

Universidade de São Paulo
Faculdade de Medicina de Ribeirão Preto
Faculdade de Filosofia Ciências e Letras de Ribeirão Preto
Informática Biomédica

**Desenvolvimento de sistema *web* para
visualização dos dados da análise *In Silico*
de biblioteca de cDNA de glândula de
peçonha da aranha *Parawixia bistriata***

Camila Santana Justo Cintra Sampaio

Ribeirão Preto – SP

2008

Universidade de São Paulo
Faculdade de Medicina de Ribeirão Preto
Faculdade de Filosofia Ciências e Letras de Ribeirão Preto
Informática Biomédica

**Desenvolvimento de sistema *web* para
visualização dos dados da análise *In Silico*
de biblioteca de cDNA de glândula de
peçonha da aranha *Parawixia bistriata***

Monografia apresentada como parte
dos requisitos para obtenção do título
de bacharel em Informática Biomédica

Camila Santana Justo Cintra Sampaio

Orientadora: Prof^a. Dr^a. Silvana Giuliatti

Co-orientadora: Prof^a. Dr^a. Alessandra Alaniz Macedo

Ribeirão Preto - SP

2008

“Aprender é a única coisa de que a mente nunca se cansa, nunca tem medo, e nunca se arrepende.”

Leonardo da Vinci

Dedico

**A minha mãe, pelo amor,
dedicação e incentivo. Sempre!**

Agradecimentos

Aos meus pais, **Marilza e Paulo Sérgio**, pelo apoio, incentivo e ajuda constante, sempre me incentivando a voar mais alto. Em especial a minha mãe, que sempre esteve ao meu lado, incentivando terminar o curso.

A minha querida irmã **Carolina** que sempre acreditou em mim, passando força e carinho.

A minha madrinha **Marinês**, pelo constante incentivo acadêmico.

A minha avó **Aparecida**, por seu abraço apertado e as incontáveis “marmitinhas” preparadas.

A toda **família Santana Justo** pela atenção, incentivo e ajuda nos estudos, me apoiando quando fui para Uberlândia na busca por uma preparação para o ingresso na Universidade pública.

Aos meus **amigos** de faculdade, pelas longas noites de estudos em grupo e alguns momentos de descontração.

Aos meus **amigos** de Franca, pelo carinho e presença constante, mesmo que *online*, e pelos importantes momentos de descontração.

A minha orientadora **Profª Drª Silvana Giuliatti**, pela confiança em meu trabalho e pelos valiosos ensinamentos.

A minha co-orientadora **Profª Drª Alessandra Alaniz Macedo**.

Enfim, a todos que direta ou indiretamente contribuíram para a realização deste trabalho, **muito obrigada!**

RESUMO

Os projetos científicos, que tem como objetivo a análise de sequências de cDNA, RNA ou sequências protéicas, fazem uso de ferramentas de bioinformática para o processamento, análise, armazenamento e visualização de seus dados e resultados obtidos. Tais projetos, geralmente, envolvem pesquisadores de diferentes áreas e de diferentes localidades. Essa problemática está presente no projeto da Universidade de Ribeirão Preto (UNAERP) intitulado “Construção de uma biblioteca de cDNA de glândula de peçonha da aranha *Parawixia bistriata*” financiado pela FAPESP, sob coordenação da Prof^a Dr^a Sônia M. Zingaretti. No projeto, os cromatogramas gerados como resultado do sequenciamento de cDNA devem ser processados para análise da qualidade das bases e remoção de possíveis contaminantes (vetores e/ou adaptadores). Após esse processamento, deve ser realizado o arranjo das sequências para ser analisadas e armazenadas, para tanto, as ferramentas de bioinformática devem ser empregadas. Portanto, o presente projeto tem como objetivo a aplicação de ferramentas de bioinformática para o arranjo e alinhamento de sequências de cDNA da aranha *Parawixia bistriata*, assim como o desenvolvimento de um sistema *web* que permitiu a visualização dos dados resultantes do arranjo e alinhamento pelos pesquisadores colaboradores do projeto. O arranjo das sequências foi realizado fazendo uso das ferramentas PHRED/CROSSMATCH/CAP3. Os alinhamentos locais foram realizados através das ferramentas BLASTX e BLASTN. O sistema *web* foi desenvolvido e encontra-se disponível no endereço: <http://bioinfo1.fmrp.usp.br/~ccintra/>

Palavras chave: bioinformática, peçonha, cDNA;

CAPÍTULO 1: INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO	01
1.2. OBJETIVO	02
1.3. ORGANIZAÇÃO DA MONOGRAFIA	03

