Universidade de São Paulo
Escola de Artes, Ciências e Humanidades

ANÁLISE DE MODELOS COMPUTACIONAIS BASEADOS EM POSITION WEIGHT MATRIX PARA A CARACTERIZAÇÃO DE FATORES DE TRANSCRIÇÃO EM *Drosophila melanogaster*

Orientado:

Lauro Hiroshi Pimentel Masuda

Orientador:

Prof. Luiz Paulo Andrioli

1 Introdução

Os fatores de transcrição (FTs) são proteínas que medeiam o processo transcricional ligando-se em trechos específicos do DNA, conhecidos por sítios de ligação de fatores de transcrição (SLFTs). Esses sítios possuem tipicamente de cinco a trinta nucleotídeos e representam sequências de nucleotídeos relativamente bem conservadas, mas que permitem algumas variações (STEWART; HANNENHALLI; PLOTKIN , 2012). Os SLFTs fazem parte dos módulos cis-reguladores da transcrição dos genes, também conhecidos por *enhancers* (LEON; DAVIDSON , 2009). Ao se ligarem nos SLFTs, os FTs modulam a expressão gênica, podendo estimular ou inibir o processo da transcrição, interagindo com o complexo transcricional. Os SLFTs normalmente aparecem próximos uns aos outros, formando agrupamentos, constituindo assi os módulos cis-reguladores (CRMs - *cis-regulatory modules*).

Existem ferramentas computacionais para a predição de SLFTs. A técnica mais utilizada para a caracterização dos motivos é a PWM (*Position Weight Matrix*) (STORMO et al. , 1982), que resumidamente é uma matriz que representa, de forma probabilística, as sequências conhecidas (chamadas de sequências de treinamento) para um dado FT. A PWM é então utilizada para analisar sequências genômicas e, por meio da atribuição de um escore probabilístico, predizer novos locais de SLFTs.

Porém, um dos problemas enfrentados pelos pesquisadores que precisam utilizar essas PWMs para a predição de SLFTs, é que podem existir várias e distintas PWMs para um mesmo FT, cada uma derivada de um conjunto de sequências de treinamento obtidas por distintas técnicas de biologia molecular que não permitem juntar todas as sequências em um único conjunto, como por exemplo, *footprinting*, ChIP-chip ou ChIP-seq (técnicas de imunoprecipitação da cromatina seguida de reconhecimento por chip ou sequenciamento), SELEX (GOLD et al. , 1995) e PBM (*Protein Binding Microarrays*) (BERGER et al, 2006). De posse dessa variedade de PWMs para um mesmo FT, a escolha de qual delas utilizar não é clara.

2 Justificativa

Este projeto é continuidade da colaboração entre os grupos de Genética do Desenvolvimento e de Bioinformática respectivamente sob responsabilidade dos professores Luiz Paulo Andrioli e Ariane Machado-Lima. Os grupos investigam mecanismos de regulação da expressão gênica utilizando a cascata de segmentação responsável pela padronização do corpo no organismo modelo *Drosophila melanogaster* como objeto de estudo. Essa colaboração fez parte de auxílios concedidos pela FAPESP e da publicação de artigos conjuntos ANDRIOLI et al., 2012; RIBEIRO et al., 2010).

3 Objetivo

O objetivo deste projeto de Iniciação Científica é propor uma metodologia de análise de PWMs de um mesmo fator que permita a escolha de uma delas para ser utilizada na predição de novos SLFTs. Também será testada uma forma de criação de uma nova PWM, quando possível, ba que também será comparada com as PWMs já existentes.

4 Metodologia

Para elaboração da proposta, a pesquisa focará no FT denominado *tailless* (*tll*) da espécie *Drosophila melanogaster*. Esse foco servirá como um projeto piloto, que posteriormente deverá ser utilizado para outros FTs com o objetivo de verificar a reprodutibilidade da abordagem delineada.

A hipótese deste trabalho é que uma PWM deveria ser capaz de predizer SLFTs que já foram identificados em experimentos *in vivo*, que representam os dados mais confiáveis da interação entre TFs e SLFTs. Dessa forma, a proposta se baseará em, para cada PWM de *tll*, realizar predições em regiões de CRMs de genes de *Drosophila melanogaster* e comparar essas predições com os picos de ligação de *tll* de experimentos *in vivo* (ChIP-chip) disponíveis no banco BDTNP (*Berkeley Drosophila Transcription Network Project*) (LI et al. , 2008; MACARTHUR et al. , 2009). Então será possível a estimação de medidas de desempenho de cada PWM, como acurácia, sensibilidade, precisão, etc., que por sua vez poderão ser utilizadas para a escolha da PWM mais apropriada.

