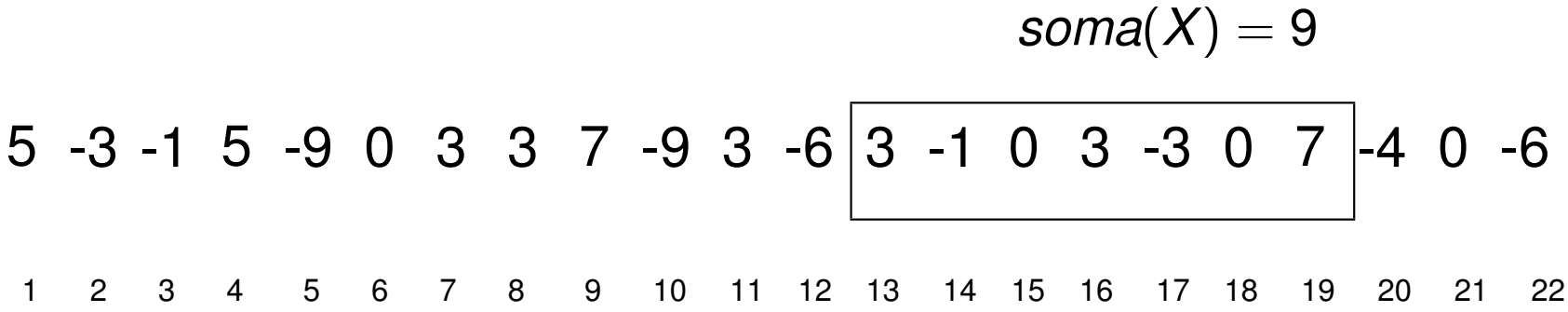
$$X = A_{L(X)}^{R(X)} = A_{12}^{19} = (a_{13}, ..., a_{19})$$

5	-3	-1	5	-9	0	3	3	7	-9	3	-6	3	-1	0	3	-3	0	7	-4	0	-6
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22

Seja a seqüência $A = (a_1, a_2, \ldots, a_n)$.

- Subseqüências de A são indicadas por $A_i^j = (a_{i+1}, ..., a_j)$.
- j indica a posição mais à direita na subseqüência, ao passo que i é um a menos da posição mais à esquerda.
- Seja X uma subseqüência de A. O início e o final de X em A são indicados por $L(X)$ e $R(X)$.
- Para ficar coerente com a notação, temos $X = A_{L(X)}^{R(X)} = (a_{L(X)+1}, ..., a_{R(X)})$.

Notações: Soma



- A soma dos valores de uma subseqüência X é indicada por $soma(X)$.

