

# KM-Finder: Uma Ferramenta para Detecção de Motivos

Luciana Montera<sup>1</sup>, Lucas Akayama Vilhagra<sup>1</sup>, Tainá Raiol<sup>2</sup>

<sup>1</sup>Faculdade de Computação – Universidade Federal de Mato Grosso do Sul (UFMS)  
Caixa Postal 549. CEP 79070-900. Campo Grande - MS - Brasil

<sup>2</sup>Instituto Leônidas e Maria Deane – Fundação Oswaldo Cruz  
Manaus – AM – Brasil

montera@facom.ufms.br, akayamalucas1@gmail.com, tainaraiol@amazonia.fiocruz.br

**Abstract.** *The identification of patterns within specific regions of the HPV genome may contribute to the understanding of the viral pathogenesis. In this work, a tool for searching nucleotide patterns (motifs) in one or a group of DNA sequences was developed, which continues to be a computational challenge.*

**Resumo.** *A identificação de padrões em regiões específicas nos genomas de HPV pode contribuir para o entendimento da patogênese viral. Neste trabalho, foi desenvolvida uma ferramenta para busca por padrões de nucleotídeos (motivos) em uma ou grupos de sequências de DNA, tarefa que continua a ser um desafio computacional.*

## 1. Introdução

Os papilomavírus humanos (HPV), fatores etiológicos do câncer do colo do útero, são classificados em cutâneos ou mucosotrópicos. O genoma viral codifica proteínas estruturais necessárias à replicação e proteínas codificadas por oncogenes, que estão diretamente associadas com o desenvolvimento de câncer. O genoma viral possui cerca de 7.900 pares de bases (pb), sendo dividido em 8 genes codificadores de proteínas (L1, L2, E1, E2, E4, E5, E6 e E7) e duas regiões não codificadoras: NCR (*Noncoding Region*) e LCR, que se localiza entre L1 e E6, e possui 883 pb. A expressão dos oncogenes virais é controlada pela LCR (*Long Control Region*), que apresenta diversos sítios de ligação para fatores transcricionais celulares e virais [Bernard 2013]. Acredita-se que o *enhancer* epitélio específico, localizado na LCR, contribui para o tropismo celular, propriedade importante na patogênese viral.

O objetivo deste trabalho é a identificação de padrões de nucleotídeos nas LCRs de HPV de um mesmo grupo e, possivelmente, inter-grupos que podem estar relacionados ao tropismo viral. Para isso, foi desenvolvida uma ferramenta de busca e outras duas ferramentas *online* foram utilizadas para comparação. Estão disponíveis 26 genótipos de genomas de HPV cutâneo e 13 de HPV mucosotrópico.

## 2. Fundamentação Teórica

Motivos são pequenas sequências que se repetem ao longo de uma molécula de DNA (ou entre moléculas distintas de DNA) as quais presume-se que tenham alguma função biológica [D'haeseleer 2006]. [Rajasekaran 2005] define três versões para o problema da busca por motivos em sequências biológicas:

**Definição 1** *Planted( $l,d$ )-Motif Search Problem* - Dadas  $n$  sequências de mesmo tamanho e dois inteiros  $l$  e  $d$ , o problema consiste em encontrar o motivo de maior score de tamanho  $l$  presente em todas as sequências  $n$ . Variações do motivo também são retornadas desde que difiram do motivo por uma distância de Hamming [Hamming 1950] não superior a  $d$ .

**Definição 2** *Edited Motif Search Problem* - Dadas  $n$  sequências e três inteiros  $l$ ,  $d$  e  $q$ , são buscados todos os padrões das sequências de entrada de tamanho  $l$  que ocorram em pelo menos  $q$  das  $n$  sequências. Um padrão  $U$  é considerado uma ocorrência de outro padrão  $V$  se a distância de edição [Levenshtein 1966] entre eles é no máximo  $d$ .

**Definição 3** *Simple Motif Search Problem* - Dadas  $n$  sequências e um inteiro  $l$ , padrões definidos como uma string de símbolos juntamente com o caracter "?" (curinga) são buscados nas sequências. Um padrão não começa ou termina com "?". O objetivo é identificar todos os padrões de tamanho máximo  $l$ , com número de curingas podendo variar de 0 a  $\lfloor \frac{l}{2} \rfloor$ , bem como o número de ocorrências destes padrões.

