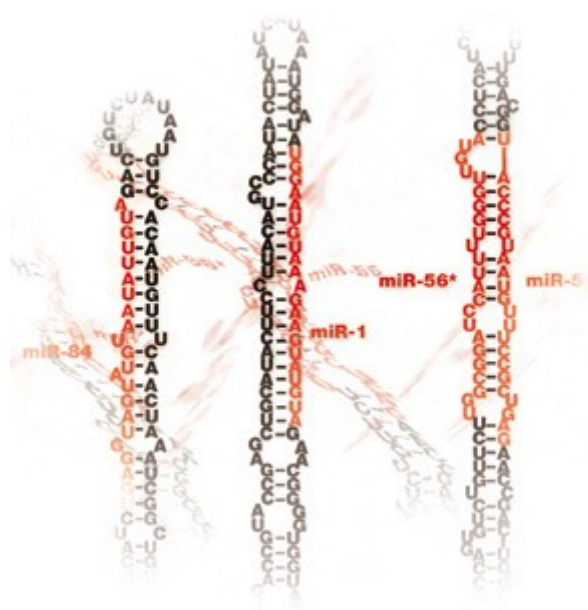


UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO E
FACULDADE DE MEDICINA DE RIBEIRÃO RETO

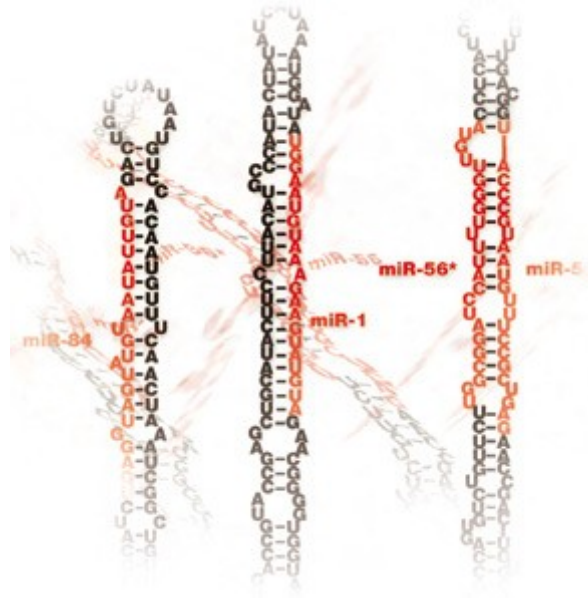


DESENVOLVIMENTO DE UMA FERRAMENTA PARA
IDENTIFICAÇÃO DE MICRORNAS

Raony Guimarães Corrêa Do Carmo Lisboa Cardenas
raony@informaticabiomedica.com.br

RIBEIRÃO PRETO – SP
NOVEMBRO DE 2008

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO E
FACULDADE DE MEDICINA DE RIBEIRÃO RETO



DESENVOLVIMENTO DE UMA FERRAMENTA PARA
IDENTIFICAÇÃO DE MICRORNAS

Monografia apresentada à Faculdade de Filosofia Ciências e Letras e à Faculdade de Medicina de Ribeirão Preto para obtenção do título de bacharel em Informática Biomédica.

Orientadora: Prof^a. Silvana Giuliatti

Co-orientador: Prof. Geraldo Aleixo Passos

RIBEIRÃO PRETO – SP

NOVEMBRO DE 2008

"E não sabendo que aquilo era impossível,
ele foi lá e fez"
(JEAN COCTEAU)

AGRADECIMENTOS

A Professora Dra. Silvana Juliatti, pela orientação, dedicação e apoio durante todas as fases importantes ao longo de toda a minha graduação. Em especial pelo voto de confiança que recebi e por sempre estar preocupada com a opinião dos alunos na decisão de assuntos importantes relacionados ao curso de Informática Biomédica.

Ao Professor Dr. Geraldo Aleixo Passos pela atenção, inspiração, pelas broncas, e excelentes sugestões nesse trabalho. Por servir de exemplo de uma carreira excepcional de pesquisador.

A Allyne Oya Chiromatzo, por ter discutido bastante comigo sobre o projeto e pelas centenas de artigos que me emprestou sobre microRNAs.

Ao Luciano Ângelo de Souza Bernardes, pela paciência, atenção e amizade que foram essenciais no desenvolvimento deste trabalho e durante minha formação acadêmica.

Ao Francis de Moraes Franco Nunes, por toda a atenção ao trabalho, em especial pela ajuda na compreensão dos processos biológicos e pelas sugestões extremamente valiosas sobre a importância do trabalho e as possíveis análises futuras.

Ao Jerônimo Ceron Ruiz, por ter sido a pessoa que me ajudou a dar os primeiros passos no mundo da Bioinformática.

Agradeço ao Prof Eric C. Lai, pelos e-mails trocados e pela atenção dedicada.

Agradeço ao Dr. Kave Mohtashami, CEO da Rubisco Bioinformatics pela compreensão de alguns processos biológicos.

Agradeço a todos que, de alguma forma, permitiram que esse trabalho fosse realizado e deram o apoio necessário para a sua concretização.

DEDICATÓRIA

Este trabalho é dedicado a minha querida avó Edna Guimarães Corrêa, por ter acreditado em mim desde o começo, pelos longos anos de dedicação e pela sua enorme insistência.

Pelas brigas, discussões e puxões de orelha desde a infância. Pelos momentos difíceis que passamos juntos e por nunca ter deixado de acreditar que um dia eu alcançaria meus objetivos.

Por me ensinar a correr atrás dos meus sonhos e me mostrar que nada é impossível de ser realizado quando se tem determinação.

Dedico a minha mãe por tudo que passamos juntos, por ter procurado me dar a melhor educação que fosse possível e apesar das brigas por nunca ter desistido de lutar.

Dedico este trabalho a minha namorada Danúbia Midori Berto Fujita, pois sem seu carinho, apoio e enorme paciência eu não teria conseguido chegar até aqui. Por todo o amor que recebo todos os dias e por ter superado junto comigo todos os momentos difíceis que passamos juntos. Por me fazer ser uma pessoa melhor e por todos os conselhos e recomendações que sempre recebo com muito carinho.

Dizem que por trás de um grande homem sempre existe uma grande mulher, no meu caso foram três.

Dedico este trabalho a todos os meus familiares que acreditaram e me deram suporte ao longo de toda a minha formação.

Dedico este trabalho aos amigos da faculdade pelas noites que passamos acordados estudando e as vezes bebendo pelos bares da cidade.

Dedico este trabalho a todos aqueles que fazem parte da minha vida e que de alguma forma contribuíram para a concretização deste trabalho.

Dedico este trabalho a todos aqueles que trabalham com Bioinformática ou

microRNAs e que dedicam sua pesquisa a este ramo da ciência.

Dedico este trabalho a todos que sonham em mudar o mundo e tentam fazer com que esse sonho se realize todos os dias da sua vida.

ABREVIATURAS

DNA – Ácido Desoxirribonucleico

RNA – Ácido Ribonucleico

kB – kilobase

pB – Pares de Base

Script – Linguagem de computador interpretada, uma série de instruções formais escritas para um interpretador.

Formato FASTA – Formato de apresentação de seqüências biológicas, onde para cada seqüência existe uma linha de identificação começando com o símbolo ">" e que descreve a seqüência com um nome e outras informações, sendo seguida por outras linhas contendo a seqüência propriamente dita em um total de 60 a 80 caracteres por linha.

Perl (Practical Extraction and Reporting Language) – Linguagem de programação interpretativa, bastante popular que vem sendo extensivamente utilizada em diferentes áreas como programação de web e Bioinformática.

Bash – É um interpretador de comandos, uma espécie de tradutor entre o Sistema Operacional e o usuário, normalmente conhecido como *shell*. Permite a execução de seqüências de comandos diretamente no *prompt* do sistema ou escritas em arquivos de texto, conhecidos como *shell* scripts.

BLAST (Basic Local Alignment Search Tool) – é um programa de pesquisa que procura regiões de similaridade local, sendo muito utilizado para busca em bancos de dados, sendo a ferramenta mais popular para a análise de similaridade de seqüências.

RESUMO DO PROJETO

O objetivo deste projeto é o desenvolvimento de uma ferramenta computacional capaz de identificar microRNAs em organismos que ainda não possuam este tipo de estrutura determinada. Para isso foram utilizados microRNAs já identificados em outros organismos, considerando que essas estruturas são extremamente conservadas, até mesmo em organismos geneticamente distantes.