Notações: Soma de Prefixos

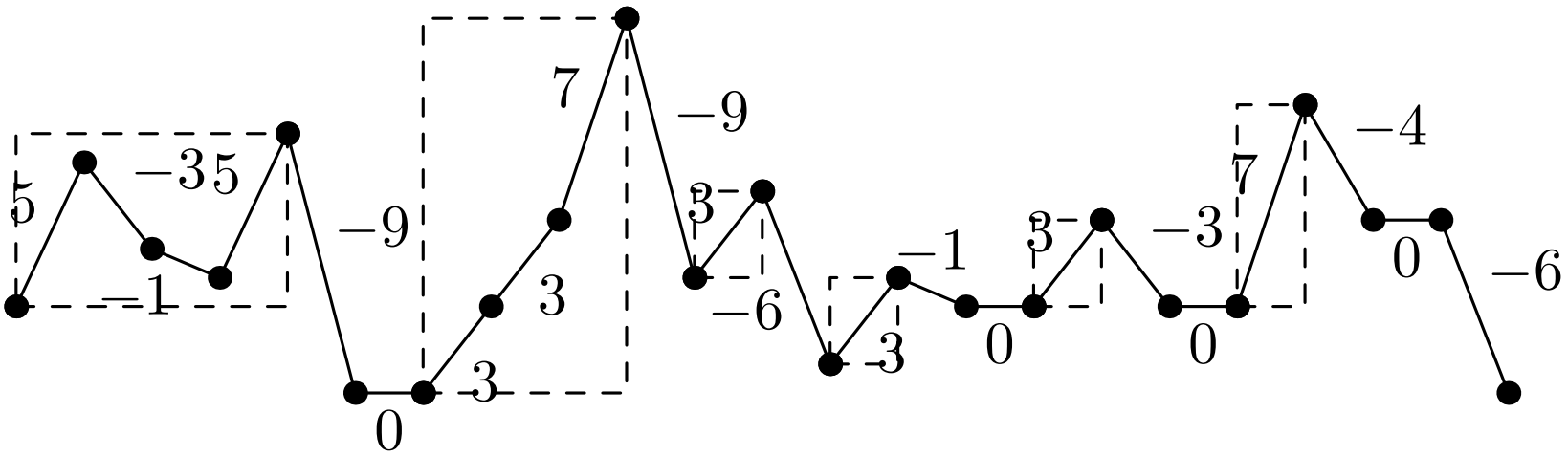
$PS(8) = 3$

5	-3	-1	5	-9	0	3	3	7	-9	3	-6	3	-1	0	3	-3	0	7	-4	0	-6
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22

- A soma de prefixos (*prefix sum* ou *PS*) de A: $PS(j)$ é a soma dos primeiros j elementos de A. Isto é, $PS(j) = soma(A_0^j) = a_1 + a_2 + \ldots + a_j$.
- Note que $soma(A_i^j) = PS(j) - PS(i)$.

Representação Gráfica: Soma de Prefixos

5	-3	-1	5	-9	0	3	3	7	-9	3	-6	3	-1	0	3	-3	0	7	-4	0	-6
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22



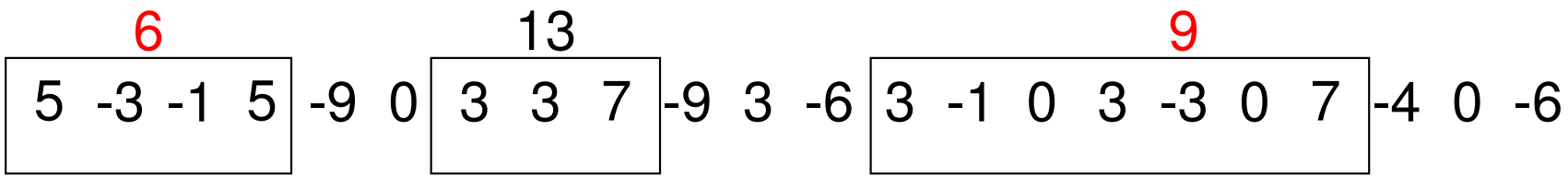
Note que $soma(A_6^9) = PS(9) - PS(6) = 10 - (-3) = 13$.

Exemplo: Dada a Seqüência

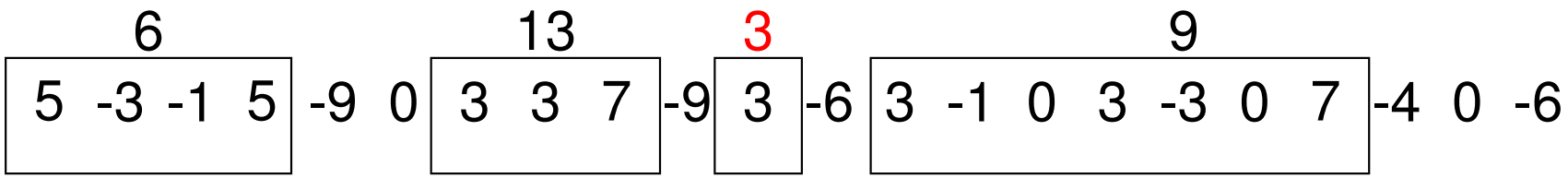
5 -3 -1 5 -9 0 3 3 7 -9 3 -6 3 -1 0 3 -3 0 7 -4 0 -6

Obtenha a Subseqüência Máxima

Obtenha as Demais Subseqüencias Maximais



E Assim Sucessivamente



Uma Outra Forma de Definir o Problema

Ruzzo e Tompa 1999 definem o problema de todas subseqüências maximais de forma seguinte. Indicamos por $soma(X)$ a soma dos valores da subseqüência X.