CAPÍTULO 2: METODOLOGIA

2.1. CONSIDERAÇÕES INICIAIS	04
2.2. LEVANTAMENTO DOS DADOS	05
2.3. DEFINIÇÃO DO ARRANJO DAS SEQUÊNCIAS	06
2.4. DEFINIÇÃO E USO DE FERRAMENTAS DE BIOINFORMÁTICA ..	08
2.4.1. PHRED	08
2.4. 2. CROSSMATCH	13
2.4. 3. CAP3	14
2.4. 4. BLAST	20
2.5. DESENVOLVIMENTO DO SISTEMA <i>WEB</i>	23
2.6. CONSIDERAÇÕES FINAIS	24

CAPÍTULO 3: RESULTADOS

3.1. CONSIDERAÇÕES INICIAIS	25
3.2. RESULTADOS CAP3	26

3.3. SISTEMA <i>WEB</i>	27
3.5. CONSIDERAÇÕES FINAIS	34

CAPÍTULO 4: CONCLUSÕES

REFERÊNCIAS BIBLIOGRÁFICAS	36
----------------------------------	----

Lista de Figuras

FIGURA 2. 1. <i>Pipeline</i> utilizado no arranjo e alinhamento das sequências	07
FIGURA 2. 2. Um das sintaxes possíveis de execução do PHERD	19
FIGURA 2. 3. Arquivo de saída PHRED	20
FIGURA 2. 4. Um das sintaxes possíveis de execução do FASTA	21
FIGURA 2. 5. Arquivo de saída FASTA	22
FIGURA 2. 6. Um das sintaxes possíveis de execução do CROSSMATCH	23
FIGURA 2. 7. Arquivo de saída CROSSMATCH	24
FIGURA 2. 8. Arquivo de saída CAP3 – Contigs (reads)	26
FIGURA 2. 9. Arquivo de saída CAP3 – Contigs (sequências)	27
FIGURA 2. 10. Arquivo de saída CAP3 – Singlets	28
FIGURA 2. 11. Arquivo de saída BLASTX	31
FIGURA 2. 12. Arquivo de saída BLASTN	32
FIGURA 3. 1. Estatísticas CAP3	37
FIGURA 3. 2. Pagina inicial do site	39
FIGURA 3. 3. Resumo disponível no site	40
FIGURA 3. 4. Programas utilizados	41
FIGURA 3. 5. Resultados	42
FIGURA 3. 6. Visualização dos resultados	43
FIGURA 3. 7. Grupo	44

Lista de Siglas

ESTs: Expression Sequence Tags

DNA: Deoxyribonucleic Acid (Ácido Desoxiribonucleico)

cDNA: Ácido Desoxiribonucleico complementar

RNA: Ribonucleic Acid (Ácido Ribonucleico)

PCR: Diferencial Display Reverse Transcriptase

Web: Word Wide Web

HTML: Hypertext Markup Language

Perl: Practical Extraction and Report Language

NCBI: National Center for Biotechnology Information

BLAST: Basic Local Alignment Sequence Tool

A: adenine

C: citosina

G: guanine

T: timina

1. 1. CONTEXTUALIZAÇÃO

Nos últimos anos, a comunidade científica vem concentrando esforços em pesquisas na área de genômica com o intuito de se verificar a sequência de DNA e a identificação das regiões codificadoras de genes em genomas nos mais variados organismos. Isso tem possibilitado avanços significativos no entendimento filogenético de espécies, de funções metabólicas, até então desconhecidas, bem como a possibilidade de um avanço biotecnológico pronunciado nas mais variadas áreas de ciências médicas e biológicas.

O projeto da UNAERP, “Construção de uma biblioteca de cDNA de glândula de peçonha da aranha *Parawixia bistriata*”, sob a coordenação da Prof^a Dr^a Sônia M. Zingaretti e, financiado pela FAPESP, objetiva selecionar genes que codifiquem possíveis peptídeos com função biológica a partir de tecido da glândula da peçonha de *P. bistriata*, bem como a criação de um banco de dados de cDNA que possa fornecer informações biológicas importantes do tecido da glândula da peçonha da *P. bistriata*. Para a identificação e a caracterização funcional de ESTs, fez-se necessário o uso de ferramentas de Bioinformática.

