Universidade Federal de Lavras

GCC109 Algoritmos e Estruturas de Dados III

Trabalho III

João Paulo Costa 201220621

Jonhy Geraldo da Silva 201220909

Casamento de Padrões em Sequência de DNA

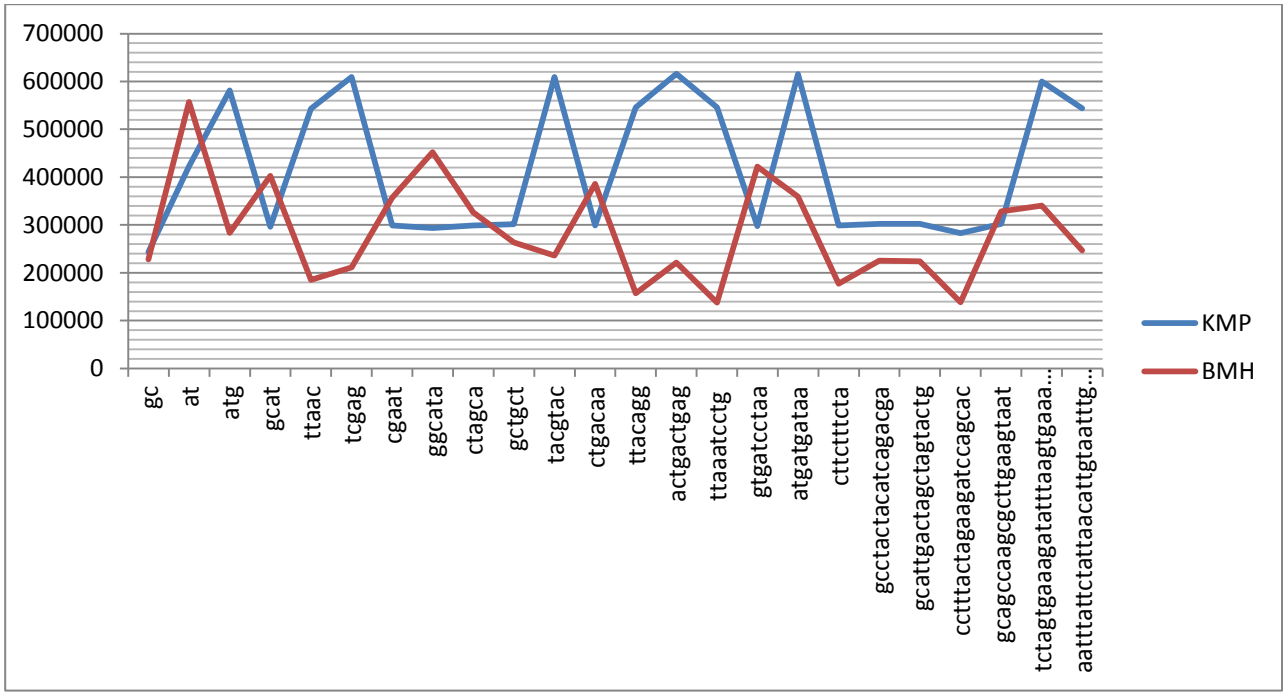
Foram escolhidos para ser implementados, os Algoritmos Boyer-Moore-Horspool – BMH, e Knuth-Morris-Pratt – KMP, esses algoritmos a princípio foram escolhidos porque entendemos que sejam os métodos, mais eficientes para o projeto. Foi interessante também comparar a funcionalidade e eficiência de um algoritmo que casa a sequência por sufixos, BHM , com um que casa por prefixos, KMP.

O algoritmo BMH consiste em pesquisar o padrão P em uma janela que desliza ao longo do texto T. Para cada posição desta janela, o algoritmo faz uma pesquisa por um sufixo da janela que casa com um sufixo de P, com comparações realizadas no sentido da direita para a esquerda. Se não ocorrer uma desigualdade, então uma ocorrência de P em T foi encontrada. Senão, o algoritmo calcula um deslocamento em que o padrão deve ser deslizado para a direita antes que uma nova tentativa de casamento se inicie , esse deslocamento é calculado através de uma tabela em que todas as possíveis ocorrências são preenchidas com um padrão M, e os caracteres presentes no padrão são preenchidos com o numero referente a distância do mesmo com o final do padrão, o ultimo caractere, por exemplo no padrão ABCAB, A = 1, pois esta a uma posição de distância de B, C = 2, pois esta a duas posições de B, e B = 3, o B do final do padrão, não é considerado na contagem por se tratar de ser o ultimo caractere, esses valores da tabela indicaram quanto o padrão irá pular quando o padrão não casar naquele caractere.

O algoritmo KMP, consiste em pesquisar o padrão, através do texto, da esquerda para a direita, ou seja, através do seu prefixo, é criada uma tabela onde cada caractere do padrão se equivale a um valor, esse valor é definido levando em consideração o quantidade de caracteres iguais presentes no padrão, por exemplo, em ABABCB, a tabela começará com os dois primeiros caracteres com 0, já que estão iniciando o padrão, como o próximo caractere é o A, e já existe na tabela, então o seu valor será 1, se os caracteres seguintes já existirem essa contagem será crescente, então o B valerá 2, depois do B tem o caractere C, como ele não existe ainda no padrão o seu valor na tabela volta a ser 0, repetindo esse processo até o fim do padrão.

A pesquisa é feita comparando o padrão com o texto, se n houver nenhum erro até o final do padrão, o casamento foi realizado, caso ocorra uma desigualdade a tabela criada é analisada, e é levada em conta o numero de caracteres já casado, através dessa tabela e que se tem a quantidade de caracteres q pode ser saltado.

Padrão	KMP	BMH
gc	243627	227834
at	422871	557190
atg	581458	283314
gcat	296187	402744
ttaac	542821	185028
tcgag	609063	211595
cgaat	298666	357097
ggcata	293771	452465
ctagca	298774	326614
gctgct	301540	263377
tacgtac	609376	235770
ctgacaa	299285	386179
ttacagg	546038	157068
actgactgag	615804	221613
ttaaatcctg	545666	138034
gtgatcctaa	297490	421818
atgatgataa	615661	358877
cttctttcta	299074	177235
gcctactacatcagacga	302213	225606
gcattgactagctagtactg	302102	224309
cctttactagaagatccagcac	283196	138551
gcagccaagcgcttgaagtaat	302099	328116
tctagtgaaagatatttaagtgaaaatatatacgattaat	600036	340583
aatttattctattaacattgtaatttgctttcgtttgtatattat	543994	247045



Observando o gráfico podemos constatar que o algoritmo KMP realiza menos comparações que o algoritmo BMH em poucos casos, esses casos ficam mais raros a medida que aumentamos o tamanho do padrão de busca. Logo podemos afirmar a partir dos testes feitos que em termos de comparações, o algoritmo BMH é mais eficiente que o algoritmo KMP.