Uma subseqüência X da seqüência A é maximal em A (chamada A-maximal) se e somente se ela possui as duas propriedades seguintes:

- Propriedade Pr1: Para cada subseqüência própria Y de X, $soma(Y) < soma(X)$.
- Propriedade Pr2: Não há superseqüência própria de X que tem Propriedade Pr1.

Pr1-Subseqüências

Repetimos a definição:
Uma subseqüência X da seqüência A é maximal em A (chamada A-maximal) se e somente se ela possui as duas propriedades seguintes:

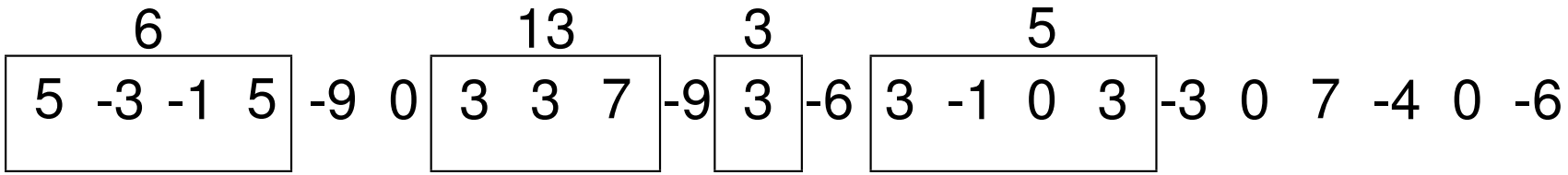
- Propriedade Pr1: Para cada subseqüência própria Y de X, $soma(Y) < soma(X)$.
- Propriedade Pr2: Não há superseqüência própria de X que tem Propriedade Pr1.

Subseqüências de A que têm a Propriedade Pr1 serão denominadas *Pr1-subseqüências*.
Das 4 Pr1-subseqüências, as três primeiras são $A-$ *maximais*.
A última não é pois não satisfaz a segunda prorpriedade.
(Pode ser estendida até 7 para dar uma soma maior.)

Pr1-Subseqüências

Repetimos a definição:
Uma subseqüência X da seqüência A é maximal em A (chamada A-maximal) se e somente se ela possui as duas propriedades seguintes:

- Propriedade Pr1: Para cada subseqüência própria Y de X, $soma(Y) < soma(X)$.
- Propriedade Pr2: Não há superseqüência própria de X que tem Propriedade Pr1.

