WB2016
Workshop de Bioinformática da UTFPR - 2016

Cornélio Procópio - PR
5 a 7 de outubro de 2016

ISBN: 978-85-7014-180-4
EDITORA UTFPR

Prefácio

O Programa de Pós-Graduação em Bioinformática da UTFPR (Câmpus Cornélio Procópio) promoveu o Workshop em Bioinformática da UTFPR 2016 com a proposta de reunir estudantes, pesquisadores e o setor privado para tratar do tema Bioinformática e suas aplicações. Tendo como base que a Bioinformática envolve diversas disciplinas das áreas de ciências exatas e biológicas, foram abordados temas, desde conceitos introdutórios a aplicações práticas, os quais são fundamentais para compreensão de diferentes problemas biológicos, como também as soluções computacionais envolvidas. As palestras e trabalhos apresentados no workshop abordaram diferentes áreas da bioinformática tais como predição de motifs em sequências, inferência de redes gênicas, integração de dados, predição de genes, montagem de genomas, anotação de transcriptomas, predição de estrutura de proteínas, modelagem de sistemas biológicos, dentre outros. O Workshop em Bioinformática da UTFPR 2016 tem como finalidade contribuir para a introdução dos participantes na área de bioinformática. Em particular, também tem o objetivo de apresentar o programa para a comunidade, ampliar as colaborações entre os pesquisadores da área e fomentar potenciais candidatos ao programa de Mestrado em Bioinformática .

5-7 de Outubro de 2016

Fábio F. da R. Vicente
Comitê de Programa WB2016

Organização

O WB2016 é organizado pelo Programa de Pós-graduação em Bioinformática (PPGBIO) com suporte do DACOM (Departamento de Computação) - UTFPR - Cornélio Procópio.

Comitê de Organização

André Yoshiaki Kashiwabara:	PPGBIOINFO - UTFPR
Douglas Silva Domingues:	Instituto de Biociências - UNESP
Fábio Fernandes da Rocha Vicente:	PPGBIOINFO - UTFPR
Fabricio Martins Lopes:	PPGBIOINFO - UTFPR
Laurival Antonio Vilas-Boas:	CCB - UEL
Márcio Dorn:	INF - UFRGS
José Eduardo de Lima Simão:	Secretário PPGBIOINFO - UTFPR

Comitê Científico

Alexandre Rossi Paschoal:	UTFPR (Cornélio Procópio)
André Yoshiaki Kashiwabara:	UTFPR (Cornélio Procópio)
Arthur Gruber :	ICB-USP
Carlos Henrique Aguena Higa:	UFMS
Dalcimar Casanova:	UTFPR (Pato Branco)
Danilo Sipoli Sanches:	UTFPR (Cornélio Procópio)
Douglas Silva Domingues:	UNESP
Fábio Fernandes da Rocha Vicente:	UTFPR (Cornélio Procópio)
Fabricio Martins Lopes:	UTFPR (Cornélio Procópio)
Flávio Augusto Vicente Seixas:	UEM (Maringá)
Francismar Corrêa Marcelino-Guimaraes:	Embrapa Soja
Laurival Antonio Vilas-Boas:	CCB-UEL
Luiz Filipe Protasio Pereira:	Embrapa/IAPAR
Marcio Dorn:	INF-UFRGS
Mariangela Hungria da Cunha:	Embrapa Soja
Mauro Antônio Alves Castro:	UFPR
Ronaldo Fumio Hashimoto:	(IME-USP)

Apoio

CAPES	CNPq
Embrapa	IAPAR
Belagrícola	

Sumário

RNA - um *app* do Cytoscape para a análise de redes Booleanas

Sibelly G. S. Cavalcante[1], Franklin M. Barbosa[1] e Carlos H. A. Higa[1]
[1]Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Campo Grande - MS
sibellycavalcante@gmail.com

Neste trabalho, estudamos um modelo de redes de regulação gênica, conhecido como rede Booleana, e implementamos um *app* do Cytoscape, que chamamos de RNA - Regulatory Network Analyzer, para a simulação e análise desse modelo. Existem alguns modelos de redes citados na literatura; o modelo Booleano é um modelo discreto e que pode ser analisado mais facilmente, em comparação com modelos quem envolvem equações diferenciais ordinárias, por exemplo. Apesar de mais simples, as redes Booleanas possuem o potencial para explicar diversos fenômenos biológicos, como o ciclo celular da levedura.

O Cytoscape é uma plataforma de código aberto específica para a visualização de redes complexas, e por isso foi utilizada no desenvolvimento do aplicativo de simulação. O Cytoscape possui uma API (*application programming interface*) para que programadores possam desenvolver *apps* utilizando a linguagem Java™, sendo assim independente de plataforma. O *app* RNA pode ser instalado de maneira gratuita através do *Cytoscape App Store* e seu código livremente compartilhado através do GitHub.

O *app* foi desenvolvido para simular redes contendo poucos genes, e é biologicamente plausível uma vez que, apesar de uma grande quantidade de genes estarem presentes em um organismo, para que uma determinada função celular seja realizada apenas uma pequena parte dos genes e seus produtos (proteínas) são necessários. Com o *app* RNA é possível visualizar a rede de regulação gênica, o diagrama de transição de estados e calcular medidas para analisar a estabilidade da rede, como a entropia da rede e o coeficiente de Derrida. Além disso, outras informações a respeito da rede Booleana também são exibidas, com o número de atratores da rede e o tamanho de cada bacia de atração.

Os testes foram feitos em computadores com 8GB de RAM e foi possível executar o *app* com no máximo 20 genes. Isso ocorre pois o diagrama de transição de estados cresce de maneira exponencial de acordo com o tamanho da entrada. Sendo assim, o *app* é adequado quando se deseja analisar sub-redes de genes/proteínas que têm um papel importante em um determinado fenômeno biológico de interesse.

Palavras-chave: Redes Booleanas, cytoscape.

Agradecimentos: Trabalho realizado sem fomento externo/interno.

Modelagem do ciclo celular da levedura através de redes Booleanas com limiar

Mariana C. de Souza[1] e Carlos H. A. Higa[1]

[1]Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Campo Grande - MS

mariana.caravanti@gmail.com

As redes de regulação gênica (GRN) representam as interações entre genes/proteínas de um organismo no contexto de um fenômeno biológico em estudo. Um modelo matemático popular de GRN são as redes Booleanas (BN). Neste trabalho, modelamos o ciclo celular da levedura usando uma BN e tomando como base um modelo já proposto. Uma rede Booleana é formada por um conjunto de genes e um conjunto de funções Booleanas, onde cada gene pode ser representado por dois valores: 0 ou 1. As funções Booleanas governam os estados dos genes.

O modelo já proposto possui 11 genes, responsáveis pelo controle do processo biológico. Ao simular esta rede Booleana, podemos construir o seu diagrama de transição de estados. Neste modelo determinístico, certos estados (uma configuração binária de todos os genes) são revisitados de maneira cíclica, dependendo de um estado inicial. Tais estados são chamados de atratores. No modelo da levedura, foi identificado um atrator que corresponde ao estado biológico denominado G_1 estacionário, onde a célula permanece até atingir um certo tamanho para que a divisão celular ocorra. O problema desta modelagem é que, uma vez que a rede se encontra no estado G_1 estacionário, ela permanece neste atrator *ad eternum*.

Tomando como base este modelo já proposto, foram realizadas modificações de modo que, quando a levedura realiza a divisão celular, ao invés de permanecer no estado G_1 estacionário, ela realiza uma nova rodada de divisão celular. Primeiramente, foram geradas todas as possíveis BNs consistentes com este ciclo celular modificado. Em uma segunda parte, filtramos as redes mais parecidas com a rede original, levando em consideração as interações entre os genes. Por último, foram gerados os diagramas de transição de estados, sendo possível filtrar os diagramas que possuíam uma bacia principal, em que a probabilidade da levedura permanecer ali fosse maior do que a probabilidade dela permanecer em outras bacias. Esta informação foi obtida através do cálculo da Entropia darde e do Coeficiente de Derrida. Com nossos experimentos, observamos que, apesar de existir uma grande quantidade de redes Booleanas com limiar contendo 11 genes (3^{11^2}), poucas delas são capazes de representar o ciclo celular modificado. Esse número se torna ainda menor quando selecionamos as redes mais estáveis através da Entropia e do Coeficiente de Derrida. Como sugestão de trabalho futuro, propomos um estudo detalhado das redes obtidas no que diz respeito às interações entre os genes, para investigar se as interações obtidas fazem sentido biologicamente.

Palavras-chave: Redes Booleanas, ciclo celular, modelagem, levedura.

Agradecimentos: CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico.

Análise e Classificação de Sequências Genômicas Utilizando Redes Complexas

Isaque Katahira[1,2], Fabrício Martins Lopes[1]

[1]Universidade Tecnológica Federal do Paraná – Campus Cornélio Procópio
Departamento de Computação
Programa de Pós – Graduação em Bioinformática – PPGBIOINFO
fabricio@utfpr.edu.br

[2]Centro Estadual de Educação Tecnológica Paula Souza– Escola Técnica Estadual Prof. Mário Antônio Verza
isaque.katahira@etec.sp.gov.br

A integração da computação com outras áreas do conhecimento vem se tornando cada vez mais importante para a evolução dos estudos em diversas áreas da ciência. Notavelmente, a Bioinformática é exemplar desse fato, uma vez que é intrinsecamente multidisciplinar ao aliar conhecimentos de biologia, estatística e informática dentre outras ciências aplicáveis. Nesse contexto, a extração simultânea de dados moleculares de milhares de genes é um grande avanço na área, porém representa também um desafio, uma vez que gera um intenso volume de dados biológicos a serem analisados. Em particular, torna-se essencial o desenvolvimento de novos métodos e técnicas a fim de compreender a influência da expressão dos genes no estado funcional dos organismos. Nesse contexto, um caminho que pode contribuir é a redução do volume de dados sem que haja prejuízo da informação contida nos mesmos. Nesse sentido, podem ser aplicadas as técnicas de extração de características, as quais buscam representar os dados por meio de suas propriedades mais relevantes sem fazer uso de todo o conjunto de dados. Dessa forma, é proposto o desenvolvimento de uma abordagem baseada no mapeamento de sequências genômicas em suas respectivas redes complexas. A partir dessas redes podem ser extraídas medidas para especificá-las em vetores de características, os quais podem ser usados para sumarizar os dados coletados, de modo a quantificar as semelhanças topológicas entre as redes geradas. Justifica-se, portanto, a utilização de redes complexas por sua aproximação com as redes reais no que tange a não-linearidade, em que a dinamicidade dos elementos envolvidos é determinante para a compreensão das interações, funções e estruturas. A observação das características extraídas, por meio de nós, arestas e outras medidas, podem proporcionar a identificação de padrões contidos nas redes. Logo, espera-se que as medidas extraídas permitam distinguir diferentes classes de sequências genômicas, verificar mutações, construir árvores filogenéticas, entre outras análises. Assim, existe um potencial de observar diferentes padrões contidos nas estruturas biológicas e propor aplicações de classificação em larga escala no contexto da biologia sistêmica, isto é, considerando a modelagem de um organismo como um todo.