Através de metodologias computacionais como o alinhamento de seqüências e a predição de estrutura secundária, foi desenvolvida uma ferramenta computacional capaz de selecionar candidatos utilizando o *e-value*, o *Score* do alinhamento entre as seqüências, o cálculo de energia livre, a penalidade e a temperatura de “fold” da predição da estrutura secundária. Foi utilizado um conjunto de dados de *Epstein barr*, *Paracoccidioides brasiliensis*, *Aspergillus nidulans*, *Aspergillus fumigatus* para que a ferramenta fosse validada e para que fossem identificados novos candidatos a possíveis microRNAs nesses organismos.

Para esse estudo foram utilizados programas como o BLAST, o Mirfold e o RNA Vienna Package que são capazes de alinhar uma seqüência, predizer sua estrutura secundária e gerar uma imagem de representação, respectivamente. Com isso, foi proposto uma automatização da análise, para que os pesquisadores possam submeter suas seqüências e selecionar os melhores candidatos a microRNAs e, posteriormente, fazer uma validação com experimentos de expressão de microRNAs identificados, como por exemplo, Northern-blot, RT-PCR e microarrays. A ferramenta encontra-se disponível para teste no seguinte endereço: <http://gbi.fmrp.usp.br/mirtorch> .

Palavras chave: Bioinformática, MicroRNAs, Alinhamento de Seqüências, Predição de Estrutura Secundária.

Sumário

RESUMO DO PROJETO.....	8
1. INTRODUÇÃO.....	10
1.1 BIOINFORMÁTICA.....	10
1.2 MICRORNAS.....	12
1.3 MOTIVAÇÃO.....	14
1.4 OBJETIVOS GERAIS.....	16
1.5 OBJETIVOS ESPECÍFICOS	17
2. MATERIAL E MÉTODOS.....	19
2.1 OBTENÇÃO DOS DADOS.....	19
2.2 ALINHAMENTO.....	22
2.5 PREDIÇÃO DA ESTRUTURA SECUNDÁRIA.....	24
3. RESULTADOS	26
3.1 CONSTRUÇÃO DA FERRAMENTA.....	26
3.2 BANCO DE DADOS.....	28
3.3 SUBMISSÃO DAS SEQUÊNCIAS.....	34
3.4 VERIFICAÇÃO DOS DADOS.....	37
3.5 ALINHAMENTO.....	38
3.6 EXTRAÇÃO DAS REGIÕES CONSERVADAS.....	41
3.7 PREDIÇÃO DA ESTRUTURA SECUNDÁRIA.....	41
3.8 REPRESENTAÇÃO GRÁFICA.....	41
3.9 VISUALIZAÇÃO DOS RESULTADOS.....	42
3.10 ANÁLISE DOS RESULTADOS.....	48
4. CONCLUSÃO	60
5. TRABALHOS FUTUROS.....	61
6. BIBLIOGRAFIA.....	62

Índice de ilustrações

FIGURA 1: FASES DE UM MICRORNA.....	13
FIGURA 2: NÚMERO DE MICRORNAS IDENTIFICADOS.....	15
FIGURA 3: NÚMERO DE GENOMAS SEQUENCIADOS.....	16
FIGURA 4: MER (MODELO ENTIDADE RELACIONAMENTO) DO BANCO DE DADOS DESENVOLVIDO.....	29
FIGURA 5: ÁREA DE PROJETOS.....	42

1. Introdução

1.1 *Bioinformática*

No início do século XXI, houve uma revolução na era da informação biológica, com o avanço da tecnologia e do aumento exponencial da disponibilidade de informação pública através da internet. Através desse aumento, a enorme quantidade de informação disponível tem gerado, desde então, novos conhecimentos através da fusão de algumas áreas e do trabalho multidisciplinar dos profissionais envolvidos.

Dentro desse contexto, passando pela transição entre as áreas, surgiu uma nova

área de atuação que conecta dois ramos de conhecimento (Biologia e Informática) com o objetivo de aumentar a velocidade de análise dos dados biológicos, extraindo informação de grandes quantidades de dados e facilitando a vida dos pesquisadores no processo de análise de seqüências e estruturas moleculares.

Através do uso de ferramentas de Bioinformática, novas drogas puderam ser descobertas e houve um aumento na velocidade de análise dos dados biológicos. Tratamentos como a terapia gênica com o uso de microRNAs, tornaram-se algo plausível [1]. Surgiram novas possibilidades na área da saúde, como a predição de doenças através de uma análise do genoma [2,3] e, também, a clonagem de células a partir de organismos adultos para auxiliar no tratamento de algumas doenças.

Com o crescimento da disponibilidade de dados biológicos, tornou-se necessário o aparecimento de um profissional que fosse capaz de analisar grandes quantidades de dados em um curto período de tempo, para resolver problemas que muitas vezes exigem um grande poder de processamento computacional e que apresentam uma grande dificuldade de compreensão e solução. Para isso surgiu o Bioinformata, que, utilizando e construindo ferramentas de programação, conseguiu extrair informação e gerar novos conhecimentos a partir de grandes quantidades de dados.

Essa nova área de conhecimento que integra biologia, bioquímica, biologia molecular, matemática, física, química, entre outras ciências, utiliza os recursos computacionais mais avançados disponíveis na atualidade para o gerenciamento, estudo e análise dos dados biológicos.

Seguindo essa visão de estudo, esse projeto tem como objetivo analisar os métodos de identificação de microRNAs e propor uma abordagem computacional para sua validação utilizando como conjunto de teste, dados de seqüências contidos em Bases Públicas.

1.2 MicroRNAs

Os microRNAs são uma nova classe descoberta de pequenas moléculas de RNA, eles possuem aproximadamente 22 nucleotídeos de tamanho no seu estado maduro e estão envolvidos no processo de regulação da expressão gênica da célula. Essa regulação ocorre através da ligação da molécula de microRNA com o RNA mensageiro interrompendo, assim, sua tradução e impedindo a expressão de genes nos organismos que possuem essa estrutura [4].

O estudo de microRNAs tem se mostrado muito importante, pois eles estão envolvidos em alterações funcionais da célula e muitas vezes em alguns tipos de câncer como, por exemplo, a expressão do *cluster-17-92* que atua em conjunto com o gene *c-myc* para diminuir a velocidade do desenvolvimento de tumores em células B de ratos [5].

Outro estudo mostrou que existem microRNAs capazes de inibir a produção de proteínas, como por exemplo a proteína *E2F1*, que está relacionada com a proliferação celular [3].

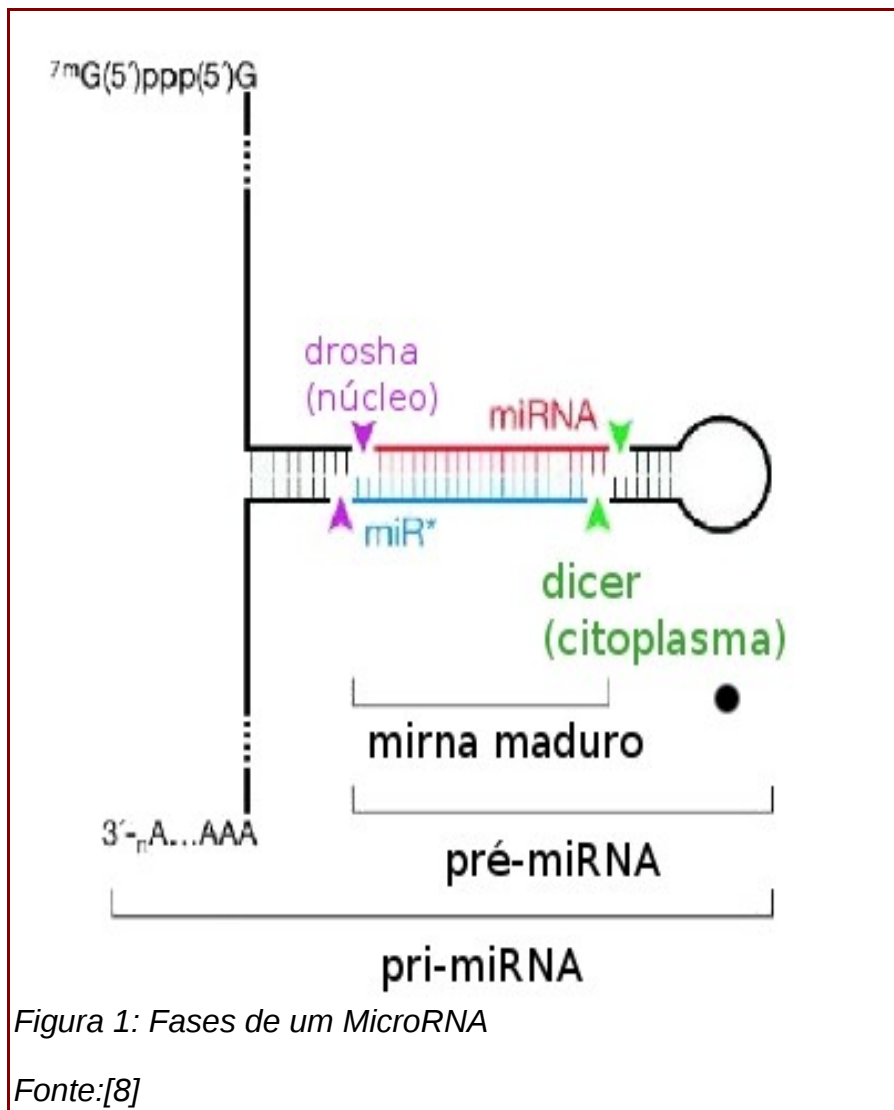


Figura 1: Fases de um MicroRNA

Fonte:[8]