O processo de sequenciamento, inicia-se pela extração das glândulas da peçonha da *P. Bistriata*, seguido pela extração do RNA total e isolamento do mRNA,

obtendo, assim, uma biblioteca com os DNA complementares (cDNA), que foram inseridos o sequenciador, resultando nos cromatogramas. Após este processo, há necessidade de fazer o arranjo dessas sequências. Ferramentas de bioinformática auxiliam nesse processo, analisando a qualidade das bases e realizando o “assembly” (arranjo) das sequências. Após a obtenção das sequências “contigs” e “singlets”, os resultados do arranjo de sequências, alinhamentos globais devem ser realizados na busca por similaridades com sequências já depositadas no banco de dados público. A partir desta etapa, dá-se início á análise dos resultados na busca pela identificação de fins e funções biológicas.

1. 2 OBJETIVO

O presente projeto tem como objetivo a aplicação de ferramentas de bioinformática para o arranjo e alinhamento de sequências de cDNA da aranha *Parawixia bistriata*, assim como o desenvolvimento de um sistema *web* que permitiu a visualização dos dados resultantes do arranjo e alinhamento das sequências pelos pesquisadores colaboradores do projeto.

1. 3 ORGANIZAÇÃO DA MONOGRAFIA

As realizações desse trabalho estão descritas nos próximos capítulos desse documento da seguinte forma: o Capítulo 2 apresenta a metodologia utilizada no desenvolvimento deste projeto, citando as ferramentas utilizadas, bem como as linguagens para o desenvolvimento do sistema *web*; o Capítulo 3 apresenta os resultados obtidos bem como o sistema *web*. O Capítulo 4 destina-se a explorar as conclusões do trabalho desenvolvido.

Capítulo 2 - Metodologia

2. 1. CONSIDERAÇÕES INICIAIS

Neste capítulo será apresentada a metodologia utilizada para o arranjo de sequências obtidas da glândula de peçonha da aranha *Parawixia bistriata*, utilizando as ferramentas PHRED (EWIN & GREEN, 1998), CROSSMATCH (EWIN & GREEN, 1998) e CAP3 (HUANG, & MADAN, 1999), assim como a metodologia utilizada na busca por similaridade em bancos de dados públicos, utilizando a ferramenta Blast (ALTSCHUL ET AL, 1997), para o alinhamento das sequências. São softwares livres disponibilizados, mediante controle, para fins acadêmicos.

Todas as sequências e resultados foram armazenados no servidor local da Faculdade de Medicina de Ribeirão Preto, disponibilizado pelo Laboratório de Bioinformática, sob coordenação da Prof^a. Dr^a. Silvana Giuliatti, do Departamento de Genética, e disponibilizados para visualização no sistema *Web (Word Wide Web)*.

Para os *scripts*, a linguagem selecionada foi a Perl (WALL, L. 1987), além de ser executada praticamente em toda parte, é disponibilizada gratuitamente e, ainda, não impõe limitações arbitrárias sobre seus dados, sendo utilizada diariamente em variados campos de pesquisa que vão desde a engenharia aeroespacial até a biologia molecular.

Para o desenvolvimento do sistema *Web* foi utilizado o HTML (*HyperText Markup Language*), que utiliza etiquetas com comandos de formatação da linguagem, o Flash,

software que suporta animações interativas embutidas em um navegador *Web*, e o PHP (*Hypertext Preprocessor*), que permite gerar conteúdo dinâmico.

2. 2. LEVANTAMENTO DOS DADOS UTILIZADOS

A coleta das amostras e o processo de sequenciamento foram realizados pelos pesquisadores do Departamento de Biotecnologia da Universidade de Ribeirão Preto (UNAERP) sob a coordenação da Prof^a Dr^a Sônia M. Zingaretti.

A prospecção de genes de interesse em glândula de peçonha da aranha *Parawixia bistriata* disponibilizou um total de 1223 placas. A partir desses cromatogramas gerados deu-se início a fase de análise das bases e arranjos das sequências. Todo esse processo, como dito anteriormente, foi realizado através de ferramentas de bioinformática e será descrito com maiores detalhes nas próximas seções.

2. 3. DEFINIÇÃO DO ARRANJO DAS SEQUÊNCIAS

Nesta fase do projeto foi realizado o *pipeline* para o arranjo das sequências. Uma coleção de dados de um programa é utilizada como coleção de entrada de outro programa, com uma semântica apropriada. Este conceito é denominado *pipeline*.

Desta forma, de posse dos cromatogramas e das sequências de vetor utilizadas no sequenciamento, o arranjo de sequências foi realizado através do *pipeline* PHRED/ PHD2FASTA/ CROSSMATCH/ CAP3. A seguir, é possível observar um diagrama (FIGURA 2.1.) com o *pipeline* executado neste projeto.