Palavras-chave: Bioinformática, inferência de redes, reconhecimento de padrões.

Inferência de rede de regulação da expressão de genes relacionados ao biofilme de *Candida albicans* influenciados pelo ácido lático.

Alexandre Tadachi Morey[1,2], Eliandro Reis Tavares[1], Sérgio Paulo Dejato da Rocha[1], Sueli Fumie Yamada-Ogatta[1], Fabrício Martins Lopes[3].

[1]Departamento de Microbiologia, CCB, Universidade Estadual de Londrina, Londrina. [2]Bolsista PNPD/CAPES, Programa de Pós-Graduação em Microbiologia, Universidade Estadual de Londrina, Londrina. [3]Campus Cornélio Procópio, Universidade Tecnológica Federal do Paraná, Cornélio Procópio.

RESUMO

Candida albicans é uma levedura comensal de diferentes sítios anatômicos do homem, porém, em casos de imunodebilidades do hospedeiro torna-se patogênica, acometendo principalmente mucosas orofaríngeas e do trato urogenital. Um dos principais problemas enfrentados atualmente em relação às infecções causadas por este fungo é a formação do biofilme, uma comunidade microbiana composta por células de uma ou mais espécies protegida por uma matriz extracelular e que apresenta fenótipo diferente em relação às células planctônicas. No caso de mulheres, além de *C. albicans*, a bactéria *Streptococcus agalactiae*, produtora de ácido lático, pode causar infecções na mucosa vulvovaginal e estudos iniciais indicam que estes dois micro-organismos podem formar biofilmes mistos. Assim, o presente estudo utilizou dados de transcriptoma presentes no banco GEO (Gene Expression Omnibus, NCBI) de *C. albicans* (SC5314) cultivada na presença de ácido lático e inferiu, utilizando o software multiplataforma (Java) de código aberto com seleção de características genéticas para reconhecimento de padrões DimReduction, redes de regulação da expressão de genes relacionados à formação de biofilme nesta levedura a partir 3 genes alvos: *GPD2* (orf19.691), *ACH1* (orf19.3171) e *GCN4* (orf19.1358), exclusivamente expressos na presença do ácido lático. Após a análise das 3 redes de regulação foi possível afirmar que a inferência estatística que apresentou maior correlação com a inferência biológica da formação do biofilme em *C. albicans* foi a que continha os genes *GCN4* (orf19.1358), *SEC6* (orf19.5463), *GPX2* (orf19.85), orf19.7490, *CTP1* (orf19.5870) e orf19.6668. Estes genes estão relacionados, principalmente, a adesão, formação de hifas, produção da matriz extracelular, via metabólica de carboidratos e ciclo celular, necessários para a formação e maturação do biofilme. Apesar do número significativo de genes associados à formação de biofilme por *Candida albicans* publicado na literatura, os mecanismos de regulação deste processo, a correlação com outros genes e a sinalização entre espécies do biofilme misto ainda não foram totalmente elucidados, assim, estes dados abrem perspectivas para o entendimento molecular da influência do ácido lático, um metabólito produzido por algumas bactérias da microbiota, na formação do biofilme de *C. albicans* e no biofillme misto.

Palavras-chave: Biofilme misto, transcriptoma, redes de regulação.

Agradecimentos: PNPD/CAPES

Aplicação de Abordagens Heurísticas na Descoberta de Motifs Biológicos

Jader M. Caldonazzo Garbelini[1], André Yoshiaki Kashiwabara[1] e Danilo Sipoli Sanches[1]

[1]PPGBINFO, Universidade Tecnológica Federal do Paraná, Cornélio Procópio

jaderg@alunos.utfpr.edu.br

A identificação dos sítios de ligação dos fatores de transcrição em sequências de DNA é o primeiro passo para a compreensão da regulação gênica. Reconhecer estes padrões nas regiões promotoras de genes co-expressos é um fator determinante para isso. Apesar de existirem vários algoritmos com este propósito, o problema ainda está longe de ser resolvido devido a grande diversidade da expressão gênica e a baixa especificidade dos locais de ligação. Algoritmos do estado-da-arte possuem limitações, tais como elevado número de falsos positivos e baixa precisão para identificação de motifs fracos. Neste artigo nós apresentamos o Discovery Mofifs by Memetic Algorithm, um algoritmo memético desenvolvido utilizando computação evolutiva juntamente com as heurísticas simulated annealing e variable neighborhood search. Para atestar sua capacidade, foram realizados testes em quatro datasets - dois reais e dois sintéticos - e os resultados foram comparados com outras abordagens encontradas bem conhecidas da literatura.

Palavras-chave: Motifs, heurísticas, fatores de transcrição.

Agradecimentos: Os autores gostariam de agradecer a PPGBIOINFO, CAPES e Universidade Tecnológica Federal do Paraná pelo apoio financeiro concedido a esta pesquisa.

DESENVOLVIMENTO DE UMA FERRAMENTA PARA IDENTIFICAR REGIÕES CODIFICADORAS NOS TRANSCRITOS DO FUNGO *Phakopsora pachyrhizi*

Cynara Leão GARCIA[1], Francismar Corrêa MARCELINO-GUIMARAES[2], André Yoshiaki KASHIWABARA[1]

[1] Programa de Pós-graduação em Bioinformática (PPGBIOINFO), Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Brasil
cynara@alunos.utfpr.edu.br
[2] Empresa Brasileira de Pesquisa Agropecuária - Embrapa, Londrina, Brasil

A soja representa uma das principais culturas agrícolas no Brasil em termos de economia, desenvolvimento e empregabilidade, sendo líder nacional em produção e área cultivada. No entanto, inúmeros fatores podem colocar em risco esse cenário, fatores esses que estão relacionados às condições climáticas e doenças. Dentre as doenças conhecidas, a mais preocupante é a ferrugem asiática da soja (FAS), pelo fungo *Phakopsora pachyrhizi*, que pode provocar perdas na produção da soja de 30% a 75%. Apesar de sua importância, o genoma completo desse fungo ainda não está disponível; assim como outras espécies de ferrugem, espera-se que o genoma *P. pachyrhizi* seja altamente complexo, com conteúdo altamente repetitivo, o que dificulta então sua montagem, além de ser estimado em cerca de 500 a 800 Mb, representando um extremo quando comparado aos genomas de outros fungos já sequenciados. Por outro lado, vários trabalhos visando identificar sequencias expressas do fungo já foram descritos, como o transcriptoma composto por 36.350 *contigs*, resultantes de montagem *ab-initio* de sequência única de *P. pachyrhizi*, obtidos de microdissecção a laser de lesão aos 10 dias de infecção. Apesar do número significativo de contigs descritos, muitos ainda não apresentam anotação e quantidade de *non hits* ainda é elevada. Dessa forma, ferramentas de bioinformática que possam auxiliar na predição de sequências codificadoras expressas se tornam extremamente úteis. Neste trabalho, dois programas mais citados pelo Google Scholar: ESTscan e OrfPredictor foram comparados, utilizando como entrada a base de dados de sequências contendo UTRs e CDSs de genes preditos que compõe o genoma da planta *Arabidopsis thaliana* obtida do banco de dados TAIR. Ambos os programas ESTScan e o OrfPredictor apresentaram uma taxa elevada de falsos positivos (41,90% e 87,85% respectivamente) de modo que faz-se necessária a implementação de um novo método para a análise de sequências transcritas. Assim, este projeto visa desenvolver um modelo de análise de transcritos utilizando o ToPS, com base na representação de regiões codificadoras utilizando Cadeias Generalizadas de Markov (GHMM). Esta abordagem amplamente utilizada em preditores de genes ainda foi pouco explorada por preditores de regiões codificadoras em sequências expressas. Adicionalmente, será também utilizado o algoritmo Viterbi para segmentar os transcritos em regiões codificadoras e regiões não traduzidas, utilizando-se de uma heurística capaz de efetuar a identificação dos erros de sequenciamento. Finalmente, o modelo deve ser capaz de ser aplicado tanto à problemática do *P. pachyrhizi* e generalizável para outras espécies.

Palavras-chave: Transcritos, ESTs, *Phakopsora pachyrhizi,* GHMM, bioinformática.

Agradecimentos: EMBRAPA

LncRNAplant-Finder :Uma ferramenta para predição de lncRNA em plantas.