A Figura 1 mostra as diferentes fases de um microRNA. O pri-miRNA é transcrito pela Polimerase II e, após ser processado pela *Drosha* ou *Pasha*, se transforma no pré-miRNA que, por sua vez, sofre um processo de clivagem devido a proteína *Dicer* e finalmente dá origem ao mirna (microRNA) maduro [4].

Apesar das proteínas *Exportin* e *Dicer* serem encontradas em fungos, a proteína responsável pela formação do pré-miRNA neste organismo ainda não foi identificada. Em animais esta proteína é a *Drosha* e em plantas é a *Pasha*.

Os microRNAs são pequenas estruturas codificadas por seqüências que são transcritos do DNA (pri-miRNA), mas que não são traduzidos em proteínas.

Ao invés disso os transcritos primários (pri-miRNA) são processados pela *Drosha* ou *Pasha* dando origem ao pré-miRNA, que possui estruturas curtas de *stem-loops* e são geralmente chamados de *hairpins*.

O pré-miRNA (70-100pb) é processado pela *Dicer* (ou RNAase III) formando o microRNA maduro (21-23pb). Por fim o microRNA maduro atua negativamente na regulação pós transcricional de um determinado gene formando um duplex de RNA entre o RNA mensageiro e o microRNA através da sua complementariedade de bases, interrompendo, assim, a tradução do RNAm em geral pela região 3'UTR em animais [4,7,8].

Então, o microRNA maduro é extraído do pré-microRNA formando sua estrutura final e ativa. Estas estruturas, de acordo com determinadas características, podem ser usadas para identificar novos microRNAs a partir de microRNAs já conhecidos [20].

Diversas pesquisas têm mostrado a importância dessas moléculas no organismo, desde o início do seu desenvolvimento [9], pois uma característica fundamental do seu estudo é a sua conservação, mesmo em organismos geneticamente distantes [10]. O estudo dessas estruturas tem contribuído bastante para compreensão dos processos análogos que são regulados pelos microRNAs na maioria dos organismos vivos existentes [21,22,24].

1.3 Motivação

O estudo de microRNAs é importante pois eles deram origem a uma revolução na compreensão dos processos regulatórios de organismos vivos. Apesar de serem moléculas extremamente pequenas quando comparadas ao tamanho do genoma de um organismo, os microRNAs desempenham um papel fundamental no controle da inativação de determinados genes [11].

Atualmente há 8273 seqüências que correspondem a microRNAs *maduros* e 8619 seqüências de *hairpins* de microRNAs depositadas no *Mirbase* (Release 12.0: Sept 2008) que é um Banco de Dados Mundial de armazenamento de seqüências de microRNAs identificados.

Conforme indicado na Figura 2, o número de novos microRNAs descobertos vem crescendo exponencialmente a cada ano, e espera-se que esse crescimento continue aumentando devido a grande quantidade de organismos que ainda não possuem esse tipo de estrutura identificada e ao importante papel desempenhado pelo microRNA como regulador de importantes processos biológicos [12].

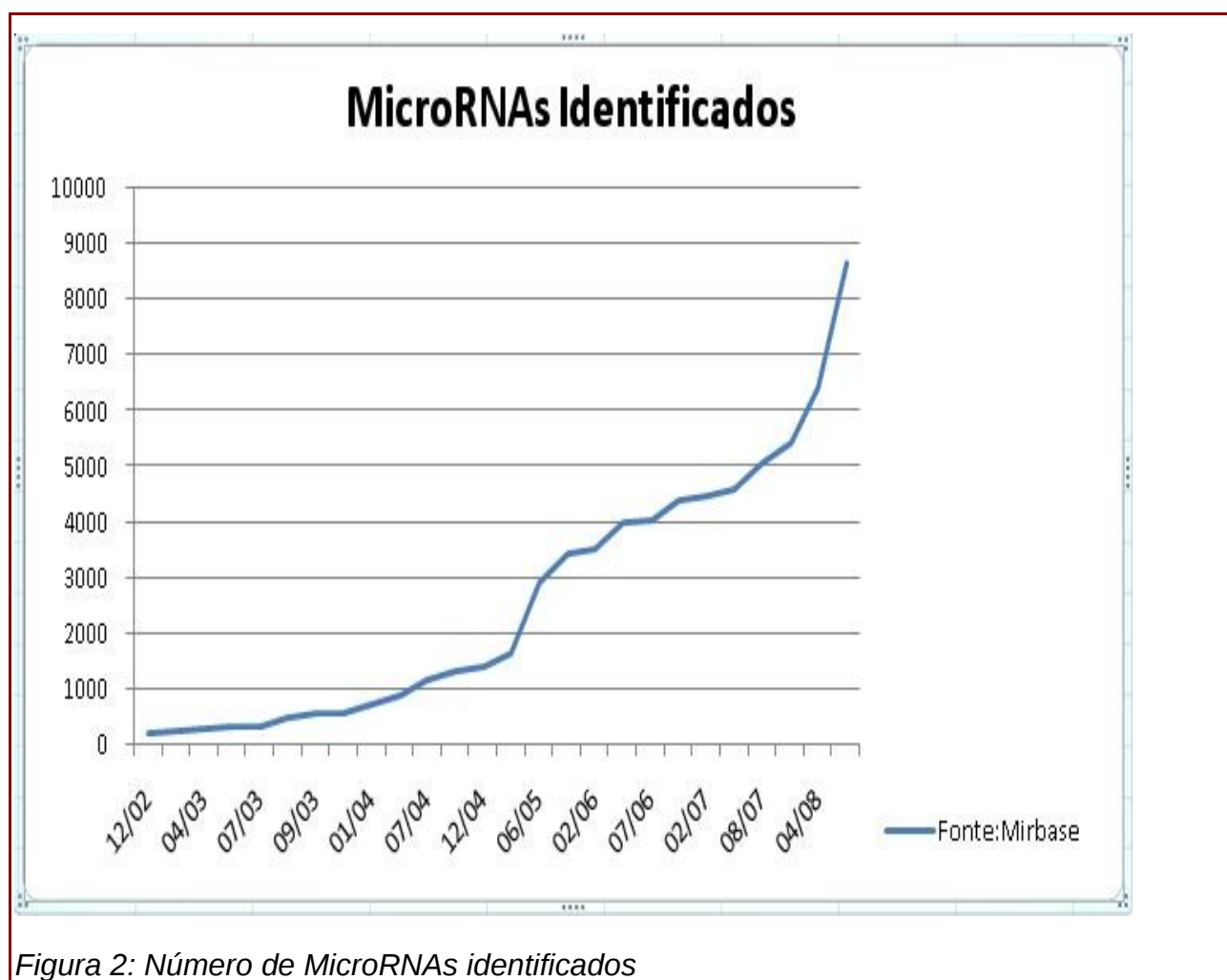


Figura 2: Número de MicroRNAs identificados

Atualmente o número de Genomas completamente seqüenciados é de 3737 e 470 em fase de sequenciamento, totalizando 4207 organismos envolvidos em projetos

genoma atualmente. A situação completa sobre o andamento dos projetos genoma pode ser visualizada na Figura 3 [Fonte: Genome Online Database].