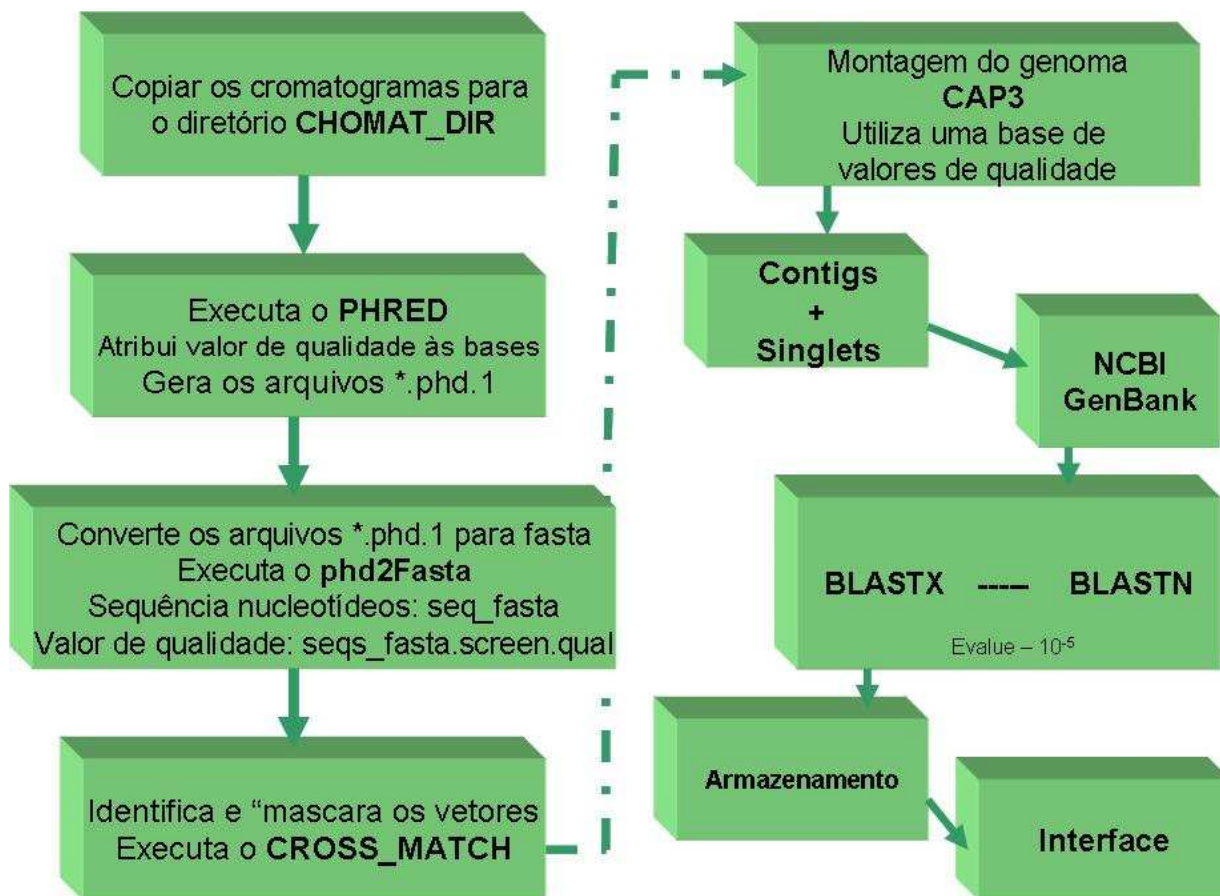


Figura 2.1. – *Pipeline* utilizado no arranjo (PHRED, PHD2FASTA, CROSSMATCH e Cap3) e alinhamento das seqüências (BlastX e BlastN), e em seguida armazenamento e disponibilização por interface gráfica.

É possível observar que se inicia com a cópia dos cromatogramas no diretório especificado para a execução do PHRED. Esses dados foram utilizados pelo PHD2FASTA, que converte os dados gerados em FASTA para que o CROSSMATCH identifique e mascare os vetores contaminantes. Os dados provenientes do CROSSMATCH foram montados pelo CAP3 que separa os dados em *singlets* e *contigs* para, em seguida, serem comparados, pelo BlastX e BlastN, com banco de dados internacionais disponibilizados pelo NCBI. A apresentação dos resultados seguirá a ordem na qual os programas foram utilizados no *pipeline*.

2. 4. DEFINIÇÃO E USO DE FERRAMENTAS DE BIOINFORMÁTICA

2. 4. 1. PHRED

Esta ferramenta pode ser associada a um digitalizador de leituras de DNA, sendo uma forma estruturada de resolver problemas em uma sequência lógica.