Tatianne da Costa Negri[1], Pedro Henrique Bugatti[1], Priscila Tiemi Maeda Saito[1], Douglas Silva Domingues[1,2] e Alexandre Rossi Paschoal[1*]

[1]Programa de Pós-Graduação em Bioinformática - PPGBIOINFO, Universidade Tecnológica Federal do Paraná, Cornélio Procópio

[2]Departamento de Botânica, Instituto de Biociências, Universidade Estadual Paulista, campus de Rio Claro

*paschoal@utfpr.edu.br

Longos RNAs não-codificantes (lncRNA) representam uma classe de reguladores de expressão genica com mais de 200 nucleotideos, encontrados em vários eucariotos; No entanto, o conhecimento de lncRNAs em plantas ainda é muito limitado. Apesar da descrição de vários lncRNAs em plantas com importantes papéis regulatórios nos últimos anos, a sua caracterização é diferente dos demais eucariotos. Além disso, a inexistência de abordagens de predição exclusiva de plantas fez com que fosse criado o LncRNAplant-Finder, uma abordagem para a identificação de lncRNAs em plantas. Para desenvolvimento do programa foram utilizadas informação de lncRNAs e transcritos de seis espécies de vegetais: *Arabidopsis thaliana*, *Cucumis sativus*, *Glycine max*, *Oryza sativa*, *Populus trichocarpa* e *Setaria italica*. Esses dados foram obtidos dos bancos de dados públicos (PLNlncRbase, GREENC e Phytozome) dos quais usou-se 22.543 lncRNAs e 29.960 transcritos (mRNAs). Técnicas de reconhecimento de padrões foram aplicadas em 85 características de composição e estrutura das sequências de lncRNAs e transcritos (e.g. conteúdo GC, ORF, distribuição dinucleotidica, trinucleotidica). A partir destas, testes foram realizados para seleção das melhores características. Nesta tarefa foram usados os seguintes programas: i- CD- Hit- Est, para retirada de sequencias similares; ii- txCDSPredict, para levantamento das regiões codificantes; iii- script em linguagem PERL para contagem de dinucleotídeos, trinucleotídeos, conteúdo GC, normalização dos dados e criação do arquivo ".arff"; iv- seleção de características, estudo e escolha do melhor classificador via Weka 3.8.0. Após a etapa de seleção de características, 16 delas foram escolhidas para a construção do classificador. Seis métodos de classificação foram testados, dos quais o método J48 apresentou o melhor resultado com os seguintes valores: i- correta classificação, ≃ 97%; ii- erro de classificação, ≃ 3%; iii- acerto de lncRNA, 22.021 (≃97,7%); iv- acertos de transcritos, 28.812 (≃96,25) ; v- erros de lncRNAs, 522 (≃2,3%); vi- erros de transcritos, 1.148 (≃3,9%). A partir da pesquisa realizada e os devidos resultados obtidos até o momento, pode-se afirmar que esta abordagem adotada contribuirá para a identificação de lncRNAs em genomas de vegetais.

Palavras-chave: lncRNA, mRNA, identificação, non-coding.

Agradecimentos: Edital Universal - CNPq Nº. 14/2014 - Projeto: 454505/2014-0, Bolsa - Fundação Araucária Edital Nº 019/2015.

Título: Aplicação do modelo oculto de Markov para identificar e classificar genes *cry* de Bacillus thuringiensis.

GONÇALVES, K. C. B[1]., VILAS-BOAS, G.T[1]. WOLF, I. R, VILAS-BOAS, L. A[1].
[1]Departamento de Biologia Geral, Centro de Ciências Biológicas, Universidade Estadual de Londrina.
kbrumatti@gmail.com.

Resumo:

O *Bacillus thuringiensis* (Bt) é uma bactéria formadora de esporos que produz a toxina Cry como cristais paraesporais, amplamente usada no controle de pragas agrícolas, bem como vetores de doenças. Triagem de cepas de Bt e sequenciamento do genes *cry* levou à identificação de mais de 700 sequências que foram classificadas de acordo com a identidade de sequência de aminoácidos em pelo menos 75 grupos diferentes (Cry1, Cry2,...Cry75). A cada ano novas proteínas são descobertas e adicionadas a estes grupos. Um dos grandes problemas relacionados aos genes *cry* é a sua correta localização e classificação. O presente trabalho teve como objetivo utilizar métodos de bioinformática para a criação de uma base de dados de genes *cry*, e utilização desta para identificação e classificação em família, utilizando como entrada a sequência de nucleotídeos isolada ou um conjunto de contigs. A estratégia proposta foi delineada a partir de um set de sequências obtidas por meio do banco *B. thuringiensis* Toxin Nomenclature. Ferramentas computacionais foram empregadas na construção do banco de dados curado. O alinhamento das sequências foi gerado com o algoritmo ClustaW, seguindo da anotação com o software Ugene, identificando as ORFs que codificam para a proteína Cry. Estas sequências foram conferidas com o Blast determinando as regiões de similaridade entre sequências. Estes passos foram realizados para a criação de um Banco de dados curado de genes *cry*. O passo seguinte se deu por uma estruturação usando comandos em Linux e scripts personalizados elaborados na linguagem Bash, juntamente com o programa Hmmer para criação de modelos e identificação dos genes *cry* e suas classificações. Nossos resultados mostraram que a estratégia computacional de bioinformática foi capaz de promover a descoberta de candidatos a genes *cry*, a partir da sequência de nucleotídeos de cepas distintas, apresentando relatório de saída caracterizado contendo dados importantes como, nome, score, e-value, posição (início e fim) do gene, sentido da fita.

Mais estudos serão realizados a partir deste trabalho, incluindo a criação de uma ferramenta web e uma base de dados online, na identificação de proteínas Cry, contribuindo desta forma para o trabalho da comunidade científica.

Palavras chave: Bioinformática, genes *cry*, *Bacillus thuringiensis*.

Identificando Eventos de Splicing Alternativos em Dados de RNAseq Utilizando Grafos de De Bruijn e Bloom Filters

Ricardo Medeiros da Costa Junior[1] e André Yoshiaki Kashiwabara[1]

[1]Universidade Tecnológica Federal do Paraná - Campus Cornélio Procópio, Departamento de Computação, Programa de Pós-Graduação em Bioinformática - PPGBIOINFO

Alternative Splicing (AS) é um mecanismo pós-transcricional em que múltiplos transcritos funcionais podem ser produzidos a partir um único gene. Em particular, um gene que codifica proteína pode produzir diferentes proteínas através de eventos de AS do pre-mRNA. Nesse processo, alguns exons podem ser incluídos ou excluídos do final do RNA mensageiro (mRNA). Por consequência disso, proteínas traduzidas de AS mRNA contém diferenças em suas sequências de amino ácido e, frequentemente, em suas funções biológicas. Nota-se, que o processo AS permite o genome humano sintetizar diretamente muitas proteínas que poderiam ser esperadas proveniente do seus 20.000 genes codificantes de proteínas. Estudos recentes relacionam abnormally spliced mRNAs com células cancerígenas.

Em 2012, é proposto um algoritmo para identificação e quantificação de polimorfismos para dados provenientes de RNA-seq quando o genoma de referência não está disponível, sem realizar a montagem de todos os transcritos. Apesar desse algoritmo identificar tanto approximate tandem repeats, SNPs (single nucleotide polymorphism) e AS, ele é focado em quantificar apenas AS. Por meio desse método, foi possível identificar que anotação de eventos AS tem sido subestimada, pois 56% dos AS identificados no conjunto de dados testados não estavam presente nas anotações atuais. No entanto, o algoritmo tem algumas limitações. Assim como a maioria de montadores de novo baseados em DBG,(De Bruijn Graphs) a construção do grafo requer um custo de memória muito alto e deve ser executado em um cluster.

Em 2016 foi publicado um artigo cujo propõe uma melhoria para um montador baseado em DBG. Foi retirado o MPI (message-passing system) e implementado o Bloom Filter, uma estrutura de dados probabilística na construção do DBG. Foi possível realizar a montagem em um computador pessoal ao invés de um cluster. Bloom filter é uma estrutura de dados probabilística criada por Burton Howard Bloom em 1970, que é usada para testar se um elemento é membro de um conjunto. Combinações falso positiva são possíveis, mas falso negativas não são, devido a isso Bloom filter é considerado com 100% de taxa de recall. Ou seja, é retornado 100% dos resultados relevantes.

Como a construção do DBG desse montador é muito semelhante ao do algoritmo de identificador e quantificador de AS, a proposta desse trabalho é implementação do Bloom Filter no algoritmo de identificação e quantificação de AS, reduzindo o custo de memória para criação do DBG, permitindo que esse seja executado de maneira eficiente em um computador pessoal.

Palavras-chave: Splice Alternativo, Bloom Filter, De Bruijn Graph

Agradecimentos: CNPq 476689/2013-9

RNAs NÃO CODIFICANTES EM *COFFEA CANEPHORA*: IDENTIFICAÇÃO VIA SIMILARIDADE.

S. M. C. De Lemos[1], A. R. Paschoal[1], D. S. Domingues[1,2]

[1] Programa de Pós Graduação em Bioinformática – PPGBIOINFO - Universidade Tecnológica Federal do Paraná – Câmpus Cornélio Procópio

[2] Departamento de Botânica, Instituto de Biociências, Universidade Estadual Paulista – Câmpus Rio Claro

RNAs não-codificantes (ncRNAs) constituem um importante componente de genomas e transcriptomas de eucariotos. Quando transcritos, eles são usualmente gerados por regiões de íntrons e intergênicas do genoma. Apesar de possuírem papéis relevantes na regulação da expressão gênica, a enorme maioria dos estudos em ncRNAs são voltados para análises no genoma humano ou de organismos-modelo. Processos fundamentais para a fisiologia e o desenvolvimento vegetal, como o florescimento e a maturação de frutos são regulados por ncRNAs em plantas, o que motiva a caracterização deste componente genômico em espécies de interesse agronômico.