IMG Genomes		
	finished/draft	Total
Bacteria	632/446	1078
Archaea	53/3	56
Eukarya	19/21	40
Plasmids	803/0	803
Viruses	2230/0	2230
All Genomes	3737/470	4207

[IMG Statistics](#)

Figura 3: Número de Genomas Sequenciados

Portanto, acredita-se que ainda exista um grande número de organismos que ainda não possuam microRNAs identificados e que esse número continue diminuindo até que todos os genomas sejam completamente mapeados e decifrados [13].

Para comparação, pode-se citar o número de microRNAs identificados de *Homo sapiens* (695) em relação ao número de microRNAs conhecidos de *Gorilla gorilla* (65) que são dois organismos geneticamente próximos.

Portanto, espera-se que o aparecimento de novas ferramentas de Bioinformática e uma investigação mais profunda nos genomas dos organismos levem à descoberta de novos microRNAs ainda desconhecidos [14,15].

1.4 Objetivos gerais

O Objetivo do Projeto é desenvolver uma ferramenta computacional capaz de

identificar novos microRNAs em organismos que ainda não possuem essa estrutura caracterizada, utilizando para isso microRNAs já identificados em outros organismos.

Através de metodologias que foram aplicadas para plantas e animais será utilizado o valor do *e-value* de um alinhamento, juntamente com o cálculo de energia livre e a predição de estrutura secundária, para identificação de novos candidatos a microRNAs [16,23]. Além disso, será desenvolvida uma ferramenta que possa ser utilizada em diferentes organismos e que será disponibilizada online, com a criação de um Banco de Dados, de forma que pesquisadores possam submeter suas seqüências e identificar novos candidatos à microRNAs no seu organismo de estudo ou fazer análises comparativas.

1.5 Objetivos Específicos

1) Desenvolver uma ferramenta para identificar microRNAs utilizando análises genômicas.

2-) Tentar identificar novos microRNAs a partir de microRNAs já conhecidos.

3-) Extrair os melhores resultados do alinhamento contra as seqüências de microRNAs.

4-) Criar um *pipeline* para a execução dos *scripts* desenvolvidos no projeto.

5-) Predizer a estrutura secundária das seqüências obtidas no alinhamento, através do programa Mirfold, e utilizar a penalidade da predição e o cálculo de energia livre para selecionar as estruturas que serão formadas[17,18].

6-) Representar graficamente os resultados, de forma que seja possível demonstrar a região conservada do microRNA e os *stem-loops* formados.

7-) Automatizar todo o processo desenvolvendo uma ferramenta computacional para ser disponibilizada *online* onde os pesquisadores poderão submeter suas seqüências ao mesmo processo e armazenar os resultados em um Banco de Dados

biológicos.

8-) Testar a ferramenta com dados de Bases Públicas.

2. Material e Métodos

2.1 Obtenção dos dados

Nesse trabalho foram utilizados dados de *Paracoccidioides Brasiliensis* fornecidos pelo Prof. Geraldo Aleixo Passos em colaboração com seu Laboratório, 2 genomas obtidos do Broad Institute (<http://www.broad.mit.edu>), 1 genoma de vírus (Epstein bar) e seqüências de microRNAs identificados e disponibilizadas pelo Mirbase.

Foram utilizadas as seguintes versões dos genomas analisados:

- a) *Paracoccidioides brasiliensis* – Esse arquivo contém 500 seqüências de ESTs de *Paracoccidioides brasiliensis* obtidas diretamente do NCBI (www.ncbi.nlm.nih.gov) que foram depositadas pelo Prof. Geraldo A. Passos.
- b) *Aspergillus nidulans* – Esse arquivo contém 248 seqüências dos supercontigs deste organismo. O sequenciamento do genoma de *Aspergillus nidulans* foi parte do projeto “Fungal Genome Initiative”, desenvolvido pelo Broad Institute. Seu objetivo inicial era disponibilizar uma cobertura de 10X do genoma deste fungo utilizando a linhagem FGSC A4.

Com a ajuda da Monsanto foi possível liberar uma cobertura de 13X do genoma deste fungo, o que resultou na publicação deste conteúdo em março de 2003.

Tabela 1: Dados do Broad Institue

	Size	Chrs	%GC	Genes	tRNAs	rRNA S
<i>A. Nidulans</i>	30.07 Mb	8	50.32	10,701	N/A	N/A

Chrs: Número de cromossomos

%GC: Porcentagem de GC

Genes: Número de genes que codificam proteínas preditos no genoma

tRNAs: Número de genes que codificam tRNAs preditos no genoma

rRNAs: Número de genes que codificam rRNAs preditos no genoma

- c) *Aspergillus fumigatus* – Esse arquivo contém 8 seqüências dos supercontigs que representam os cromossomos deste organismo. O genoma de *Aspergillus fumigatus* (linhagem Af293) foi sequenciado através de um esforço comum entre o TIGR, o Sanger Centre e o Instituto Pasteur, através do financiamento fornecidos pelo Instituto Nacional de Alergia e Doenças Infecciosas (NIAID). Os dados foram liberados em 2005 em um artigo publicado na Nature (Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature 2005, 438:1151-6).

Tabela 2: Dados do Broad Institute

	Size	Chrs	%GC	Genes	tRNAs	rRNAs
<i>A. Fumigatus</i>	29.38 Mb	8	48.82	9,887	N/A	N/A

Chrs: Número de cromossomos

%GC: Porcentagem de GC

Genes: Número de genes que codificam proteínas preditos no genoma

tRNAs: Número de genes que codificam tRNAs preditos no genoma

rRNAs: Número de genes que codificam rRNAs preditos no genoma

O Mirbase contém atualmente 8619 registros representando *hairpins* precursores de microRNAs e 8273 microRNAs maduros em primatas, roedores, pássaros, peixes, minhocas, aves, plantas e vírus (Release 12.0: Sep 2008).

Atualmente o Mirbase possui um total de 34 organismos que contêm microRNAs já identificados e a lista com os nomes dos organismos pode ser encontrada na figura 4:

I

As versões das seqüências de microRNAs utilizadas foram as seguintes :

- a) mature.fasta – 8273 Seqüências de microRNAs maduros em formato FASTA. Release (12.0).
- b) hairpin.fasta – 8619 Seqüências de *hairpins* de microRNAs em formato fasta. Release(12.0).
- c) Epstein.fasta – 23 Seqüências de *hairpins* de microRNAs do vírus Epstein-Barr.

Na figura 5 são apresentados os arquivos FASTA utilizados para gerar os Bancos de Dados de microRNAs para realização do BLAST

2.2 Alinhamento

Para fazer o alinhamento das seqüências foi utilizado o programa BLAST (Basic Local Alignment Search Tool), que procura por regiões de similaridade entre as seqüências. O programa faz a comparação de uma seqüência de nucleotídeos ou de proteínas, com um Banco de Dados e calcula a significância estatística dos resultados. O BLAST também pode ser utilizado para inferir relações evolucionárias entre os organismos e identificar membros de famílias de genes. Fonte: (www.ncbi.nlm.nih.gov/BLAST/)

Entre os parâmetros do alinhamento, o BLAST atribui um valor de *e-value*, e um *Score* para cada alinhamento. Esses dois parâmetros podem ser utilizados para analisar a similaridade e a qualidade do alinhamento obtido.

O *e-value* é um parâmetro que mede a similaridade entre as seqüências, ou seja ele considera a probabilidade do alinhamento ter ocorrido ao acaso e o *Score* de cada pareamento para calcular o quanto uma seqüência se parece com outra. Como são utilizadas para estudo seqüências muito pequenas (~22nt), este valor será na maioria das

vezes alto.

Ele fornece uma idéia sobre a significância estatística de um par de alinhamentos, reflete o tamanho do Banco de Dados e qual foi o sistema de pontuação utilizado.

Quanto menor for o valor do *e-value* mais significativo será o alinhamento. Mesmo que os valores dos cálculos estatísticos sejam significantes, não podemos confiar apenas nisso, uma análise da estrutura secundária deve ser necessária para aumentar o grau de confiança no resultado.

A equação para o cálculo de *e-value* é apresentada na figura 6:

Os valores de K e λ são relacionados aos *Score* de HSPs (*high-scoring segment pairs*) ou seja do *Score* atribuído as regiões completamente pareadas. O K representa um parâmetro de escala para o espaço de procura e o λ um parâmetro de escala para o sistema de pontuação. Os valores de m e n estão relacionados ao tamanho da seqüência e a dimensão da base de dados, respectivamente, e o *Score* está relacionado a soma dos valores atribuídos a cada pareamento de bases do alinhamento.