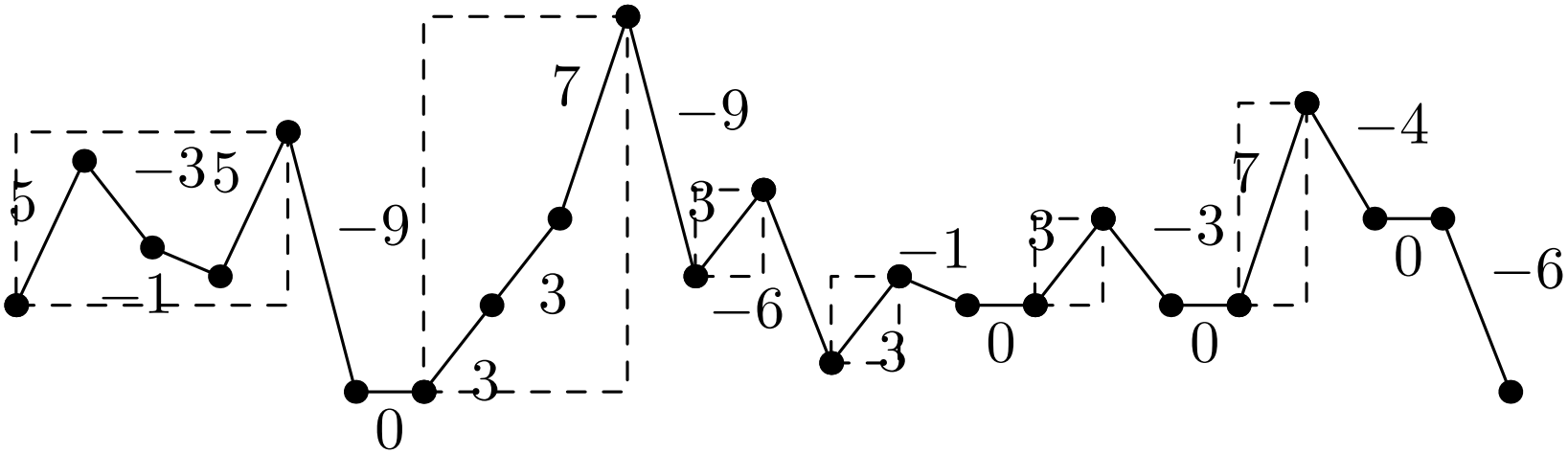


Subseqüências de A que têm a Propriedade Pr1 serão denominadas *Pr1-subseqüências*.
Das 4 Pr1-subseqüências, as três primeiras são $A-$ *maximais*.
A última não é pois não satisfaz a segunda prorpriedade.
(Pode ser estendida até 7 para dar uma soma maior.)

Característica de uma Pr1-Subseqüência

Uma subseqüência A_i^j é Pr1-subseqüência se e somente se para todo m, $i < m < j$, $PS(i) < PS(m) < PS(j)$.

Isto é, a curva da soma de prefixos numa Pr1-Subseqüência fica compreendida entre o seu início e o fim.

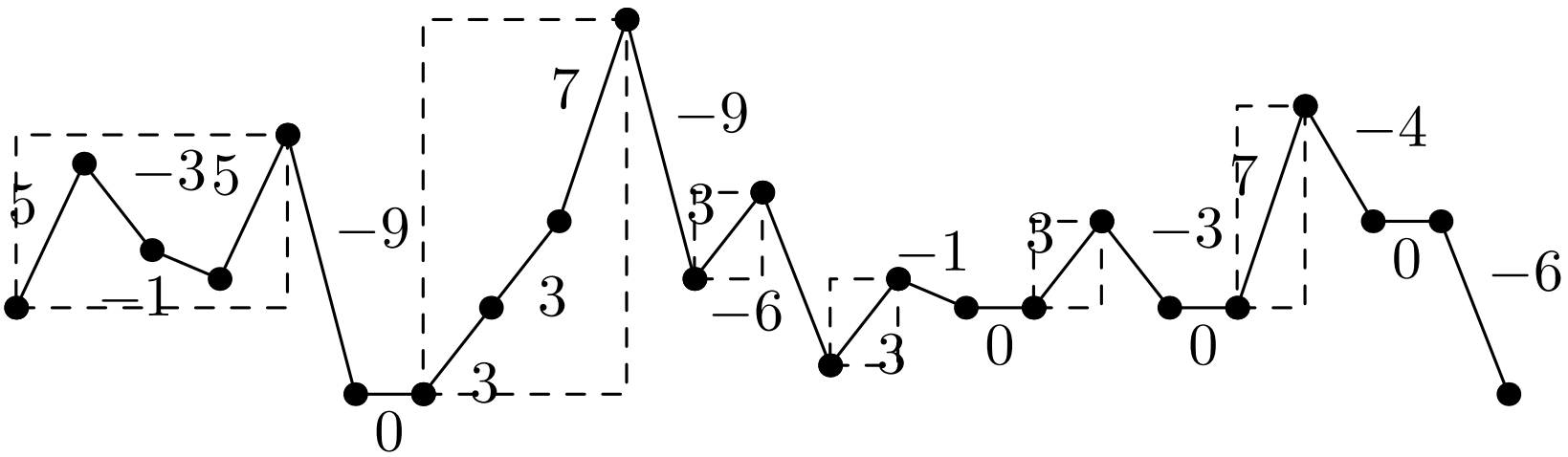


As 3 primeiras Pr1-Subseq. são $A - maximais$. As últimas 3 não são (podem ser estendidas).