O café é uma das principais *commodities* agrícolas mundiais, da qual o Brasil é o principal produtor e segundo maior mercado consumidor. Nos últimos anos ocorreram enormes ganhos na geração de bancos de dados públicos de sequências de genoma e transcriptoma no cafeeiro, sobretudo para a espécie *Coffea canephora* (conhecido no Brasil como Robusta). Com isso, surgiu a demanda da caracterização funcional desses dados gerados. Neste contexto, o objetivo deste trabalho foi identificar ncRNAs do genoma recém sequenciado de *C. canephora*. Regiões intergênicas foram extraídas utilizando *pipeline* desenvolvido com scripts *Shell* e PERL, com exclusão de regiões de baixa qualidade de sequenciamento. Por meio de BLAST, foi feito o alinhamento destas regiões intergênicas contra ncRNAs de 31 espécies vegetais disponíveis no repositório ENSEMBL – plantas - (versão 32). Foi possível identificar 288 alinhamentos com pelo menos 80% de identidade. Desses, 23 tiveram 100% de cobertura e identidade sendo classificadas como: 5 microRNAs de *Oryza sativa* (arroz), 8 RNAs transportadores (tRNAs)

de *Solanum lycopersicum* (tomate), 7 tRNAs de *Solanum tuberosum* (batata), 1 tRNA de *Amborella trichopoda* e 2 tRNAs de *Triticum urartu* (trigo). Um total de 2 miRNAs em *Oryza sativa* e 2 tRNAs em *Amborella trichopoda* contiveram sequências idênticas, mas não obtiveram total cobertura. *Coffea canephora* também apresentou um índice de 100% de cobertura e identidade acima de 80% com 2 pequenos RNAs nucleolares (snoRNAs) e 1 miRNA de *Oryza glaberrima* (espécie africana de arroz), 5 miRNAs de *Oryza sativa*, 3 snoRNAs de *Solanum tuberosum*, 4 tRNA de *Prunus persica* (pêssego), 2 snoRNA de *Solanum lycopersicum*, 1 tRNA de *Medicago truncatula* (luz cortada) e 2 snRNA de *Hordeum vulgare* (cevada). Esses resultados preliminares fornecem um ponto de partida para outras pesquisas na mesma área, ajudando a compreender como funciona a base molecular de regulação de importantes processos de interesse agronômico em cafeeiro, (como a frutificação), e em plantas em geral.

Palavras-chave: NCRNAs, café, similaridade.

Agradecimentos: PPGBIOINFO.

Identificação *in silico* de pequenos RNAs não codificantes em genoma de *Proteus mirabilis* uropatogênico

Sérgio Paulo Dejato da Rocha[1]; Ivan Rodrigo Wolf [2]; Laurival Antônio Vilas-Boas[2]; Alexandre Rossi Paschoal[3]

1: Departamento de Microbiologia, Centro de Ciências Biológicas, Universidade Estadual de Londrina
2: Departamento de Biologia Geral, Centro de Ciências Biológicas, Universidade Estadual de Londrina
3: Universidade Tecnológica Federal do Paraná-Campus Cornélio Procópio
Contato: rochaspd@uel.br

A infecção do trato urinário (ITU) caracteriza-se pela multiplicação da bactéria em qualquer parte deste local, seja nos rins, ureteres, bexiga ou uretra. A infecção do trato urinário associada a cateter (ITU-CA) é a infecção nosocomial mais comum dentre todas, e representa mais de 80% das ITU nosocomiais. *Proteus mirabilis,* uma bactéria Gram-negativa, não é um comum causador de ITU em pacientes normais, sendo mais relacionado à ITU complicada, principalmente à ITU-CA. *P. mirabilis* possui dezenas de genomas sequenciados cujo tamanho varia de 3 a 5 Mb. Pequenos RNAs (sRNAs) são sequências curtas de RNA que regulam a expressão gênica. Estas moléculas atuam através do seu pareamento com sequências alvo por meio de complementaridade específica ou parcial de bases. Em procariotos os sRNAs desempenham papéis fundamentais em redes reguladoras de expressão nas respostas à estímulos ambientais, inclusive em bactérias patogênicas. A abordagem computacional é considerada uma das mais eficientes para localização de candidatos a sRNAs em sequências de genomas. Podendo ser dividida de acordo com o método de busca utilizado; dentre os quais, podem ser citados os que procuram estruturas secundárias consensuais; os que efetuam na busca por sinais de transcrição; e os que aplicam genômica comparativa e *ab initio.* Diante disso, este trabalho realizou a identificação e caracterização de pequenos RNAs não-codificantes no genoma de *P. mirabilis* por meio de análises de bioinformática (*in silico*). Em suma, os candidatos a sRNA do genoma da cepa *P. mirabilis* HI4320 foram preditos utilizando-se os programas INFERNAL/Rfam e nocoRNAc. Após uma análise manual dos resultados de anotação, foram considerados um total de 29 sRNA. Em geral, os tamanhos variaram de 41 a 505 nucleotídeos, com média de 273 nucleotídeos, sendo que oito tem anotação funcional relacionados com descrito na literatura. Por exemplo, seis já foram descritos em *Escherichia coli* e duas em *Lactococcus lactis*. Em relação as famílias, quatro são cis-regulatórios e quatro são sRNA, sendo que todos se localizam no cromossomo bacteriano. Quanto à função, todos estão envolvidos na regulação da expressão de proteínas de vias metabólicas de procariotos. Portanto, estes resultados podem contribuir no conhecimento da regulação da expressão gênica deste uropatógeno. Para que estas moléculas deixem de ser somente preditas uma futura validação experimental *in vitro* torna-se necessária bem como a busca pelos seus salvos e o entendimento de suas funções regulatórias.

Palavras-chave: pequenos RNAs, *Proteus mirabilis,* non-coding RNA.

MONTAGEM *DE NOVO* E ANOTAÇÃO FUNCIONAL DO TRANSCRIPTOMA DE *ANTICARSIA GEMMATALIS*

Larissa Forim Pezenti[1], Kátia Brumatti Gonçalves[1], Rogério Fernandes de Souza[1], Laurival Antônio Vilas-Boas[1], Carlos Roberto Maximiano da Silva[1], Daniel Ricardo Sosa-Gomez[2], Renata da Rosa[1].

[1] Universidade Estadual de Londrina, Departamento de Biologia Geral, Laboratório de Bioinformática, Londrina-PR

[2]Embrapa Soja Londrina, Londrina-PR

laripez@hotmail.com

Anticarsia gemmatalis Hübner 1818 (Lepidoptera: Noctuidae), conhecida como lagarta-da-soja é o desfolhador mais comum da soja no Brasil, ocasionando perdas consideráveis na produtividade da cultura. Informações relevantes têm sido obtidas a respeito das bases genéticas e fisiológicas associadas à resistência a vários tipos de inseticidas, através da aplicação de técnicas moleculares e análises de bioinformática, para determinar padrões de expressão gênica. Entre elas, destaca-se o RNA-seq que permite tanto quantificar o nível de expressão gênica, quanto analisar a estrutura do transcriptoma sem a necessidade de um conhecimento prévio do genoma estudado. Dessa maneira, neste trabalho reportamos a montagem e a anotação funcional de transcritos de diferentes populações de *A. gemmatalis* resistentes e suscetíveis, tratadas com a proteína bioinseticida de *Bacillus thuringiensis,* Cry1Ac. Foram construídas seis bibliotecas de cDNAs (Illumina TruSeq Stranded mRNA LS) a partir de seis amostras de RNAs extraídas conforme protocolo adaptado utilizando TRIzol® Reagent e o SV Total RNA Isolation System Promega. O sequenciamento foi realizado na plataforma Illumina® HiSeq 2500 com leituras *paired-end* de 125pb. A qualidade dos *reads* obtidos foi verificada com FastQC v0.11.4 e trimados com Prinseq v0.20.4. A montagem do transcriptoma foi realizada na plataforma Trinity v2.2.0 com a estratégia *de novo*, pois não existe um genoma de referência para o mapeamento dos transcritos e construção dos *contigs*, sendo que, em seguida, foi utilizado o CD-HIT-EST para redução da redundância. Neste primeiro momento somente os transcritos com mais de 1.000 pb foram anotados e categorizados usando o software Blast2go v3.2.7, após a busca por similaridade dos transcritos contra banco de dados do NCBI. O transcriptoma das seis bibliotecas de *A. gemmatalis* apresentou 332.700 transcritos, montados a partir de 241.776.712 *reads*, com uma média de 40.296.119 milhões de *reads* por biblioteca, e N50 variando de 1146 a 1520. Os dados apresentados são preliminares, porém este estudo nos permite reportar a montagem do transcriptoma de *A. gemmatalis* e apresentar dados de anotação funcional. Estes resultados abrem novas perspectivas para o estudo genômico de *A. gemmatalis*, e possibilitará a posterior identificação de genes candidatos relacionados aos mecanismos de resistência de importantes pragas agrícolas.

Palavras chave: lagarta-da-soja; resistência a inseticidas; RNA-seq.

Apoio Financeiro: CAPES/CNPq; EMBRAPA Soja, CNPSO

Distribuição e impacto de elementos transponíveis em genes de *C. reinhardtii* e *V. carteri*

Gisele S. Philippsen[1,2], Juliana S. Avaca-Crusca[1], Ana P. U. Araujo[1] e Ricardo DeMarco[1]
[1]Departamento de Física e Ciência Interdisciplinar, Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos – SP – Brasil.
gistrieder@ufpr.br
[2]Universidade Federal do Paraná, Jandaia do Sul – PR – Brasil.

Elementos transponíveis (TEs) são sequências de DNA que possuem a capacidade de transposição no genoma hospedeiro, característica esta que possibilita aos mesmos influenciar a trajetória evolutiva das espécies. Neste contexto, o principal objetivo deste trabalho reside na investigação acerca da distribuição e impacto de TEs em diferentes regiões gênicas das algas Chlamydomonas reinhardtii e Volvox carteri. Análises para apurar a distribuição dos elementos em regiões gênicas foram estabelecidas, nas quais a frequência observada dos elementos foi comparada à frequência esperada segundo a distribuição randômica de TEs no genoma, simulada computacionalmente. Foram constatadas regiões em que a presença dos elementos encontra-se abaixo do esperado, a exemplo de intervalos adjacentes ao início e ao término dos genes, o que provavelmente reflete a seleção negativa de eventos de integração em virtude dos efeitos deletérios associados à disrupção de estruturas de regulação da expressão gênica. O estudo da distribuição de TEs nos íntrons indicou a preservação destas regiões quando à fixação de TEs, sendo a representatividade abaixo do esperado mais evidente em intervalos adjacentes ao éxon, o que minimiza a chance de disrupção do padrão de splicing dos genes. Em sequências codificantes, a escassez de TEs – esperada devido ao provável efeito deletério destes eventos à função do gene – foi constatada para as duas espécies. No entanto, inovações decorrentes da integração de TEs em regiões codificantes podem resultar em efeitos evolutivos positivos, embora estes eventos sejam raros. Considerando a espécie C. reinhardtii, foram identificados onze genes com regiões codificantes derivadas de TEs, sendo o gene Cre06.g262800 de especial interesse por possuir um domínio PHD-finger derivado de uma cópia do elemento Gypsy-5_CR. Apesar da baixa representatividade de TEs no genoma de C. reinhardtii quando comparada à representatividade de TEs no genoma humano, foi observado que a proporção de sítios de poliadenilação derivados de TEs nesta alga é similar à descrita na literatura para a espécie humana. Considerados conjuntamente, os resultados deste estudo sugerem que a modesta representatividade de TEs no genoma das algas C. reinhardtii e V. carteri está relacionada com um rigoroso processo de seleção negativa, associado à retenção de cópias que contribuem positivamente com as estruturas gênicas.