Após a entrada dos cromatogramas, o software PHRED¹ reconhece a sequência de nucleotídeos a partir do arquivo de dados brutos provenientes dos seqüenciadores automáticos de DNA, lê os dados, monta a estrutura de nomenclatura da placa sequenciada para facilitar a identificação e localização de cada clone, e decodifica cada sequência, atribuindo à qualidade do sinal, chamado *base calling*, que são valores de qualidade atribuídos às bases constituintes da sequência nucleotídica e, por fim, gera arquivos de saída contendo informações sobre o *base calling* e os valores de qualidades.

Para montar a estrutura, decodificar a sequência e atribuir valor de qualidade, o software utiliza algoritmos de tratamento de sinais, e para atribuir os valores de qualidade da base, o algoritmo passa por quatro fases na qual determina o pico ideal e o pico observado, compara os dois picos através de programação dinâmica e por fim, quantifica a qualidade do sinal (*base calling*). Tal valor é estimado através do erro de

¹ <http://bozeman.genome.washington.edu/phrap.docs/phred.html>

leitura calculado para cada base e corresponde a um inteiro entre 0 e 99, e quanto maior o valor de qualidade, menor o erro.

A execução do PHRED é feita através de linha de comando. Na FIGURA 2.2. pode-se observar um exemplo de sintaxe, montado de acordo com a necessidade do projeto. Entretanto, vários valores “default” podem e, dependendo do projeto, devem ser alterados na busca por resultados mais precisos.



FIGURA 2.2. Um das sintaxes possíveis de execução do PHRED

Em seguida, a ferramenta PHRED gera um único arquivo de saída (*.phd) contendo as bases lidas pelo seqüenciador com suas respectivas qualidades. Para separar as sequências em arquivos individuais, fez-se uso de um script em Perl.

Na FIGURA 2.3. é apresentado um arquivo de saída .phd, da ferramenta PHRED, que possui um cabeçalho e três colunas, sendo que a primeira indica a base, a segunda indica o valor de qualidade e a terceira indica a posição da base no traço.

```

BEGIN_SEQUENCE 01PB4A01.g

BEGIN_COMMENT

CHROMAT_FILE: 01PB4A01.g
ABI_THUMBPRINT: 102136020110002170212314215255
PHRED_VERSION: 0.990722.f
CALL_METHOD: phred
QUALITY_LEVELS: 99
TIME: Wed Feb 27 16:17:49 2008
TRACE_ARRAY_MIN_INDEX: 0
TRACE_ARRAY_MAX_INDEX: 14058
TRIM: -1 -1 0.0500
CHEM: unknown
DYE: unknown

END_COMMENT

BEGIN_DNA
c 7 5      a 10 369      a 8 2180      g 8 13810
c 10 15    t 7 388      c 9 2190      g 8 13825
t 9 34     t 7 397      g 8 2207      a 10 13838
a 10 52    g 8 2220      g 8 2220      c 10 13847
t 7 73     g 11 2241     g 11 2241     a 10 13860
c 7 81     c 8 2253      c 8 2253      c 8 13869
c 7 94     t 8 2264      t 8 2264      c 7 13883
t 8 108    g 13 473      a 9 2278      t 7 13896
g 8 117    t 11 487      t 11 2295     g 8 13906
a 8 129    g 13 503      c 9 2312      t 10 13915
t 8 145    a 16 516      a 10 2321     a 10 13932
a 8 156    a 9 530      a 10 2337     a 10 13944
a 8 173    c 8 545      a 13 2353     g 6 13949
g 8 182    g 8 556      g 16 2371     c 6 13965
c 11 195   a 11 573      a 16 2387     g 8 13973
g 6 208    c 10 592      t 9 2402      c 9 13983
c 6 219    g 17 605      t 7 2419      c 7 14004
t 6 242    a 17 619      g 7 2430      c 6 14011
c 6 249    c 14 633      g 8 2437      g 9 14024
g 10 261   a 14 645      a 8 2455      c 9 14035
c 9 278    a 11 659      a 9 2473      c 9 14048
a 9 285    a 11 678      c 9 2484      a 4 14057
END_DNA

END_SEQUENCE

```

Cabeçalho

Figura 2.3. Arquivo de saída PHRED (.phd)

A próxima etapa é a utilização do PHD2FASTA que converte o arquivo (*.phd) para FASTA após executar, entre as sintaxes possíveis, a seguinte sintaxe (FIGURA 2.4.):

ph2fasta -id phd_dir **-os** seqs_fasta **-oq** seqs_fasta.qual

Escreve um arquivo de seqüências com o nome de "seq_fasta".