Notação: Min e Max de Soma de Prefixos

Para uma subseqüência $X = A_i^j$, o mínimo e o máximo de todos os valores de $PS(k)$, para $i \leq k \leq j$, será denotado por $Min(X)$ e $Max(X)$, respectivamente.

5	-3	-1	5	-9	0	3	3	7	-9	3	-6	3	-1	0	3	-3	0	7	-4	0	-6
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22



Considere a seqüência A acima.
Temos $Min(A) = -3$ e $Max(X) = 10$.

Pr1-Subseqüência

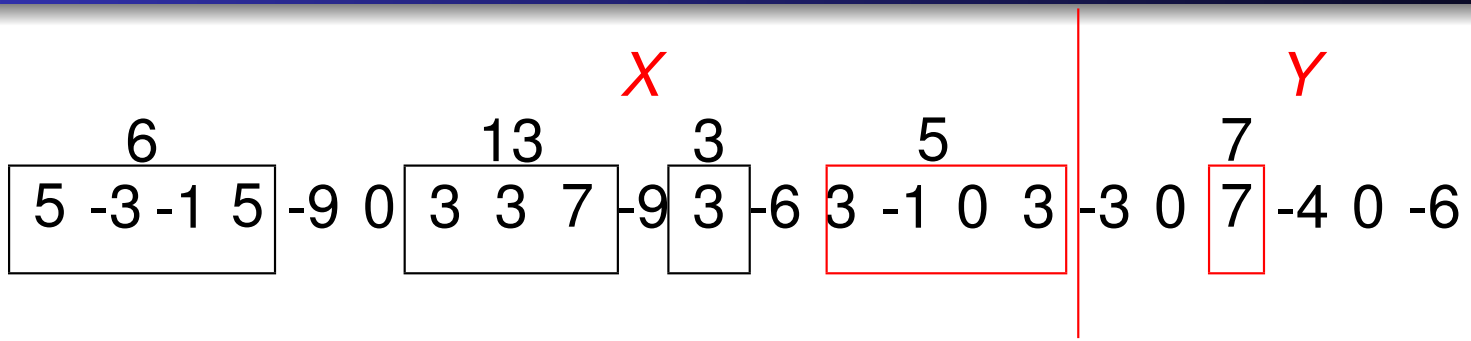
Informalmente, uma Pr1-Subseqüência com soma S é portanto

- Maximal no sentido de não possuir dentro dela uma outra subseqüência com soma maior que S.
- Mas pode haver uma subseqüência que a envolve que tenha soma maior que S.
- Uma Pr1-Subseqüência pode portanto se um prefixo ou sufixo de uma maximal maior.
- Tanto o algoritmo seqüencial como o algoritmo paralelo calculam as Pr1-Subseqüências e depois procurarm juntá-las para formar maximais maiores.

Subseqüências Maximais Não se Sobrepõem

Dada uma seqüência A, quaisquer duas subseqüências maximais distintas de A não se sobrepõem ou tocam uma na outra.

Lema sobre Junção de Pr1-Subseqüências



Seja $Z = \langle X, Y \rangle$ (concatenação de X e Y). Então, há no máximo uma Z-maximal M que sobrepõe ambas X e Y. Se existe tal M, então ela tem uma X-maximal como prefixo e uma Y-maximal como sufixo. As X-maximais à esquerda de M e as Y-maximais à direita de M são também Z-maximais.

- O algoritmo seqüencial concatena simplesmente um outro número a X em cada passo, portanto $|Y| = 1$.
- No algoritmo paralelo, a seqüência A é dividida em subseqüências que são tratadas separadamente em cada processador. Suas subseqüências maximais (locais em cada processador) são usadas mais tarde para obter as A-maximais.

Algoritmo Seqüencial e Algoritmo Paralelo

Vamos agora apresentar os algoritmos seqüencial (de Ruzzo e Tompa 1999) e o algoritmo paralelo (de Alves, Cáceres e Song 2006).