Palavras-chave: Elementos transponíveis, domesticação de elementos transponíveis, evolução gênica.

Agradecimentos: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Pró-reitoria de pesquisa da Universidade de São Paulo (PRP-USP).

Caravela: um navegador para metagenomas

Gianluca Major Machado da Silva[1] e João Carlos Setubal[1,2]
[1]Curso de pós-graduação Interunidades em Bioinformática, IME/USP, São Paulo
gianlucamajor@gmail.com
[2]Departamento de Bioquímica, IQ/USP, São Paulo

Análises de diversidade taxonômicas (baseadas em reads) e análise funcional (baseadas em contigs/genes) a partir de estudos metagenômicos comumente geram informações complementares. No entanto, as ferramentas que geram anotação funcional dos genes (ex. *IMG/M*) e classificadores taxonômicos (ex. *MyTaxa*) não permitem a fácil integração dos resultados. Motivados por essa separação nós estamos desenvolvendo uma plataforma web que facilita a integração, busca e visualização das informações geradas pelas análises baseadas em reads e análises baseadas em contigs/genes. Atualmente, a versão beta da ferramenta já foi disponibilizada para alguns integrantes do grupo. A Caravela está apta a exibir uma lista de contigs e para cada contig, é possível visualizar os genes anotados, reads que participam da sua composição e táxon associado a cada read. Na mesma área de visualização também é possível identificar regiões de sobreposição entre reads associados a diferentes táxon. Tais funcionalidades permite a Caravela encontrar contigs mal montados assim como reads mal classificados, de forma manual através da navegação e visualização dos contigs, ou ainda, de forma automatizada através dos relatórios gerados a partir da ferramenta. A plataforma está habilitada a aceitar arquivos de saída de várias ferramentas, desde que sigam certos padrões. O conjunto de dados que estamos utilizando para testar e validar a ferramenta são reads e contigs/genes obtidos a partir da compostagem do Parque Zoológico de São Paulo.

O desenvolvimento da CARAVELA está sendo feito sobre a plataforma *Java, HTML5, CSS e JavaScript*. Para armazenamento dos dados usamos o sistema gerenciador de banco de dados *Mysql Server*. Por se tratar do desenvolvimento de uma plataforma web, fazemos uso do servidor de aplicação web *Apache Tomcat* para abrigar a ferramenta desenvolvida. A implantação foi feita em máquina virtual (VM) do parque computacional do laboratório do Prof. João Carlos Setubal. Atualmente a VM está configurada com 4 núcleos de processamento, 8 GB de memória, 150 GB de disco e sistema operacional *Ubuntu Server*.

Palavras-chave: Metagenômica, comunidade microbiana, ferramenta, bioinformática, montagem, classificação taxonômica, quimera

Agradecimentos: FAPESP, CAPES, CNPQ.

DETERMINAÇÃO DE MARCADORES DE PCR PARA DIAGNÓSTICO MOLECULAR DE *Aeromonas hydrophila*

APPEL, R.J.C.[1]; FONTE, N.[1]; ALTRÃO, C.S.[1]; PRETTO-GIORDANO, L.G.[1]; VILAS-BOAS, L.A.[1]

[1]Departamento de Biologia Geral – CCB, Laboratório de Genética e Taxonomia de Microrganismos, Universidade Estadual de Londrina – UEL, Londrina.

renanappel@gmail.com

A piscicultura corresponde hoje a um dos setores da produção de alimentos que mais cresce em todo o mundo. No entanto, este crescimento baseia-se em algumas características como produção intensiva, alta estocagem dos animais e altas densidades de arraçoamento, o que propicia o desenvolvimento de enfermidades, principalmente de origem bacteriana, que acometem os animais e trazem grandes prejuízos para a produção, além de colocar em risco a saúde do consumidor. Bactérias do gênero *Aeromonas*, principalmente *Aeromonas hydrophila,* têm sido descritas como um dos agentes etiológicos de doenças associadas aos cultivos de peixes. Dentre as variedades de espécies cultivadas na piscicultura brasileira a tilápia do Nilo (*Oreochromis niloticus*) destaca-se, correspondendo a 41,9% do total da produção de peixes em cativeiro. Problemas de sanidade têm levado os produtores, afim de evitar prejuízos, a utilizar antibacterianos para controle das doenças onde o agente etiológico muitas vezes é desconhecido, ocasionando impactos ao ambiente, danos econômicos e risco na saúde pública. Técnicas de biologia molecular, como a PCR, propiciam rapidez e um incremento na sensibilidade e eficiência do diagnóstico auxiliando no manejo da doença. O presente trabalho objetivou a descrição de iniciadores específicos para diagnóstico e estudo de ocorrência de *A. hydrophila*. Para tanto, foram recuperadas, a partir de bancos de dados, sequências de DNA dos genes *recA* e *gyrB* de *A. hydrophila* e de outras quatro espécies relacionadas. Com o auxílio do programa Mega6, as sequências foram alinhadas e regiões com diversidade em cada gene foram escolhidas para a seleção dos marcadores de PCR específicos para *A. hydrophila*. Os iniciadores propostos foram otimizados e testados para especificidade e eficiência de amplificação. Como resultado, os primers foram capazes de amplificar fragmentos do tamanho esperado para a espécies e não foi observado amplificação cruzada com as espécies avaliadas. O marcador descrito será utilizado para monitoramento e diagnóstico da presença de *A. hydrophila* em peixes de criação, auxiliando no manejo e controle das doenças em sistemas de cultivo intensivo de tilápia do Nilo.

Palavras chave: Patogenicidade, primer, tilápia do Nilo.

ANÁLISE *in silico* E CLASSIFICAÇÃO DE GENES *rap-phr* NO GRUPO *Bacillus cereus*

Cardoso, P. F.[1], **Fazion, F. A. P.** [1,2], Vilas-Boas, L. A.[1], Sanchis, V.[2], Perchat, S.[2], Lereclus, D.[2], Vilas-Bôas, G. T.[1]
[1]Universidade Estadual de Londrina, Deptº. Biologia Geral, Londrina, Brasil.
[2]INRA, Unité Micalis UMR 1319, Jouy-en-Josas, França.

O grupo *Bacillus cereus* é composto por bactérias Gram positivas formadoras de esporos, que podem colonizar uma diversidade de hospedeiros e apresentam importância biotecnológica ou médica. Dentre elas estão *B. thuringiensis* e *B. cereus*, que apesar de fenotipicamente diferentes, são geneticamente muito similares, sendo suas principais características fenotípicas codificadas por genes de localização plasmidial. Além dos genes responsáveis pela patogenicidade dessas bactérias, os plasmídeos das espécies do grupo também contêm genes essenciais ao metabolismo da célula, como por exemplo, os genes *rap-phr* que regulam as vias de esporulação, competência e formação de biofilme, através da proteína Rap (regulador de resposta aspartato fosfatase), a qual é inibida pelo peptídeo Phr, um sensor de *quorum*. O objetivo deste trabalho foi identificar, caracterizar e classificar os sistemas *rap-phr* nos genomas das espécies do grupo *B. cereus,* visando a compreensão do papel desses sistemas nestas bactérias e dos plasmídeos na evolução do poder patológico dessas espécies. Para isso, genomas com sequência completa disponível no banco de dados do NCBI foram selecionados para o levantamento e obtenção das sequências nucleotídicas e proteicas dos sistemas *rap-phr*. As sequências foram analisadas quanto ao tamanho, localização (plasmidial ou cromossomal), posição dentro do replicon e quanto à similaridade entre elas. Para agrupar os genes em classes foi realizado o alinhamento por Muscle e a árvore filogenética por Neighbor-Joining. Foram analisados os genomas de 49 linhagens das sete espécies do grupo *B. cereus*. Os sistemas *rap-phr* estão presentes em todas linhagens analisadas, sendo que o número de genes *rap* variou de dois a 16 entre elas e a média de genes *rap* foi maior em *B. thuringiensis* do que em *B. cereus*. Cerca de 80% dos genes *rap* apresentam o gene *phr* downstream. Destaca-se a grande quantidade de genes plasmidiais nas linhagens de *B. thuringiensis*, seis vezes maior do que em *B. cereus* e a presença de sistemas *rap-phr* em 36,8% dos plasmídeos de *B. thuringiensis* e em 24% dos plasmídeos de *B. cereus*. A partir da árvore filogenética foi possível separar os 302 genes *rap* em 17 grupos. A presença de genes *rap-phr* plasmidiais em um grande número de plasmídeos do grupo *B. cereus*, especialmente em linhagens de *B. thuringiensis*, sugere um papel fisiológico importante de seus produtos e dos plasmídeos.

Palavras-chaves: *Bacillus thuringiensis*, *Bacillus cereus* (*stricto sensu*), plasmídeo, *quorum sensing*.
Suporte financeiro: Programa CAPES/COFECUB (processo 816/14).