Ao encontrar uma seqüência semelhante à aquela que foi submetida ao Banco de Dados, é muito importante considerar a qualidade do alinhamento para saber se ele possui alguma relação biológica ou se a semelhança observada ocorreu ao acaso.

O valor do *Score* de um alinhamento é uma indicação da sua qualidade, portanto, quanto maior o valor do *Score*, melhor será o alinhamento. Em termos gerais esse índice é calculado a partir de uma equação que leva em consideração o alinhamento de resíduos similares ou idênticos, além de possíveis *gaps* (espaços não pareados)

introduzidos para que a seqüência fosse alinhada. O elemento chave deste cálculo é o uso da matriz de substituição que atribui uma pontuação para alinhar os eventuais pares de resíduos. O valor do *Score* é normalizado para que possa ser comparado entre diferentes tipos de alinhamento, até mesmo se matrizes de substituição diferentes tiverem sido utilizadas.

A equação para o cálculo do *Score* é apresentada na Figura 7 a seguir:

O *Score* do alinhamento é calculado através da soma dos *Scores* atribuídos a cada pareamento. O *Score* de um pareamento é atribuído de acordo com a identidade entre as seqüências, os *mismatches* (regiões não pareadas) e os *gaps* (regiões de extensão da seqüência) conforme representados na figura 7.

2.5 Predição da Estrutura Secundária

Para fazer a predição da estrutura secundária foi utilizado o software Mirfold para verificar a existência de *hairpins* e com isso validar os resultados obtidos nos

alinhamentos.

O Mirfold é um software que prediz as estruturas secundárias ótimas e possíveis de microRNAs através do cálculo de energia mínima de moléculas e das características específicas de microRNAs [2].

Os critérios para calcular a energia mínima de uma molécula dependem, principalmente, do modelo de dobramento utilizado e das energias atribuídas a cada tipo de dobramento.

Os parâmetros envolvidos na predição da estrutura secundária são o número de nucleotídeos extras “upstream” e “downstream” que o algoritmo considera na hora de atribuir a conformação de cada base, o número de passos que o algoritmo vai utilizar, (esse parâmetro está relacionado com a complexidade do algoritmo) e a temperatura de *fold* e de enovelamento da seqüência [17,18].

Para fazer predição da estrutura secundária o Mirfold utiliza bibliotecas do Vienna Package que é um software bastante utilizado para esse tipo de análise [2].

3. Resultados

Neste capítulo é apresentada a ferramenta que foi construída e os detalhes sobre a sua construção.

3.1 Construção da Ferramenta

Primeiramente as seqüências dos microRNAs foram preparadas para que pudessem ser alinhadas utilizando o BLAST contra as seqüências dos organismos de estudo. Os arquivos formatados para executar o BLAST podem ser visualizados na Figura 8:

Após a preparação dos dados foi desenvolvido um diagrama para representar todas as fases da análise do projeto, de forma que fosse possível criar o *pipeline* dos *scripts* desenvolvidos e executar os programas necessários em cada fase de execução.



Na figura 9 podemos observar que cada fase corresponde a um diferente estado do projeto, onde os dados são submetidos a diferentes *scripts* para processamento de acordo com a sua fase atual.

Para cada fase do projeto foram desenvolvidos *scripts* em Perl para a execução das análises que podem ser visualizados na Figura 10 a seguir:

Os *scripts* apresentados na figura 10 serão explicados detalhadamente nos tópicos a seguir.

3.2 Banco de Dados

Para armazenar os dados gerados pela ferramenta, foi desenvolvido um Banco de Dados em Mysql com tabelas referentes ao projeto, a seqüência, ao alinhamento, ao mapa do alinhamento, a predição da estrutura secundária , além dos microRNAs maduros e *hairpins*.

O Modelo Entidade Relacionamento descreve de maneira conceitual, os dados a serem utilizados para a construção do Sistema. A Figura 11 apresenta os dados armazenados nas tabelas do Banco de Dados.

Figura 4: MER (Modelo Entidade Relacionamento) do Banco de Dados desenvolvido

Sendo que a partir desse Modelo os dados armazenados em cada tabela são os seguintes :

Tabela alignment

idalignment INTEGER – Chave primária atribuída ao alinhamento

sequence_project_idproject INTEGER – Id atribuído ao projeto

sequence_name VARCHAR(255) – Nome da seqüência

sequence_mirna_index INTEGER – Índice do microRNA usado no mapa do alinhamento

mirna_mature_mature VARCHAR(45) – nome do microRNA maduro

algorithm VARCHAR(45) – Algoritmo utilizado no alinhamento

sequence_length INTEGER – Tamanho da seqüência submetida

sequence_description TEXT – Descrição da seqüência submetida

database_name VARCHAR(255) – Nome do Banco de Dados utilizado

database_letters INTEGER – Tamanho do Banco de Dados utilizado

database_entries INTEGER – Número de Sequências contidas no Banco de Dados utilizado

num_hits INTEGER – Número de Alinhamentos da seqüência submetida

mirna_length INTEGER – Tamanho do microRNA alinhado

mirna_description TEXT – Descrição do microRNA alinhado

raw_Score INTEGER – Score atribuído ao alinhamento

significance FLOAT – Grau de significância do alinhamento

num_hsp INTEGER – Número de HSPs de um alinhamento

evaluate FLOAT – Valor de e-value do alinhamento

frac_identical FLOAT – Fração idêntica entre a seqüência e o microRNA

frac_conserved FLOAT – Fração conservada entre a seqüência e o microRNA

gaps INTEGER – Número de Gaps do alinhamento

query_string TEXT – String da seqüência alinhada

hit_string TEXT – String do microRNA alinhado

homology_string TEXT – String de homologia

length_total INTEGER – Tamanho total do alinhamento

length_query INTEGER – Tamanho total da query

num_identical INTEGER – Número de pareamentos idênticos

num_conserved INTEGER – Número de pareamento conservados

rank INTEGER – Rank atribuído ao alinhamento

Score INTEGER – Score atribuído ao alinhamento

range_query – Intervalo da query

range_hit – Intervalo do Hit

percent_identity FLOAT – Porcentagem de identidade

strand_hit INT – Direção do hit

strand_query INT – Direção da query

start_query INTEGER – Início da query

end_query INTEGER – Fim da query

start_hit INTEGER – Início do Hit

end_hit INTEGER – Fim do Hit

mirna_tag VARCHAR(255) – tag atribuída ao microRNA para gerar o mapa do alinhamento

Como chaves primárias para essa tabela foram utilizados os campos: idalignment, sequence_project_idproject, sequence_name, sequence_mirna_index e mirna_mature_mature.

Como chaves estrangeiras para essa tabela foram utilizados os campos: sequence_project_idproject, sequence_name e mirna_mature_mature.

Tabela map_alignment

sequence_name VARCHAR(255) – Nome da seqüência submetida

sequence_project_idproject INTEGER – Chave primária atribuída ao projeto

map TEXT – Código HTML para gerar o mapa do alinhamento

image TEXT – Endereço para a imagem do mapa do alinhamento

Como chaves primárias para essa tabela foram utilizados os campos: sequence_name e sequence_project_idproject.

Como chaves estrangeiras para essa tabela foram utilizados os campos: sequence_name e sequence_project_idproject.

Tabela mirna_hairpin

hairpin VARCHAR(45) – Nome do hairpin do microRNA

mirna_mature_mature VARCHAR(45) – Nome do microRNA maduro

accession VARCHAR(45) – Accession Number para o hairpin de microRNA

status_sequence VARCHAR(45) – Estado atual da seqüência

sequence TEXT – Sequência do microRNA maduro

image_upstream VARCHAR(255) – Imagem Upstream do hairpin do microRNA

image_downstream VARCHAR(255) – Imagem Downstream do hairpin do microRNA

Como chaves primárias para essa tabela foram utilizados os campos: hairpin e mirna_mature_mature.

Como chaves estrangeiras para essa tabela foi utilizado o campo: mirna_mature_mature.

Tabela mirna_mature

mature VARCHAR(45) – Nome do microRNA maduro

accession VARCHAR(45) – Accession Number para o microRNA maduro

sequence VARCHAR(45) – A seqüência do microRNA

specie VARCHAR(255) – a espécie à qual o microRNA pertence

mirna VARCHAR(45) – String que correnpodne a família do microRNA

length_bp VARCHAR(45) – número de pares de base

Como chave primária para essa tabela foi utilizado o campo: mature.