Escreve um arquivo de qualidade com o nome de "seq_fasta.qual".

FIGURA 2.4. Um das sintaxes possíveis de execução do FASTA

O arquivo FASTA apresenta uma linha de comentário indicada por ">" seguida pelo nome e origem da sequência, com seqüências em padrão de símbolos de uma letra, apresentado na FIGURA 2.5. O arquivo FASTA é um arquivo único, e para separar as seqüências em arquivos individuais, fez-se uso de um script em Perl.

```
>O6Pb05DILA09.g CHROMAT FILE: O6Pb05DILA09.g PHD FILE: O6Pb05DILA09.g.phd.1
CHEM: unknown DYE: unknown TIME: Wed Feb 27 16:17:27 2008
ngcgengancctagacacctcgnacgctgtccttacgatccnnaatcccg
ggtcgactctatggngatgggcccagcgccttgtgctcgctggccaatgcc
gccgcggtcacctgngcgatcatgaagccggagttcacaccaccggttgcc
caccaggaatggcggcagttgcgacatgtgcttgtccagcagcaacgana
tacgacgctcgctcagggagccgatttccgcgatggccaacgccatgtgg
tcggggcccatggccaccggttcagcgtggaagtggccggcggaaatcac
gtcaccttcagccgcaaacaccaacgggttatccgatacggcgtnggctt
ccaccaccagcacttcggccgcttggcggaaactgggtcaggcaggcggcc
atgacttggcgctggcaacgcagagagtacgggtcttggacctgtcgca
gttctcgtgggatggngacacttcgctgctttcaccagcagggcgcgat
atggggcgcctctagaggatccaagcttacgtacgcgtgcatgcgacgtc
atagctcttctatgtgtcacctataattcattcaatggcgcgtcttttac
aacgtcgtgatggaaaacctggcgttaccaacttatccgctgcaacaca
tcccccttccgcagtgccggttatagcgagaaggccgaccgatcgccttc
caacagttgcccaccttgatggggcgtgggacgcgcctgttacgcgcatt
nagcgcgcgggtgtgggtggttaccgcacaggtaacgctaaaattgcacag
gccttacgcccgttccctcggettctcccttcccttctcgcagttcgcc
ggtttcccgtaacetctaaatcggggctcctttaggttccatttatggt
tacggccttcgacccaaaaactgattaggtgaaggtcactangggcatcc
ccctgtttaaggttcgcctttgagtgagtgccattccataatgggcctg
gtccaagggacaaaatcaaactttcggcttctttgatataaagatttg
cgctttgccctctggtaaaaggccggttccaaattacgcgattttacaaa
tttagcccaattggggccttgggaaggcgggaaccattggttttctaa
aatcaaatgtccgcgggaaacaaccgtaaggtctatttgaagggttgg
tttcatccgtgcccttcccttgggttccctcgttggcaccaaaacggg
agaaaaaggaactggtcgatggcccagggtccagcgaacctnatcccc
aaattccaagaaaataatggtggggcgtatcccggggc
```

Comentário

Sequências

Figura 2.5. Arquivo de saída no formato FASTA

A saída do PHD2FASTA, no formato FASTA será utilizada pelo CROSSMATCH para identificar e mascarar as bases contaminantes.

2. 4. 2. CROSSMATCH

No CROSSMATCH² os dados são submetidos a uma filtragem para remoção das bases contaminantes. O algoritmo compara cada sequência no formato FASTA com as sequências contidas no arquivo vetor e “mascara” as bases contaminantes substituindo-as por um “X”. A ferramenta pode ser adquirida gratuitamente para fins acadêmicos.

O CROSSMATCH também é executado na linha de comando. Pode-se observar, a seguir a sintaxe escolhida para executar a ferramenta (FIGURA 2.6.):

```
CROSS_MATCH SEQS_FASTA VECTOR.SEQ -MINMATCH 12 -  
MINISCORE 20 -SCREEN > SCREE.OUT
```

FIGURA 2.6. Um das sintaxes possíveis de execução do FASTA

Pode-se observar, na FIGURA 2.7., as bases contaminantes marcadas com “X”.