Incorporação da distância de Robinson-Foulds em algoritmos genéticos multi-objetivo para o problema de inferência filogenética

M. Villalobos-Cid[1], M. Dorn[2], R. Ligabue-Braun[3] e M. Inostroza-Ponta[2]

[1]Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Santiago, Chile
[2]Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil
[3]Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

A inferência filogenética se refere ao processo utilizado para estabelecer uma hipótese sobre as relações evolutivas entre um grupo de organismos. O problema de inferência filogenética pode ser tratado como um desafio de otimização onde, se tem como objetivo encontrar a *"melhor árvore"* dentre todas as suas possíveis topologias de acordo com um critério definido. Devido ao elevado número de possíveis soluções, este problema tem sido classificado na área de complexidade computacional como *NP-hard*. Ao longo das últimas décadas, foram propostas diferentes abordagens baseadas em heurísticas para inferir árvores filogenéticas. Frequentemente, estas abordagens estão baseadas em um único objetivo/critério. Não obstante, os resultados obtidos por estas abordagens variam dependendo do método utilizado e do critério selecionado, podendo em alguns casos apresentar topologias conflitantes entre si. Recentemente, foram propostos diversas estratégias computacionais baseadas em otimização multi-objetivo para o problema de inferência filogenética. Estas abordagens tornaram possível a redução do viés associado com a seleção de um critério em particular. Além disto, apresentam em suas fronteiras de Pareto novas soluções não consideradas pelos métodos baseados em otimização de um único objetivo. Apesar deste progresso, os extremos das fronteiras de Pareto resultantes dos métodos baseados em multi-objetivos ainda não conseguem superar os resultados oriundos de métodos clássicos que consideram um único objetivo. Adicionalmente, os métodos de um único objetivo consideram diversidade de soluções em relação aos critérios utilizados, independentemente da diversidade de topologias das árvores. Este trabalho propõe um novo algoritmo genético multi-objetivo para o problema de inferência filogenética considerando os critérios de parcimônia e verossimilhança. Com objetivo de assegurar a diversidade de soluções, essa proposta, além da distância de aglomeração (*crowding distance*), incorporou uma distância de edição entre topologias de árvores (*distância Robinson-Foulds*). Além de levar em consideração árvores com diferentes topologias, as soluções obtidas com essa nova proposta se mostram superiores, ou ao menos não-dominadas, quando comparadas com os resultados provenientes de ferramentas baseadas em único objetivo e as recentes aproximações multi-objetivas.

Palavras-chave: inferência filogenética, otimização multi-objetivo, algoritmo genético, *Robinson-Foulds*.

Agradecimentos: MV-C and MI agradecem ao `CITIAPS-PMI USA1204` e Dicyt 061619IP pelo suporte financeiro.
MD and RL agradecem ao apoio financeiro recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (`CNPq`); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (`CAPES`); e Fundação de Amparo a Pesquisa do Estado do Rio Grande do Sul (`FAPERGS`). Esta pesquisa recebeu suporte da `Microsoft Azure` for Research Award.

SEPredictor: Um sistema para predição de secretoma e efetores de fitonematóides

FH Sano[1], VS Lopes-Caitar[2,3], FC Marcelino-Guimaraes[3] e AY Kashiwabara[1]
[1]Departamento de Computação, Programa de Pós-Graduação em Bioinformática - PPGBIOINFO, Universidade Tecnológica Federal do Paraná - Campus Cornélio Procópio
fabiosano@alunos.utfpr.edu.br
[2]Universidade Estadual de Londrina - UEL, Londrina
[3]Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA, Londrina

Efetores são moléculas secretadas por fitopatógenos que alteram a resposta da planta hospedeira, facilitando a sua infecção e parasitismo. Análises genômicas e transcriptômicas de fitonematóides têm permitido a prospecção de genes codificadores de tais proteínas, o que tem permitido a elaboração de estratégias moleculares que visam o desenvolvimento de plantas tolerantes e com resistência durável a estes parasitas. Embora haja uma ampla gama de proteínas secretadas sendo descritas, ainda é um grande desafio predizer tais moléculas. Isto deve-se a diversos fatores, incluindo a não existência de bancos de dados e algoritmos de predição específicos para genes efetores de fitonematóides. Para auxiliar a predição de proteínas efetoras de fitonematoides, este trabalho visa construir um sistema fundamentado nas características de sequências efetoras, já descritas, destes patógenos, com banco de dados guiado por ontologias e com preditores moldados para suas sequências, facilitando assim os estudos de genômica comparativa e a prospecção de novos candidatos a proteínas envolvidas no parasitismo. Na construção do sistema será utilizado o banco de dados Chado, por tratar-se de um esquema baseado em ontologias o que, consequentemente, permite representar informações biológicas associadas ao genoma. Entretanto, serão utilizados apenas os módulos e tabelas necessários para gerenciar o domínio biológico dos fitonematóides. Para os preditores será proposta uma metodologia de bioinformática que utilize abordagem ab initio, como reconhecimento de padrões, e que permite fazer predições *in silico* de genes do secretoma de fitonematóides, que constituem potenciais efetores. Esta metodología utilizará métodos computacionais como, por exemplo, modelos probabilísticos de sequência biológica, pois eles fornecem formas de representar a distribuição de probabilidade de famílias de sequências. Espera-se através deste sistema fornecer um ambiente prático e de fácil acesso para a realização de estudos na área, além de organizar e armazenar todas as sequências preditas como secretadas e efetoras destes parasitas de plantas, tanto as que já foram publicadas e previamente validadas in vivo, quanto as que serão preditas através do sistema.

Palavras-chave: Secretoma, Efetores, Fitonematóides, Predição, Modelos Probabilísticos, Reconhecimento de Padrões, Bioinformática.

Agradecimentos: Apoiado por: CNPq 476689/20139.

Investigação de respostas imunes mediadas por CD40 dinamicamente distintas que conduzem à proteção do hospedeiro ou susceptibilidade à infecção por *Leishmania donovani*

Frank Brombacher[1], Sergey Kiselev[2], Marcel Joly[3], Xiaopeng Xu[4] and Bhaskar Saha[5]

[1]Department of Pathology University of Cape Town; [2]Vavilov Institute of General Genetics RAS, Moscow; [3]UTFPR; [4]School of Life Sciences, Sun Yat-sen University; [5]National Centre for Cell Science, Ganeshkhind

Leishmania spp. são protozoários parasitas que causam a leishmaniose em 88 países, incluindo o Brazil. A elevação do número de pacientes de Leishmaniose dérmica pós-kala azar funciona como um reservatório para Leishmania e aumenta a chance de uma epidemia no futuro próximo. Portanto, elucidar os mecanismos biológicos de patogenicidade é crucial para derivação de uma nova intervenção terapêutica. A via de sinalização CD40-CD40L tem sido relacionada com a susceptibilidade à infecção por Leishmania major e Leishmania donovani, embora pouco seja conhecido acerca do(s) mecanismo(s) controlador(es). Para caracterizar o papel da sinalização CD40 sobre a resposta imune pró-leishmanial ou anti-leishmanial, uma colaboração científica interdisciplinar entre os países dos BRICS (Brasil, Rússia, Índia, China e África do Sul) foi recentemente estruturada. Em termos metodológicos, quatro abordagens experimentais serão consideradas: [1] estudo da dinâmica do processo infeccioso e de transcriptomas de células T em camundongos portadores de deleção célula-específica para CD40 em células hospedeiro importantes, macrófagos, células dendríticas (DCs) e células B, criadas empregando o sistema cre-LoxP de recombinação sítio-específica sob promotores específicos, LysMcre, CD11ccre e mb1cre, respectivamente; [2] comparação entre funções das células T mediadas por células B, macrófagos e células dendríticas em camundongos selvagens, deficientes para CD40 e deficientes para CD40 lentiviralmente reconstituídos para CD40 no contexto da infecção por L. donovani; [3] avaliação da dinâmica de infecção por L. donovani e transcriptomas de células T em camundongos BALB/c silenciados para CD40 por CD40shRNA expresso lentiviralmente; [4] avaliação da função do CD40 específica a células B, macrófagos e células dendríticas em instantes de tempo definidos após infecção por L. donovani em linhagens susceptíveis ou resistentes de cepas consanguíneas de camundongos com haplótipo similar. O grupo sul-africano, o qual gerará e caracterizará camundongos com deficiência célula-específica para CD40, e o grupo indiano irão examinar as dinâmicas da patologia e imunologia da infecção por L. donovani. Considerando-se os estados inicial e final in vivo, a modulação epigenética do genoma e a rede miRNA induzida por CD-40 serão analisados pelos grupos russo e chinês, respectivamente. O grupo brasileiro trabalhará o desenvolvimento de uma abordagem de biologia de sistemas para suportar a elucidação dos mecanismos responsáveis pela susceptibilidade ou resistência à infecção modulados por CD40. A construção e validação do modelo de biologia teórica serão executadas através da combinação de modelagem orientada por dados considerando a informação experimental a ser produzida pelos demais grupos.