Tabela prediction

idprediction INTEGER – Chave primária atribuída à predição

alignment_sequence_project_idproject INTEGER – Chave primária atribuída ao projeto

alignment_mirna_mature_mature VARCHAR(45) – Chave primária atribuída ao microRNA maduto

alignment_sequence_name VARCHAR(255) – Nome da Sequência do alinhamento

alignment_sequence_mirna_index INTEGER – Índice do microRNA

alignment_idalignment INTEGER – Id do alinhamento

direction VARCHAR(20) – Direção da predição

image TEXT – Endereço da imagem gerada pela predição

sequence TEXT – Sequência da predição

fold TEXT – String correspondente ao Fold da seqüência

penalty INT - Penalidade

prediction_length – Tamanho da predição

free_energy VARCHAR(255) – Cálculo de Energia Livre

mirna_start INTEGER – Início do microRNA

mirna_end INTEGER – Fim do microRNA

Como chaves primárias para essa tabela foram utilizados os campos: idprediction e alignment_idalignment.

Como chaves estrangeiras para essa tabela foram utilizados os campos: alignment_sequence_mirna_index, alignment_sequence_name, alignment_mirna_mature_mature e alignment_sequence_project_idproject.

Tabela project

idproject INTEGER – Chave primária atribuída ao projeto

project_name VARCHAR(45) – Nome do Projeto

email VARCHAR(255) – E-mail do pesquisador

hash VARCHAR(45) – Chave única atribuída ao projeto

ip VARCHAR(45) – IP de origem

projectstatus VARCHAR(45) – Estado atual do projeto

date VARCHAR(45) – Data do Prejeto

sequences INTEGER – Número de Sequências do Projeto

alignment_parameters VARCHAR(255) – Parâmetros do alinhamento

prediction_parameters VARCHAR(255) – Parâmetros da predição

Como chave primária para essa tabela foi utilizado o campo: idproject.

Tabela sequence

name VARCHAR(255) – Nome para a seqüência

project_idproject INTEGER – Id do Projeto

description TEXT - Descrição

sequence LONGTEXT – Sequência submetida

size INTEGER – tamanho da seqüência

Como chaves primária para essa tabela foram utilizados os campos: name e project_idproject

E como chave estrangeira foi utilizado o campo: project_idproject

Para povoar as tabelas mirna_mature e mirna_hairpin foi desenvolvido o script “import.pl” que lê os dados dos arquivos fasta e insere os dados no Banco de Dados.

3.3 Submissão das seqüências

Para que os pesquisadores possam submeter suas seqüências para análise, foi desenvolvida uma interface de submissão de projetos conforme apresentado na tabela 3:

Dados do Projeto:
Nome do Projeto – O nome do projeto é apenas para facilitar a sua identificação
Sequências – As seqüências de entrada devem estar no formato fasta e podem ser inseridas a partir de um arquivo ou de uma caixa de texto
Parâmetros do Alinhamento:
e-value – O e-value serve para restringir a qualidade dos resultados. Quanto menor o valor melhor serão os resultados
WordSize – O Worsize serve para especificar o menor valor que o algoritmo vai considerar para encontrar regiões de similaridade
Filter – O Filtro mascara regiões repetidas para não diminuir a qualidade dos alinhamentos
Parâmetros da Predição da Estrutura Secundária:
Nucleotídeos Extras – Define os nucleotídeos extras em ambas as direções (Upstream e Downstream) que serão utilizados na hora de predizer a estrutura secundária
Passos – Define a quantidade de passos que o algoritmo vai utilizar, esse parâmetro está relacionado com o custo computacional do algoritmo
Temperatura – Define a temperatura ideal para a expressão dos microRNAs no organismo

Tabela 3: Tabela de Submissão

Com isso, as interfaces desenvolvidas em PHP são apresentadas nas Figuras 12,13 e 14:

Na figura 12 pode-se observar os campos referentes ao nome do projeto e para submissão das sequências em formato texto ou através do envio de um arquivo.

Na figura 13 são apresentados os parâmetros para o alinhamento entre as seqüências de entrada e as seqüências de microRNAs maduros



Na figura 14 são apresentados os parâmetros da predição da estrutura secundária das seqüências que alinharem com os microRNAs maduros.

Após o usuário submeter um novo projeto para execução, o sistema verifica se houve o preenchimento de todos os campos necessários para a execução dos *scripts*, e caso os dados estejam corretos, submete as seqüências para a validação.

Caso os campos não estejam corretamente preenchidos, o usuário recebe um aviso sobre as correções necessárias, juntamente com o formulário de submissão.

3.4 Verificação dos Dados

Para fazer a validação dos dados foi desenvolvido um *script* em Perl “validation.pl” utilizando bibliotecas de Bioperl (<http://www.bioperl.org>) para verificar se o arquivo de entrada está no formato FASTA, ou seja, se possui um nome para cada seqüência e se a seqüência contém apenas as letras A,C,T,G e U.

Caso o arquivo passe pelas especificações, é criada uma pasta para a execução do projeto, cujo nome referente é o “Id” do Projeto. Então, o arquivo contendo as seqüências submetidas é copiado para a pasta de projetos para o início da execução das análises.

Após essa fase de preparação das sequências, os dados referentes ao Projeto são

inseridos no Banco de Dados e seu status é definido como “New”. Um *script* desenvolvido para fazer a execução dos programas e outros *scripts* que serão executados durante a análise é executado (“analysis.pl”). Ao iniciar sua execução o *script* executa outro *script* (“insert_sequences.pl”) para fazer a inserção das seqüências do arquivo fasta no Banco de Dados.

O *script* “insert_sequences.pl” utiliza bibliotecas de Bioperl para ler o arquivo fasta e inserir os dados na tabela “sequence” do Banco de Dados.

Após a inserção das seqüências o *script* “analysis.pl” altera o *status* do projeto para “Aligning”.

3.5 Alinhamento

Através do *script* “analysis.pl” é feita uma chamada para a execução do BLAST utilizando os parâmetros definidos pelo usuário.

Para validação dos resultados do projeto, o alinhamento foi feito utilizando o genoma de 3 tipos de fungos (Paracoccidioides brasiliensis, Aspergillus nidulans e Aspergillus fumigatus) e 1 vírus (Epstein-barr).

Lembrando que os fungos ainda não possuem nenhum tipo de microRNA identificado e o Epstein-barr possui 23 microRNAs já identificados e depositados no Mirbase

O objetivo dessa abordagem foi selecionar os melhores resultados dos alinhamentos e identificar as seqüências que fossem encontradas com a predição da estrutura secundária.

Portanto, obteve-se 4 alinhamentos, um pra cada genoma analisado, todos contra o banco de dados de microRNAs maduros:

Tabela 4: Alinhamentos Realizados

QUERY X DATA BANK
PB.fasta x mature
nidulans.fasta X mature
fumigatus.fasta X mature
Epstein.fasta X mature

Os alinhamentos foram feitos utilizando os seguintes parâmetros:

Exemplo: `blastall -p blastn -d ../db/mature -i ../sequencias/PB.fasta -o blastn_PBXhairpin -`

`F f -w7 -e 0.01`

-p - Tipo do algoritmo de alinhamento.

-d - Banco de dados utilizado nos alinhamentos

-i - Seqüências de entrada para o alinhamentos

-o - Nome do arquivo de saída do alinhamento

-F T - Utiliza o filtro ligado para obter os melhores resultados possíveis.

-w7 - utiliza um *wordsize* de 7 ao invés de 11 que é o padrão definido pelo BLAST. O *wordsize* é o menor tamanho de uma seqüência utilizado para a busca feita pelo BLAST.

Quanto menor for o valor do *wordsize* mais sensível será o programa, porém isso irá aumentar bastante o tempo de processamento da análise.

Na figura 15 são apresentados os arquivos de saída do BLAST gerados pelos scripts desenvolvidos.

Após o término de execução do BLAST, o *script* “analysis.pl” executa um *script* chamado “html4BLAST.pl” que gera um arquivo HTML contendo o mapa dos alinhamentos do BLAST para cada seqüência.

Então, o *script* “analysis.pl” executa o *script* “parseblast.pl” para fazer o processamento dos arquivos de saída do BLAST.