Na literatura existem diversos algoritmos de busca por motivos, dentre eles [Pevzner and Sze 2000], [Buhler and Tompa 2001], [Adebiyi and Kaufmann 2002] e [Dinh et al. 2012].

### 3. Ferramentas e Resultados

Ferramentas *online*, tais como SMILE [Marsan and Sagot 2000] e MEME [Bailey and Elkan 1994] realizam a busca por motivos em sequências biológicas. Neste trabalho, SMILE<sup>1</sup> e MEME<sup>2</sup>, bem como uma ferramenta própria<sup>3</sup> denominada KM-Finder (*K-mer and Motif Finder*) foram utilizadas para buscar por padrões nas LCRs de HPV.

SMILE implementa um algoritmo combinatório baseado na construção de uma árvore de sufixo generalizada enquanto MEME [Bailey et al. 2009] é uma suíte de ferramentas que além de permitir realizar a busca por motivos permite a comparação de novos motivos encontrados com motivos já conhecidos e predição de função biológica.

A implementação proposta neste trabalho, KM-Finder, inicialmente determina grupos de *substrings* de tamanho  $k$  (*k-mers*), comparando cada *k-mer* e adicionando ao grupo os que possuem distância de Hamming de no máximo  $m$  com relação ao *k-mer* representante. Os grupos que não tiverem ocorrências em no mínimo  $q$  e no máximo  $Q$  sequências do conjunto em estudo, são descartados. Para cada um destes grupos, são construídos *motif-profiles*, que são matrizes de frequências de cada nucleotídeo presente nos *k-mers* deste grupo. A concatenação dos nucleotídeos mais frequentes resulta em um candidato a motivo [Jones and Pevzner 2004]. A qualidade desse candidato é calculada fazendo-se a média aritmética dos valores dos nucleotídeos mais frequentes e, somente os motivos que tiverem qualidade superior a  $p$  ( $0 < p \leq 1$ ) serão mostrados na saída. Para o *motif-profile* mostrado na Figura 1(b) a qualidade é 0.81. A Figura 1 mostra como o motivo GCGACCGA foi determinado dado a busca no conjunto de LCR de mucosa com  $k = 8$ ,  $q = 13$ ,  $Q = 13$ ,  $m = 2$  e  $p = 0.8$ . Na Figura 1(a) são mostrados 5 dos

<sup>1</sup>disponível em [mobyli.pasteur.fr/cgi-bin/portal.py#forms::smile](http://mobyli.pasteur.fr/cgi-bin/portal.py#forms::smile)

<sup>2</sup>disponível em [meme-suite.org/tools/meme](http://meme-suite.org/tools/meme)

<sup>3</sup>disponível em [pintado.facom.ufms.br/hpv](http://pintado.facom.ufms.br/hpv)

23 elementos do grupo de *k-mers* utilizados para construir o *motif-profile* apresentado na Figura 1(b).

GCGACGA	Posição 1	Posição 2	Posição 3	Posição 4	Posição 5	Posição 6	Posição 7	Posição 8
GGGACCGA	0.04	0	0.39	0.87	0.04	0	0	0.43
GCAACCGT	0.04	0.83	0.04	0.13	0.87	1	0	0.26
GCGACCGC	0.91	0.17	0.57	0	0.09	0	1	0.13
GCAACCGA	0	0	0	0	0	0	0	0.17
GCAACCGG								

(a) Grupo de *k-mers*

(b) *motif-profile*

**Figura 1. Grupo de *k-mers* e respectiva *motif-profile* que resulta no motivo GC-GACCGA.**

A fim de comparar as três ferramentas citadas, as buscas foram realizadas com parâmetros iguais ou similares. Os tamanhos de motivos buscados variaram entre 6 e 19. Para o grupo de HPV cutâneo na busca por motivos de tamanho 6, KM-Finder e SMILE encontraram a sequência AATAAA, enquanto MEME, além de encontrar essa sequência, encontrou também TGCCAA. Buscando por motivos de tamanho 19, KM-Finder encontrou a sequência AGCGACCGATTTCCGGTACC, enquanto MEME encontrou GATTGTTGCCAACAATCAT e SMILE retornou um resultado inesperado com 74.325 motivos. Para o grupo de HPV mucosotrópico, buscando motivos de tamanho 12, SMILE, KM-Finder e MEME encontraram a sequência ACCGATTTCCGGT. SMILE encontrou ainda outras 8 sequências e MEME encontrou ACCGAAAACGGT e AACCGAAATCCGG.