Palavras-chave: biologia de sistemas, imunoterapia, Leishmania, modelagem matemática, otimização numérica, Pesquisa Operacional, programação matemática, simulação computacional, sinalização celular, sistema imune

WB 2016 – Workshop de Bioinformática da UTFPR (05/10/2016 – 07/10/2016)

Título: Caracterização, clonagem e avaliação entomopatogênica dos genes cry de *Bacillus thuringiensis* BR58 efetivos no controle de *Hypothenemus hampei* (Ferrari) (Coleoptera: Curculionidae: Scolytinae)

Autores, instituição de vínculo e nome do primeiro autor: ALTRÃO, C.S.; RICIETTO, A.P.S.; VILAS-BÔAS, G.T.; CARDOSO, P.F.; GONÇALVES, K.B.; da SILVA, C.R.M.; VILAS-BOAS, L.A. Universidade Estadual de Londrina – UEL. Carla Suzuki Altrão

Palavras chave: Fitossanitário, controle biológico, primers

O agronegócio do café brasileiro desempenha relação direta com a direção da economia e da política do país, que se destaca, mundialmente, por ser o maior produtor e exportador do grão. No entanto, os cafezais são suscetíveis a ataques de pragas que causam danos nas lavouras e perdas econômicas. A broca-do-café, *Hypothenemus hampei* (Ferrari, 1867) (Coleóptera: Curculinoide: Scolytinae), é considerada uma das pragas de maior risco fitossanitário e importância econômica na produção desta cultura. A utilização de agentes químicos no controle destes insetos tem sido questionada por apresentarem componentes que podem causar problemas, como a contaminação ambiental e humana, além de incrementar os custos de produção. Uma das propostas para contornar o problema seria a construção de plantas geneticamente modificadas expressando genes tóxicos para a praga. Uma das estratégias mais usadas é a produção por parte das plantas, de proteínas Cry, produzidas por *Bacillus thuringiensis,* uma bactéria entomopatogência, ativa contra a broca do café. Recentemente foi demonstrado que o isolado *B. thuringiensis* BR58, contém os genes *cry4A, cry4B*, *cry10A*, *cry11A*, *cry60A*, *cry60B* e apresenta atividade inseticida para *H. hampei.* Neste contexto, com o presente trabalho objetivou-se, a partir de um genoma produzido pelo grupo, fazer a caracterização, clonagem e avaliação da atividade entomopatogênica de cada um dos genes *cry* preditos na linhagem *B. thuringiensis* BR58. Para tanto, foram construídos iniciadores específicos desenvolvidos a partir do alinhamento das sequências dos genes da linhagem BR58 depositadas no banco de dados do NCBI com o auxílio do programa Mega6. Aos primers foram adicionados sítios de enzimas de restrição *BamHI*, *HindIII* e

SalI, dependendo da estratégia de clonagem adotada para cada gene e avaliados com o software Snapegene (GSL Biotech). Para comprovar se as seqüências dos genes codificam para as proteínas esperadas, foram realizados alinhamentos pelo algaritmo BLAST. Ademais, as seqüências foram avaliadas para a confirmação das regiões promotoras e das Orfs com a utilização das ferramentas de software Softberry e Orfinder. A avaliação dos iniciadores quanto à formação de estruturas secundárias como grampos e dímeros foi realizada com o uso da ferramenta de análise Oligonalyzer 3.1. Cada gene *cry* uma vez inserido no vetor pHT315*xyl*, utilizado para superexpressão de genes serão clonados nas *Escherichia coli* TG1 e ET12527. Os vetores serão purificados e transformados na linhagem *B. thuringiensis* 407^{-}, onde ocorrerá a indução de sua expressão para que a ação inseticida das proteínas sejaavaliada individualmente frente à *H. hampei*.

Clonagem e expressão das proteínas binárias Vip1/Vip2 de um isolado de *Bacillus thuringiensis*

Ricietto, A.P.S[1], Gomis-Cebolla, J.[2] Vilas-Bôas, L.A.[1] Ferré, J.[2] e Vilas-Boas, G.T.F.L[1]
[1]Departamento Biologia Geral, Laboratório de Genética e Taxonomia de Bactérias, UEL, Londrina, Brasil
ricietto@gmail.com
[2]Laboratório de Genética, Universidad de Valencia, Burjassot, Espanha

Bactérias entomopatogênicas apresentam um enorme potencial para o controle de insetos por nos oferecer uma grande variedade de compostos inseticidas. A mais utilizada e conhecida bactéria empregada no controle biológico é o *Bacillus thuringiensis,* que sintetiza proteínas com atividade inseticida, como as proteínas Cry e Cyt (que acumulam-se em cristais parasporais) e as proteínas entomopatogênicas Sip e Vip, que são secretadas para o meio extra-celular. Este último grupo tem sido estudado como uma nova estratégia para o controle biológico, afim de prevenir ou impedir o desenvolvimento de resistência pelos insetos. Proteínas Cry e Vip apresentam modos de ação diferentes, essa característica, possibilita a combinação de ambas proteínas no desenvolvimento de novos produtos biotecnológicos como plantas transgênicas. A linhagem utilizada nesse estudo apresenta os genes *vip1* e *vip2* que codificam para proteínas com função binária e com ação inseticida para a ordem Coleoptera. Neste contexto, com o presente trabalho objetivou-se, caracterizar, clonar e avaliar a atividade entomopatogênica desses genes. Para tanto, foram construídos iniciadores específicos com os sítios de enzimas de restrição *BamHI*, *HindIII* e *SalI*, integrados de acordo com a estratégia de clonagem. A avaliação dos iniciadores quanto à formação de estruturas secundárias foi realizada com o uso da ferramenta de análise Oligonalyzer 3.1 e a estratégia para clonagem foram avaliadas com o software Snapgene (GSL Biotech). As sequências foram avaliadas para a confirmação das regiões promotoras e das Orfs com a utilização das ferramentas de software Softberry e Orfinder. Cada um dos genes *vip,* uma vez inserido no vetor pHT315*xyl*, utilizado para superexpressão de genes serão clonados em *Escherichia coli* TG1 e ET12527. Os vetores serão purificados e transformados na linhagem *B. thuringiensis* 407⁻, onde ocorrerá a indução de sua expressão para que a ação inseticida das proteínas seja avaliada frente à diferentes insetos da ordem Coleoptera.

Palavras-chave: proteínas Vip, resistência à insetos, plantas transgênicas

Agradecimentos: CAPES.

Bioinformática para investigação de dados públicos de mirtrons

Bruno Henrique Ribeiro da Fonseca[1], Tamires Priscila da Costa[2], Douglas Silva Domingues[1,3] e Alexandre Rossi Paschoal[1]

[1] Programa de Pós-Graduação em Bioinformática - PPGBIOINFO, Universidade Tecnológica Federal do Paraná, Cornélio Procópio
[2] Universidade Tecnológica Federal do Paraná, Cornélio Procópio
[3] Departamento de Botânica, Instituto de Biociências, Universidade Estadual Paulista, Rio Claro

* Autor correspondente: paschoal@utfpr.edu.br

MicroRNA (miRNAs) é uma das classes de RNAs não-codificantes (ncRNAs) presente em genomas de eucariotos, que tem como principal papel biológico a regulação pós-transcricional dos níveis de RNA mensageiro em células. Em 2007 foram descritos os mirtrons, um tipo de miRNA que tem biogênese alternativa ao miRNA canônico, pois utiliza-se do processo de *splicing* para sua formação, ao invés da tradicional utilização de enzimas específicas de clivagem.
Para seis organismos eucariotos com genomas sequenciados foram descritos mirtrons: *Arabidopsis thaliana*, *Oryza sativa, Homo sapiens, Mus musculus, Drosophila melanogaster* e *Caenorhabditis elegans*. Os dados públicos disponíveis destes genomas foram investigados.
Na análise exploratória realizada, especificamente foram consideradas as características de sequência e estrutura dos dados de mirtrons, dentre eles: conteúdo GC, tamanho, composição de nucleotídeos (K-mer até 3), energia livre e conservação. Foram realizadas também análise comparativa dessas características com os miRNAs canônicos dos mesmos genomas. Para esta etapa, scripts PERL e em R studio foram desenvolvidos e aplicados.
Durante a análise dos dados foi identificado que todo conteúdo está distribuído em fontes distintas, sem qualquer tipo de organização ou estrutura de banco de dados ou sistema. Assim, a criação de um repositório central específico para mirtrons é uma importante ferramenta de estudos deste tipo de ncRNAs.
Este projeto objetiva duas principais contribuições: (i) análise exploratória de características (sequência e estrutura) dos mirtrons; e (ii) o desenvolvimento de um repositório web amigável para centralização e organização dos dados. O repositório permitirá exploração sobre o conhecimento público de mirtrons, bem como análises de diferenciações dos miRNA canônicos.
O desenvolvimento do repositório central de dados de mirtrons está divido em três etapas: (i) análise e coleta de dados disponíveis na literatura atual; (ii) mapeamento, integração, e consistência dos dados, e (iii) o desenvolvimento de interface web amigável para consulta dos dados de mirtrons. Para a criação do repositório de mirtrons será utilizada a linguagem PHP, a partir do esquema de banco de dados biológico Chado.
Por fim, acredita-se que esta pesquisa possibilita a futura aplicação de métodos e técnicas computacionais em bioinformática no conjunto dos dados públicos sobre mirtrons.

Palavras-chave: mirtron, RNA não-codificante, repositório.

Agradecimentos: Edital Universal - CNPq Nº. 14/2014 - Projeto: 454505/2014-0

Carboximetilcelulase (CMCase) minerada a partir de um banco de dados metagenômico

Gilberto de Aguiar Pereira[1]; Fernando Gomes Barcellos[1].

1 – Departamento de Biologia Geral, Laboratório de Genética de Micro-organismos, Universidade Estadual de Londrina, Londrina. E-mail: gilbertopperera@gmail.com

Estudos metagenômicos podem sofrer com erros de anotação relacionados a sensibilidade das ferramentas utilizadas na execução do seu *workflow*, neste sentido, a utilização de programas que não determinam alinhamentos significativos unicamente através dos símbolos observados, se colocam como alternativas para extrair informações significativas de conjuntos de dados considerados subutilizados. Assim, o objetivo deste estudo foi minerar um banco de dados de metagenomas, baseando-se nos perfis *Hidden Markov Models* (HMM) para sequências de CMCases (EC 3.2.1.4). Para isso, sequências não redundantes de CMCases obtidas a partir do banco de proteínas curado Swiss-Prot foram alinhadas, processadas e convertidas ao formato Stockholm com a utilização dos programas SATé, FigTree, AliView e Convert Sequence; em seguida, o alinhamento processado e convertido participou da construção do perfil HMM, através da utilização da ferramenta *hmmbuild* (HMMER 3.0), e o perfil HMM construído foi utilizado para minerar o banco de dados EBI metagenomics, a partir do qual, CDSs sem anotações, provenientes de projetos classificados como pertencentes ao bioma solo, foram utilizadas para o desenvolvimento de análises locais a partir da ferramenta *hmmsearch* (HMMER 3.0). Adicionalmente o *Blast* do banco de dados UniProt, o *Conserved Domain Database* e o *Blast Tree View* do NCBI foram utilizados para interpretar os resultados obtidos. Nossas análises apontaram então que 27% dos projetos metagenômicos incluídos neste estudo, apresentavam ao menos um hit para o perfil HMM pesquisado, no entanto, de acordo com o *Conserved Domain Database*, apenas o *hit* relacionado a um projeto sobre deposição de matéria orgânica no solo (Identificação EBI - ERR1303295) apresentou similaridade com um domínio pertencente à família 8 de uma glicosídeo hidrolase. Análises subsequentes indicaram ainda que a sequência de aminoácidos pertenciam a uma gama proteobactéria (*Blast Tree View*) e ainda que de acordo com *Blast* Uniprot esta enzima poderia ser nomeada como uma 1,4-D-glucanase (sinônimo de CMCase). Desta forma, é possível concluir que o programa HMMER 3.0 é uma boa alternativa na busca por similaridade de sequências de aminoácidos no banco EBI metagenomics já que conseguiu localizar rapidamente e corretamente enzimas CMCases de acordo com as sequências que lhe foram fornecidas; oferecendo ainda como um grande diferencial a disponibilização completa dos *hits* obtidos, possibilitando assim sua validação *in silico* e o desenvolvimento de novos estudos a partir destas sequências.