Para extração dos dados dos resultados dos alinhamentos do BLAST foi desenvolvido um *script* em Perl utilizando bibliotecas de BioPerl [19] que percorrem o arquivo de saída do BLAST e retiram as informações mais importantes de todos os alinhamentos gerando um arquivo do tipo CSV para que o usuário possa fazer o download da tabela contendo os alinhamentos. Os seguintes campos foram extraídos :

algorithm, query_name, query_accession, query_length, database_name, database_letters, database_entries, num_hits, name, length, accession, description, raw_Score, significance, bits, num_hsp, locus, evalue, expect, frac_identical, frac_conserved, gaps, query_string, hit_string, homology_string leng, th('total') length('hit'), length('query'), hsp_length, query->frame, hsp->frame, num_conserved, num_identical, rank, Score, bits, range('query'), range('hit'), percent_identity strand('hit'), strand('query'), start('query'), end('query'), start('hit'), end('hit')

Além de gerar um arquivo do tipo CSV, os dados são todos inseridos no Banco de Dados para que os alinhamentos possam ser visualizados junto com as predições.

Após inserir os dados dos resultados do BLAST, o *script* faz um *parsing* no arquivo de saída do script html4blast para extrair os mapas dos alinhamentos e faz a inserção de cada mapa no Banco de Dados juntamente com o nome da seqüência e o Id do projeto.

Após a finalização do *script* parseblast.pl, a execução novamente retorna para o *script* “analysis.pl” que atualiza o *status* do projeto para “Predicting” e executa o *script* “prediction.pl” para fazer a predição da estrutura secundária das regiões que alinharam.

3.6 Extração das Regiões Conservadas

Para fazer a predição da estrutura secundária foi extraído 600 pb *upstream* e *downstream* da região onde ocorreu o alinhamento com a seqüência de microRNA maduro. E com isso foi gerado um arquivo FASTA dessa seqüência contendo no cabeçalho as coordenadas da região onde houve um alinhamento com o microRNA maduro de outro organismo.

3.7 Predição da Estrutura Secundária

Continuando a execução do *script* "prediction.pl", é feita uma chamada para a execução do programa Mirfold utilizando os parâmetros definidos na submissão do projeto e o arquivo FASTA gerado anteriormente como parâmetros de entrada.

O arquivo de saída do Mirfold contém uma *string* que representa a estrutura secundária predita nas duas direções (*Upstream* e *Downstream*), de acordo com os parâmetros de entrada escolhidos e respeitando as características relacionadas aos microRNAs.

3.8 Representação Gráfica

Para a representação gráfica da estrutura secundária obtida, o *script* executa o programa RNAPLOT e passa como parâmetros o arquivo de saída do Mirfold e as coordenadas da região do microRNA maduro para que seja possível visualizar exatamente onde está essa região que alinhou.

O programa RNAPLOT devolve uma imagem no formato *PostScript* como saída que por sua vez é convertida para o formato JPG para que possa ser visualizada ou então para que seja feito o *download* do arquivo pelo pesquisador.

Após a geração das imagens, o *script* prediction.pl insere os resultados da predição no Banco de Dados para que possam ser visualizados posteriormente.

Quando o *script* “prediction.pl” é finalizado o *status* do projeto é atualizado para “Done” e o processamento dos dados chega ao fim.

3.9 Visualização dos Resultados

Para a visualização dos resultados foi desenvolvida uma interface em PHP que permite que o pesquisador tenha acesso aos resultados utilizando qualquer navegador *web*. A interface é apresentada na Figura 16 :

Ao clicar em Projetos o pesquisador é redirecionado para uma página que exhibe todos os projetos que foram submetidos a partir daquele endereço de IP e um campo para que ele possa inserir um “hash” atribuído ao seu Projeto, caso ele não esteja sendo exibido na lista logo abaixo.



Figura 5: Área de Projetos

Na figura 16 é apresentada a interface desenvolvida para a ferramenta e a localização da área de projetos no site.

Na visualização dos projetos do usuário (Figura 17), cada projeto possui uma tabela que mostra os dados referentes aos seguintes campos:

Tabela 5: Dados referentes ao Projeto

Dados do Projeto:
Nome do Projeto – Nome dado durante a submissão
Hash – Chave única atribuída ao projeto
IP – Endereço de origem de onde foram submetidos os dados
Status – Estado atual do Projeto
Data – Data em que o processo foi submetido
Parâmetros do alinhamento – Utilizados durante a execução do BLAST
Parâmetros da predição – Utilizados durante a execução do Mirfold

Quando o alinhamento das seqüências do projeto estiver finalizado, será exibido um *link* logo abaixo da tabela (“View Alignments”) para que os resultados do alinhamento possam ser visualizados.

Na figura 18 podemos observar que os alinhamentos são exibidos em uma tabela com todos os microRNAs que tiveram um alinhamento com a seqüência submetida, o *Score* e o *e-value* do alinhamento.

Logo abaixo à tabela que contém a lista dos alinhamentos, é possível observar um mapa para o alinhamento que mostra a região da seqüências onde os microRNAs alinharam. Ao clicar nessa região, correspondente ao alinhamento com o microRNA, o usuário é direcionado para a tabela contendo as informações específicas sobre o alinhamento.

A figura 19 mostra um exemplo de visualização dos dados específicos de um alinhamento entre a sequência submetida e o microRNA alinhado.

Na tabela de visualização do alinhamento é possível obter informações mais específicas sobre o alinhamento. Os dados são apresentados na tabela a seguir:

Tabela 6: Tabela com os dados sobre o Alinhamento

Tabela de Alinhamento
Query Name – O nome da seqüência submetida
Query Length – O tamanho da seqüência submetida
Query Description – A descrição da seqüência submetida
Hit Name – O nome do microRNA que alinhou com a seqüência
Hit Length – O tamanho do microRNA que alinhou com a seqüência
Hit Description – A descrição do microRNA que alinhou com a seqüência
Algorithm – O algoritmo utilizado para o alinhamento
Database – O Banco de Dados utilizados para o alinhamento
Percent Indentity – A similaridade entre as seqüências que alinharam
Lenght total – O tamanho total do alinhamento
e-value – Valor atribuído ao alinhamento
Score – Valor atribuído ao alinhamento
Homology Map – O mapa de homologia do alinhamento
Prediction – A predição de estrutura secundária do alinhamento (se houver)
Predicted Targets – Os alvos preditos relacionados ao microRNA que alinhou

Ao clicar em “View” o usuário é redirecionado para a página de visualização da estrutura secundária da seqüência alinhada, para verificar a existência de estruturas características de microRNAs.

Na figura 20 é apresentada a visualização dos dados referentes a predição da estrutura secundária além da figura que representa a predição.

A pagina de visualização da predição da estrutura secundária mostra os resultados da seguinte maneira:

Tabela 7: Tabela com os dados sobre a Predição da Estrutura Secundária

Tabela de Predição da Estrutura Secundária
Sequence – Nome da seqüência submetida
Mirna – Nome do microRNA que alinhou com a seqüência submetida
Direction – Direção do Fold (upstream ou downstream)
Sequence – Seqüência predita pelo Mirfold
Fold – String que representa o fold
Penalty – Penalidade atribuída ao fold
Prediction Length – Tamanho da seqüência predita
Free Energy – Calculo de Energia Livre da seqüência
Mirna Start – Coordenada de inicio do microRNA
Mirna End – Coordenada do final do microRNA
Image – Imagem correspondente ao microRNA

3.10 Análise dos Resultados

Para validação da ferramenta foram submetidos 4 genomas diferentes (3 Fungos e 1 Vírus) para análise dos resultados obtidos.

O primeiro genoma submetido foi o genoma de *Epstein-barr*. Os resultados dos alinhamentos mostraram 31 alinhamentos com seqüências do próprio *Epstein-barr* nas regiões 5´ e 3´ e 5 alinhamentos com seqüências de microRNAs de *Rhesus lymphocryptovirus*, que é um tipo de vírus do mesmo gênero que o *Epstein*.

Isso de certa forma faz sentido, pois por pertencerem ao mesmo gênero esses organismos devem possuir um alto grau de conservação entre os seus microRNAs.



A figura 21 mostra um resultado de alinhamento entre uma sequência de Epstein-Barr e 4 sequências de microRNAs.

Essa análise foi importante para verificar se a ferramenta estava encontrando os microRNAs que já estão identificados em um genoma que possui essas estruturas conhecidas.

Após essa análise inicial, foi utilizado o genoma de 3 fungos (*Paracoccidioides brasiliensis*, *Aspergillus nidulans* e *Aspergillus fumigatus*) para tentar encontrar novos candidatos à microRNAs.