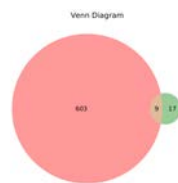
As divergências nos resultados vem do fato das ferramentas implementarem algoritmos diferentes, sendo assim, para uma mesma entrada, as saídas podem ser diferentes. Pode-se notar que quanto menos conservados são os motivos, mais os resultados apresentados pelas ferramentas divergem.

O tempo de resposta do MEME foi o melhor entre as ferramentas testadas e a apresentação dos resultados, por meio da construção de logos e outros gráficos, facilita a análise. SMILE responde rapidamente buscando motivos com tamanho até 16, porém, para tamanhos superiores, demora várias horas. Além da diferença no algoritmo implementado, um importante diferencial do KM-Finder é a possibilidade de busca simultânea em dois grupos distintos, pois, no caso do HPV, deseja-se observar possíveis semelhanças e diferenças entre cada grupo.

A Figura 2 apresenta o Diagrama de Venn resultante da busca por motivos para  $k = 9$ ,  $m = 2$  e  $p = 0.8$ . Foram encontrados 603 motivos exclusivos de mucosa e 17 exclusivos de cutâneo enquanto que 9 motivos estão presentes em ambos os grupos. Padrões que caracterizem cada grupo de vírus podem ser utilizados como método de diagnóstico para identificar a origem de determinado tumor, além de contribuir para o conhecimento sobre a patologia da doença.

#### 4. Conclusão e Trabalhos Futuros

O prognóstico de uma neoplasia depende do tecido afetado que, no caso de tumores causados pela infecção por HPV, pode ter origem no epitélio ou mucosa. Neste trabalho, foi



**Figura 2. Diagrama de Venn resultante da busca por motivos em sequências de HPV de mucosa (rosa) e cutâneo (verde).**

possível identificar padrões únicos que diferem entre os grupos de HPV, indicando potenciais marcadores para diagnóstico. Este estudo, ainda, gerou dados preliminares para a futura identificação de potenciais motivos envolvidos no tropismo viral. Para tal, existem ferramentas, como TFBind<sup>4</sup>, cujas análises podem ser incorporadas na ferramenta proposta. É possível ainda melhorar o algoritmo pela implementação de métodos mais sofisticados para mensurar a qualidade dos motivos a fim de tentar reduzir ainda mais falsos positivos e implementar outras representações gráficas para auxiliar no entendimento dos resultados.

## Referências

- Adebiyi, E. F. and Kaufmann, M. (2002). Extracting common motifs under the levenshtein measure: theory and experimentation. pages 140–156. Proc. Workshop on Algorithms for Bioinformatics (WABI).
- Bailey, T. L., Boden, M., Buske, F., Frith, M., vGrant, C., and et al. (2009). Meme suite: tools for motif discovery and searching. pages 202–208. Nucleic Acids Research.
- Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. pages 28–36. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology.
- Bernard, H.-U. (2013). Regulatory elements in the viral genome. *Virology*, 445(1):197–204.
- Buhler, J. and Tompa, M. (2001). Finding motifs using random projections. pages 269–278. Proc. Fifth Annual International Conference on Computational Molecular Biology (RECOMB).
- D’haeseleer, P. (2006). What are dna sequence motifs? pages 423–425. Nature Biotechnology.
- Dinh, H., Rajasekaran, S., and Davila, J. (2012). qpms7: A fast algorithm for finding (l,d)-motifs in dna and protein sequences. In *PLoS ONE*.
- Hamming, R. W. (1950). Error detecting and error correcting codes. The Bell System Technical Journal.
- Jones, N. C. and Pevzner, P. A. (2004). *An Introduction to Bioinformatics Algorithms*. The MIT Press.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics-Doklady.
- Marsan, L. and Sagot, M.-F. (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. pages 345–362. Journal of Computational Biology.
- Pevzner, P. and Sze, S. H. (2000). Combinatorial approaches to finding subtle signals in dna sequences. pages 269–278. Proc. Eighth International Conference on Intelligent Systems for Molecular Biology.
- Rajasekaran, S. (2005). *Algorithms for Motif Search in Handbook of Computational Molecular Biology*. Chapman and Hall/CRC. chapter 37.

<sup>4</sup>disponível em [tfbind.hgc.jp](http://tfbind.hgc.jp)