Palavras-chave: HMM, CMCase, Mineração de metagenomas.

Agradecimento: CAPES.

Análise da Diversidade Genotípica no Algoritmo Evolução Diferencial Aplicado ao Problema de Predição de Estrutura de Proteínas

Pedro Henrique Narloch e Rafael Stubs Parpinelli
Programa de Pós-Graduação em Computação Aplicada, Centro de Ciências Tecnológicas,
Universidade do Estado de Santa Catarina, Joinville
dcc6phn@joinville.udesc.br, rafael.parpinelli@udesc.br

Resumo

O problema de predição de estrutura de proteínas que utiliza somente a informação contida na cadeia de aminoácidos, conhecido como *Ab Initio*, é um dos problemas mais desafiadores da bioinformática devido ao seu alto nível de complexidade. Essa complexidade está relacionada com a grande quantidade de formas tri-dimensionais que uma proteína pode assumir, principalmente se o detalhamento da estrutura se der em nível atômico, caracterizando o problema como NP-Completo. Diversos algoritmos bio-inspirados já foram propostos para predizer a conformação nativa de proteínas utilizando representações atômicas. Entretanto, nenhum dos trabalhos existentes na literatura fazem a análise de diversidade durante o processo de otimização. O presente trabalho aplica o algoritmo Evolução Diferencial (ED) e explora o uso de dois mecanismos bastante simples de diversificação conhecidos como: *generation gap* (GG) e perturbação gaussiana (GP) com o objetivo de aumentar a diversidade genotípica da população e verificar seu impacto durante o processo de otimização. A proteína escolhida para este estudo inicial é a proteína 1PLW. A partir desse monitoramento é possível verificar se a diversidade genotípica gerada por métodos bastante simples é um fator que deve ser considerado na solução do problema de predição de estrutura de proteínas. A representação tri-dimensional adotada no trabalho é o modelo atômico que considera os ângulos e torsões da cadeia principal e da cadeia lateral. Para avaliar as soluções encontradas a função de energia CHARMM é utilizada. Por meio dos resultados obtidos em dez execuções independentes, o algoritmo que utilizou as duas técnicas de diversificação (ED_{GG-GP}) obteve energia mínima de $-35,82$ kcal mol^{-1} com média e desvio padrão de $-30,47 \pm 4,44$. O resultado obtido é competitivo com o algoritmo tido como estado-da-arte da literatura conhecido como ADEMO/D que, para a mesma proteína, alcançou valor de energia mínimo de $-30,43$ kcal mol^{-1}. Quando levado em consideração a relação entre convergência da energia e manutenção de diversidade, pôde-se observar que a manutenção de diversidade feita pela versão ED_{GG-GP} foi um fator importante para a obtenção de resultados melhores que a versão canônica da ED e competitivos com os da literatura. A aplicação desse algoritmo em proteínas com estruturas mais complexas que a 1PLW é um dos trabalhos futuros de pesquisa, bem como a exploração de novos mecanismos que possam influenciar na manutenção da diversidade genotípica.

Palavras-chave: Predição de Estrutura de Proteínas, Bioinformática, Evolução Diferencial, Diversidade Genotípica.

Técnicas de Aprendizado Ativo para Classificação do Vigor de Sementes de Soja

Douglas Felipe Pereira[1], Guilherme Camargo[1], Pedro Henrique Bugatti[1] e Priscila Tiemi Maeda Saito[1,2]

[1]Programa de Pós-graduação em Bioinformática (PPGBIOINFO), Universidade Tecnológica Federal do Paraná (UTFPR-CP)

[2]Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)

Email: {douglaspereira, gcamargo}@alunos.utfpr.edu.br, {pbugatti, psaito}@utfpr.edu.br

Oferecer grãos de qualidade ao produtor é um dos desafios do setor agroindustrial de grãos como os da soja. Para atingir tal qualidade em sementes de soja aplica-se amplamente o chamado teste de tetrazólio. Tal teste visa definir o vigor da semente, bem como apontar os tipos de danos encontrados na mesma, como por exemplo os causados por umidade, por percevejo ou mecânicos. Técnicas de processamento de imagens e de aprendizado já vêm sendo estudadas para aplicação no problema de classificação de sementes. No entanto, nenhuma delas faz o uso de técnicas de aprendizado ativo. Métodos de aprendizado ativo podem selecionar um conjunto razoavelmente pequeno de amostras relevantes para a criação de um modelo de classificação e utilizar o especialista para verificação/correção de amostras classificadas incorretamente em poucas iterações de aprendizado. Dessa forma, o presente trabalho visa explorar uma metodologia de aprendizado ativo para a classificação do vigor de sementes de soja, oriundas do teste de tetrazólio. Para tanto, foi utilizado o método de aprendizado ativo *Root Distance-Based Sampling* (RDS), o qual realiza um pré-processamento, organizando a priori as amostras do conjunto. Essa estratégia de organização consiste em realizar o agrupamento das amostras e criar, para cada cluster, uma lista ordenada de amostras de acordo com a distância para a raiz do respectivo cluster. A estratégia de seleção consiste em selecionar amostras diversas (amostras de classes distintas) e incertas (amostras mais difíceis de serem classificadas). Para tanto, é obtido um conjunto de amostras de cada lista (diversidade), e em cada lista são priorizadas amostras cujo rótulo, dado pela instância atual do classificador, é distinto do rótulo da raiz do cluster correspondente. O processo de aprendizado torna-se mais rápido, uma vez que não requer a classificação e re-organização de todas as amostras do conjunto de dados a cada iteração. Para os experimentos foram realizados comparações entre RDS e Rand, em que as amostras são selecionadas aleatoriamente do conjunto de dados. Para a classificação, foram utilizados: OPF, SVM, RF, J48, MLP, NB e k-NN. Os resultados mostraram que RDS com o classificador OPF apresentou maiores acurácias e quantidade de classes distintas mais rapidamente, reduzindo as iterações de aprendizado, bem como o tempo computacional.

Palavras-chave: Aprendizado ativo, análise de imagens, processamento de imagens, classificação, sementes de soja

Agradecimentos: Os autores gostariam de agradecer à CAPES, CNPq, Fundação Araucária, UTFPR, Belagrícola e SETI pelo apoio financeiro.

Aprendizado Ativo para Classificação de Bioimagens

Guilherme Camargo[1], Douglas Felipe Pereira[1], Pedro Henrique Bugatti[1] e Priscila Tiemi Maeda Saito[1,2]

[1]Programa de Pós-graduação em Bioinformática (PPGBIOINFO), Universidade Tecnológica Federal do Paraná (UTFPR-CP)

[2]Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)

{gcamargo, douglaspereira}@alunos.utfpr.edu.br, {pbugatti, psaito}@utfpr.edu.br

Grande avanço tecnológico tem acontecido nos últimos anos, através de novos dispositivos de captura de dados (imagens, sons ou textos). Com isso, obtêm-se grandes bases de dados diariamente. Para o armazenamento e posterior recuperação de informações a partir dessas bases, especialistas devem realizar a anotação das amostras. No entanto, a anotação pode trazer inconsistências aos dados, já que os indivíduos podem interpretá-los de maneira divergente. Além disso, o custo para realizar a tarefa é elevado e exaustivo para os especialistas. Logo, uma solução para o problema seria automatizar o processo de identificação das amostras, utilizando métodos computacionais em que um rótulo (informação textual) é atribuído a cada amostra, classificando-a de acordo com o escopo do problema. Uma maneira de desenvolver a solução é por meio de técnicas de aprendizado de máquina, para construção de um classificador de padrões. Trabalhos da literatura, em geral, consideram todas as amostras do conjunto para treinamento do classificador. No entanto, podem haver amostras redundantes na base de dados, assim como amostras mais relevantes para o aprendizado do classificador. Técnicas de aprendizado ativo são interessantes nesse contexto, já que uma quantidade reduzida de amostras mais informativas é selecionada para o treinamento do classificador. A maioria das técnicas de aprendizado ativo propostas na literatura não levam em consideração o tempo computacional, já que executam os processos de classificação, organização e seleção de amostras mais significativas a cada iteração, bem como consideram todo o conjunto de dados. Neste trabalho é proposto um novo método mais efetivo e eficiente para o aprendizado ativo, em que a redução e a organização de um subconjunto de amostras mais significativas ocorrem uma única vez. Além disso, não são classificadas todas as amostras do conjunto a cada iteração. Resultados preliminares mostram que a utilização do método proposto alcança melhores resultados em relação aos métodos de aprendizado ativo tradicionais, onde não há pré-seleção de amostras mais informativas, sendo os dados selecionados aleatoriamente do conjunto de treinamento. Os tempos computacionais necessários para treinamento e teste do classificador mostram-se equivalentes para ambos os métodos. Já o tempo de seleção das amostras mais informativas é naturalmente maior para o método proposto, já que os dados são previamente organizados. Contudo, vale ressaltar que o método proposto apresenta tempos de resposta interativos, além de apresentar conhecimento amplo da base mais rapidamente que o método aleatório.

Palavras-chave: Aprendizado ativo, análise de imagens, classificação, floresta de caminhos ótimos, bioimagens.

Agradecimentos: CAPES, CNPq, Fundação Araucária, UTFPR e SETI

Índice de autores