² <http://www.phrap.org/phredphrap/general.html>

```

>O1Pb05sarra01.g CHROMAT_FILE: O1Pb05sarra01.g PHD_FILE: O1Pb05sarra01.g.phd.1
CHEM: unknown DYE: unknown TIME: Wed Feb 27 16:17:35 2008
AGCCNCCNCCCTGGATCGTCTGAGGTACGGTCCGGAATCCCGGGTCGAA
CCACGCGTCCGATGTGTCAGAATGATTCTGAAAGTGAAGAAAGTTAGCGA
TCCTTTATGTGCAGGCATACAATAAGCACCAGCTATACTCCATGTAATTG
GAAAGAATGGTTAAGAAGACGTCTGATGATATAACAAATGATACCAGAAAAG
GAAAGTTCGGATAGAAAGAAAAGTGAAGAGTAAAGTTGTGTCTTCGAAAAAA
CAGACAATCCAAGACAGAGGTCATTGCGCTATACTAAACTGTACTACAGA
AGXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXTCTTCCTTAATGCCCTAAATTTGATTGATCCGCCACGTTA
ACGATATACTTGCTGGGACAAACCGCCGGTCCCAAAATTAACCCCTTCC
TAAATCCCCCTTCCCGGTGTCCCAATAAAAATGAACGGGGGGTCCCCT
TCGAACAGTTTGGGCTTAAAGGCAATGGCCCCACCCTTATGGGTAAA
GCCCGGGTGTGGGTTCACAAAAGAACCTTCCCTTTCGAGCCGGAACCC
CGTCTCTTCGTGGATCCTTTCACAAAAGGAAAAAACCCACCCCCCGCGA
TATTTATTTGGGGTACCGGTTTGGGCCACTCTCGCGTTGTGGGACACCCG
CCCCAAAACAAAATGGGGGGGTTCCACAGGGCCCCCCCCCAAAAAGG
GGCCTCCCGGAGATGTGGCCCCCTCTATGGGGTCTTTCCACCAGAAA
ATTCCCCCTTACGAATCCCTTATATTTTAAAGAATGCGAAAAGGCGCTT
GTTTAAAAAATGGGGGCCATAATCCTTCGGCGTATACATTTACTTCA
CACCACAAGTCTGGGGTAATATCGGACAACCACGGCGGCTGCGTTATAAA
TAATCGTCTCGCGCGCAAAAACACATGTGGCCCCGCTATCGCATCAGAG
TACTTCTTCGCCCA

```

Bases contaminantes

Figura 2.7. Arquivo de saída CROSSMATCH

2. 4. 3. CAP3

A montagem do genoma a partir dos fragmentos lidos é realizada pela ferramenta CAP3³, para isso utiliza uma base de valores de qualidade produzidos pelo PHRED, construindo um alinhamento de múltiplas sequências e gerando as sequências consensos, separando-as em *Contigs* e *Singlets*.

³ <http://genome.cs.mtu.edu/cap/cap3.html>

O CAP3 gera as *contigs* com mais de um *read*, ou seja, grupos com alinhamento de múltiplas sequências, gerando as sequências *consenso*, e os *singlets*, os quais são sequências únicas, que não conseguiram se alinhar a outra sequência para formar um *contig*.

Em seguida, o algoritmo identifica e numera as sobreposições, utilizando frente-verso para corrigir erros e montar as *contigs*. As duas leituras devem ser feitas em sentidos opostos da molécula de DNA, dentro de um determinado intervalo de distância.

O CAP3 gera o arquivo cap3.out, contendo, no início do arquivo, os *reads* de cada *contig* (FIGURA 2.8.) e, no mesmo arquivo de saída, as sequências de cada *contig* (FIGURA2.9.).

```

***** Contig 1 *****
02Pb05sarrA07.g+
    79PB4F04.g- is in 02Pb05sarrA07.g+
    15Pb02DILCO2.g+ is in 02Pb05sarrA07.g+
    03Pb01DILA02.g+ is in 15Pb02DILCO2.g+
    17Pb05sarrC03.g+ is in 02Pb05sarrA07.g+
    60Pb05sarrB12.g+ is in 17Pb05sarrC03.g+
    23Pb05sarrD03.g+ is in 17Pb05sarrC03.g+
    16Pb05sarrC08.g+ is in 02Pb05sarrA07.g+
11Pb05sarrB03.g+
    71Pb01DILD06.g+ is in 11Pb05sarrB03.g+
    82Pb05sarrF11.g+ is in 71Pb01DILD06.g+
    02Pb05DILA07.g+ is in 11Pb05sarrB03.g+
    66Pb05sarrC12.g+ is in 11Pb05sarrB03.g+
    11Pb05DILB03.g+ is in 11Pb05sarrB03.g+
27Pb02DILE02.g+
    09Pb02DILB02.g+ is in 27Pb02DILE02.g+
17Pb02DILCO3.g+
    27Pb05sarrE02.g+ is in 17Pb02DILCO3.g+
    04Pb01DILA08.g+ is in 17Pb02DILCO3.g+
***** Contig 2 *****
03Pb05sarrA02.g+
05Pb02DILA03.g+
***** Contig 3 *****
04PB4A08.g+
06PB4A09.g+
    40PB4G08.g+ is in 06PB4A09.g+
    48PB4H09.g+ is in 06PB4A09.g+
    60PB4B12.g+ is in 06PB4A09.g+
    74PB4E10.g+ is in 06PB4A09.g+
    18PB4C09.g+ is in 06PB4A09.g+
    42PB4G09.g+ is in 18PB4C09.g+
    01PB4A01.g+ is in 18PB4C09.g+
    54PB4A12.g+ is in 18PB4C09.g+
    07PB4B01.g+ is in 54PB4A12.g+
86PB4G10.g+
67Pb02DILD04.g-
    41PB4G03.g+ is in 67Pb02DILD04.g-
    25Pb05DILE01.g- is in 67Pb02DILD04.g-