Os resultados para o genoma de *Paracoccidioides brasiliensis* mostraram um total de 225 alinhamentos com microRNAs já conhecidos. Entre os alinhamentos pode-se

observar:

Tabela 8: Número de Alinhamentos obtidos

Alinhamentos de Paracoccidioides Brasiliensis
34 alinhamentos com Pinus taeda
28 alinhamentos com Physcomitrella patens
22 alinhamentos com Arabidopsis thaliana
11 alinhamentos com Homo sapiens

Esse resultados podem ser utilizados para validação de microRNAs em Paracoccidioides brasiliensis a partir dos microRNAs que alinharam. Ou então para inferir uma árvore evolucionária entre os organismos relacionados.

Outro resultado interessante é que alguma regiões alinharam com vários microRNAs diferentes, como por exemplo:



A figura 22 mostra o mapa de um alinhamento entre uma sequência de *Paracoccidioides Brasiliensis* e alguns microRNAs de outros organismos.

Isso pode ser utilizado para verificar a similaridade entre os microRNAs que já são

conhecidos e que possuem uma certa semelhança entre eles próprios.

Além disso a Figura 23 mostra como exemplo, o resultado do alinhamento com um microRNA conhecido de *Oikopleura dioica*.

Os dados específicos do alinhamento são apresentados na Figura 24:

E a predição de sua estrutura secundária pode ser visualizada na figura 25:

A análise feita com o Genoma de *Aspergillus nidulans* retornou 12 alinhamentos e, entre os resultados pode-se citar o alinhamento e a predição da estrutura secundária de uma seqüência de *Aspergillus* com um microRNA de *Vitis vinifera* apresentado na Figura 26:

Esse alinhamento deu origem a predição que é apresentada na figura 27 a seguir:



A estrutura formada pela predição é apresentada na Figura 28 a seguir:

Para o alinhamento de *Aspergillus fumigatus* não houve nenhum resultado considerável. Isso pode ter ocorrido devido a escolha dos parâmetros utilizados sendo portanto necessário um estudo mais aprofundado que não é o principal objetivo desse trabalho.

4. Conclusão

Os resultados encontrados utilizando o BLAST e o Mirfold foram considerados satisfatórios, sendo possível, portanto, identificar a formação de alguns *stem-loops* que são características fundamentais das estruturas dos microRNAs além de obter bons alinhamento com os microRNAs já existentes.

Esse estudo permitiu que fossem selecionados bons candidatos para uma possível validação da existência de microRNAs em *Aspergillus nidulans* e *Paracoidioides brasiliensis*.

A partir das seqüências selecionadas é possível utilizar novas abordagens computacionais para encontrar possíveis alvos de microRNAs no genoma dos fungos estudados ou então fazer a identificação de microRNAs em outros tipos de organismos.

Portanto, a ferramenta se mostrou muito eficiente na identificação de candidatos a microRNAs e pode ser utilizada para a análise de genomas em qualquer tipo de organismo ou então para que seja feita uma pré-seleção dos dados que serão validados.

Para uma validação dos resultados obtidos será necessária a realização de testes de expressão de microRNAs como por exemplo (Northern-blot, RTPCR, microarrays), para que a existência dessas estruturas seja realmente comprovada no organismo.

5. Trabalhos Futuros

Para uma classificação dos dados obtidos pode ser utilizada uma rede neural artificial que é um mecanismo de classificação baseado no aprendizado.

Uma rede neural artificial é um conceito bastante utilizado em computação que visa trabalhar com o processamento dos dados de maneira semelhante ao cérebro humano.

Ou seja, é um sistema capaz de tomar decisões baseadas no aprendizado (ou experiência) e disponibilizar esse conhecimento para classificar novos dados.

Como conjunto de teste para o aprendizado da rede, pode-se utilizar os microRNAs que já foram encontrados em alguns organismos. E como parâmetros de entrada os dados relacionados ao alinhamento (*e-value* e *Score*) e à predição de estrutura secundária (penalidade e o cálculo de energia livre).

Existem algumas características específicas de microRNAs em plantas e animais que podem ser consideradas para construir filtros durante o processamento com o objetivo de aproximar os resultados obtidos dos resultados esperados.

Um outro problema que pode ser abordado é a utilização de *clusters* de computadores para melhorar a performance das análises e diminuir o tempo de processamento.

6. BIBLIOGRAFIA

1. LIU G, WONG-STAL F, LI QX. Development of new RNAi therapeutics. *Histol Histopathol.* 2007;22:211-217.
2. B. BILLOUD, R. DE PAEPE, D. BAULCOMBE AND M. BOCCARA, "Identification of new small non-coding RNAs from tobacco and Arabidopsis" *Biochimie* 2005, 87 (9-10):905-910.
3. LU J, GETZ G, MISKA EA, ALVAREZ-SAAVEDRA E, LAMB J, PECK D, SWEET-CORDERO A, EBERT BL, MAK RH, FERRANDO AA, DOWNING JR, JACKS T, HORVITZ HR, GOLUB TR (2005). "MicroRNA expression profiles classify human cancers". *Nature* 435 (7043): 834–838. doi:10.1038/nature03702. PMID 15944708.
4. LEE RC, FEINBAUM RL, AMBROS V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75: 843-854, Dec 3 1993
5. HE L. A microRNA polycistron as a potential human oncogene. *Cell.* 2005 Jul 15;122(1):6-7.
7. RUVKUN G. . *Molecular biology. Glimpses of a tiny RNA world.* *Science* 294 (5543): 797-9, Oct 26 2001
8. CERVATO M. C. Análise das Associações de microRNAs e seus Alvos em Bibliotecas de SAGE Dec 2007 – Trabalho de Conclusão de Curso de Informática Biomédica
9. ROMERO DG, PLONCZYNSKI MW, CARVAJAL CA, GOMEZ-SANCHEZ EP, GOMEZ-SANCHEZ CE. "Microribonucleic acid-21 increases aldosterone secretion and proliferation in H295R human adrenocortical cells." *Endocrinology.* 2008 May;149(5):2477-83. Epub 2008 Jan 24. PMID: 18218696 [PubMed - indexed for MEDLINE]
10. AMBROS, V. The functions of animal microRNAs. *Nature*, v. 431, n. 7006, p. 350–355, Sep 2004.
11. WEILER J., HUNZIKER J., HALL J. Anti-miRNA oligonucleotides (AMOs): ammunition

to target miRNAs implicated in human disease ? *Gene Therapy* (2006) 13, 496–502., Sep 29 2005

12.RUSINOV V. ET AL *MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence* *Nucleic Acids Research* 33(Web Server Issue):W696-W700, 2005

13.WANG X. ET AL *MicroRNA identification based on sequence and structure alignment* *Bioinformatics* 21(18):3610-3614, 2005

14.RAJEWSKY , N *Computational identification of microRNA targets* *Genome Biology*, 5:P5, 2004

15.LAI E. C., TOMANCAK. P., WILLIAMS R. W , RUBIN G. M. *Computational identification of Drosophila microRNA genes* *Rubin*, Jul 30 2003

16.LEGENDRE M. , LAMBERT A., GAUTHERET D., *Profile-based detection of microRNA precursors in animal genomes* *Bioinformatics* 21(7):841-845, 2005

17.ZUKER M. *Mfold web server for nucleic acid folding and hybridization prediction*, *Nucleic Acids Research*, Vol. 31, No. 13 3406-3415, Zuker

18. ZUKER, M. D. H. MATHEWS & D. H. TURNER. *Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide*; In *RNA Biochemistry and Biotechnology*, 11-43

19.STAJICH J. E., ET AL *The Bioperl Toolkit: Perl Modules for the Life Sciences* *Genome Res.* ; 12(10): 1611–1618, 2002

20.CHIROMATZO A.O., CARDENAS R. G. C. C. L. ET AL *MiRNApath: a database of miRNAs, target genes and metabolic pathways* *Genet. Mol. Res.* 6 (4): 859-865, 2007

21.ENRIGHT A.J., JOHN B., GAUL U., TUSCHL T., SANDER C., *MicroRNA targets in Drosophila* *Marks*; ; *Genome Biology* 5(1):R1. 2003

22.LEE P. LIM ET AL *Vertebrate MicroRNA Genes* *Science*: Vol. 299. no. 5612, p. 1540, March 7 2003

23. ANDREAS R. GRUBER, RONNY LORENZ, STEPHAN H. BERNHART, RICHARD NEUBÜCK, AND IVO L. HOFACKER "The Vienna RNA Websuite" *Nucleic Acids Res.* 2008 July 1; 36(Web Server issue): W70–W74.
24. HE L. A microRNA polycistron as a potential human oncogene. *Cell.* 2005 Jul 15;122(1):6-7.