Para a realização desta proposta, as seguintes tarefas serão necessárias:

1. Levantamento dos banco de dados contendo elementos reguladores de genes da *Drosophila* a fim de identificar os CRMs alvos de *tll* assim como SLFTs específicos;
2. Levantamento e obtenção dos modelos (PWMs) disponíveis de *tll*;
3. Obtenção das sequências de pico (250 base pairs (bp)) de *tll* no site do BDTNP;
4. Levantamento e obtenção das sequências de sítios de *tll*, identificando a sequência propriamente dita, coordenada genômica e se estão presentes nos picos de dados *in vivo*;
5. Obtenção das sequências de *enhancers* dos seguintes genes:bicoid (bcd) / hunchback (hb)/ orthodenticle (otd)/ empty-spiracles (ems)/ buttonhead (btd)/ sloppy-paired 1 (slp1)/ giant (gt)/ Kruppell (Kr)/ knirps (kni) / even-skipped (eve)/ hairy (h)/ runt (run)/ fushi-tarazzu (ftz);
6. Testar todos os modelos (PWMs) nas sequências de pico de dados *in vivo*;
7. Testar todos os modelos (PWMs) nas sequências dos *enhancers*;
8. Criar uma nova PWM derivada com as sequências preditas pelas outras PWMs e que coincidiram com os picos de ligação de *tll in vivo*;
9. Testar essa nova PWM tanto nos *enhancers* e novamente nos picos de ligação in vivo;
10. Análise das PWMs: de posse de todos esses dados e resultados, serão realizadas as seguintes análises:
    a. cada PWM (incluindo a PWM derivada na atividade 8) terá as medidas de desempenho calculadas considerando como resultados positivos aquelas predições que coincidirem com os picos *in vivo*;
    b. será calculada a intersecção dos resultados de cada PWM nos enhancers do genes mencionados na atividade 5. Serão calculadas as medidas de desempenho dessa intersecção com os dados *in vivo;* Essa tarefa visa a identificar um conjunto mínimo de predições que tendam a possuir menor taxa de falsos positivos, candidatas a validação biológica.

5 Cronograma

ATIVIDADES	PREVISÃO
1 A 3	AGOSTO
4 A 5	SETEMBRO
6 A 7	OUTUBRO- NOVEMBRO.2017
8	DEZEMBRO- FEVEREIRO
9	MARÇO- ABRIL
10	MAIO- JULHO
ELABORAÇÃO DO RELATÓRIO	AGOSTO

6 Referências

ANDRIOLI, L.P.; DIGIAMPIETRI, L.A.; de BARROS, L.P.; MACHADO-LIMA, A. Huckebein is part of a combinatorial repression code in the anterior blastoderm. Dev Biol. v. 361; n. 1, p177-185, 2012.

BERGER, M. F.; PHILIPPAKIS, A. A.; QURESHI, A. M.; HE, F. S.; ESTEP, P. W; BULYK, M. L. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature Biotechnology, v. 24, p. 1429-1435, 2006.

GOLD, L. et al. Diversity of oligonucleotide functions. Annual review of biochemistry, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 64, n. 1, p. 763-797, 1995.

LEON, S. B.-T. de; DAVIDSON, E. H. Modeling the dynamics of transcriptional gene regulatory networks for animal development. Developmental biology, Elsevier, v. 325, n. 2, p. 317-328, 2009.

LI, X.-y. et al. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol, Public Library of Science, v. 6, n. 2, p. e27, 2008.

MACARTHUR, S. et al. Developmental roles of 21 Drosophila transcription factors are determined by quantitative diferences in binding to an overlapping set of thousands of genomic regions. Genome Biol, v. 10, n. 7, p. R80, 2009.

RIBEIRO, T.C.; VENTRICE, G.; MACHADO-LIMA, A.; ANDRIOLI, L.P. Investigating giant (Gt) repression in the formation of partially overlapping pair-rule stripes. Dev Dyn., v. 239, n. 11, p. 2989-2999, 2010.

STEWART, A. J.; HANNENHALLI, S.; PLOTKIN, J. B. Why transcription factor binding sites are ten nucleotides long. Genetics, Genetics Soc America, v. 192, n. 3, p. 973–985, 2012.

STORMO, G. D. et al. Use of the `perceptron' algorithm to distinguish translational initiation sites in e. coli. Nucleic Acids Research, Oxford Univ Press, v. 10, n. 9, p. 2997-3011, 1982.