```

Figura 2.8. Arquivo de saída CAP3 - Contigs (reads)

```

***** Contig 1 *****
      .   :   .   :   .   :   .   :   .   :   .   :
02Pb05arrA07.g+  AGTACGGTCCGGAATCCCGGGTTCGTACCACGCGTCCGAGAAGAGCAGCTCTGCTAACGTG
79PB4F04.g-      GCAGCTCTGCTAACGTG
23Pb05arrD03.g+      CCACGCTTCCG-GAAGAGCACTTCTGCTAACGTG
17Pb05arrC03.g+      GAGCAGCTCTGCTAACGTG
60Pb05arrB12.g+      GCTCTGCTAACGTG
16Pb05arrC08.g+      GGAATCCCGGGTTCGATCCACGCGTCC----GGAGCAGCTCTGCTAACGTG
11Pb05arrB03.g+      GCTCTGCTAACGTG
02Pb05DILA07.g+      ACGTG

consensus Sequências AGTACGGTCCGGAATCCCGGGTTCGATCCACGCGTCCGAGAAGAGCAGCTCTGCTAACGTG

      .   :   .   :   .   :   .   :   .   :   .   :
02Pb05arrA07.g+  TTTGCCAGGAGAAAATAATAATAAAAAAATATGTCGAAGTACGTGACATTTTTGATTTTGC
79PB4F04.g-      TTTGCCAGGAGAAAATAATAATAAAAAAATATGTCGAAGTACGTGACATTTTTGATTTTGC
23Pb05arrD03.g+  TTTGCCAGGAGAAAATAATAATAAAAAAATATGTCGAAGTACGTGACATTTTTGATTTTGC
17Pb05arrC03.g+  TTTGCCAGGAGAAAATAATAATAAAAAAATATGTCGAAGTACGTGACATTTTTGATTTTGC
60Pb05arrB12.g+  TCTGCCAGGAGAAAATACTAATAAAAAAATATGTCGAAGTACGTGACATTTTTGATTTTGC
16Pb05arrC08.g+  TTTGCCAGGAGAAAATAATAATAAAAAAATATGTCGAAGTACGTGACATTTTTGATTTTGC
11Pb05arrB03.g+  TTTGCCAGGAGAAAATAATAATAAAAAAATATGTCGAAGTACGTGACATTTTTGATTTTGC
02Pb05DILA07.g+  TTTGCCAGGAGAAAATAATAATMNAAAAAAATATGTCGAAGTACGTGACATTTINTGATTTGGC
11Pb05DILB03.g+      ATAATAAATAAAAAANATATGTCGAAGTACGTGACATTTGTGATTTTGC
15Pb02DILC02.g+      TAAAAAATTTTGTCAAATTACTTGACTTTTTTTGATTTTGC
SONIAPL09F_B05_b_04+      GACATTTTTGATTTTCGC
SONIAPL08F_C02_b_00+      GACATTTTTGATTTTGC

consensus TTTGCCAGGAGAAAATAATAATAAAAAAATATGTCGAAGTACGTGACATTTTTGATTTTGC

```

Figura 2.9. Arquivo de saída CAP3 – Contigs (Sequências)

O CAP3³ ainda gera os *singlets* (FIGURA 2.10.), sequências que não conseguiram se alinhar a outra sequência para formar um *contig*, ou seja, um arquivo FASTA contendo *read* sem nenhum *match*.

```
>CHROMAT_FILE: PBO1001B07F.g PHD_FILE: PBO1001B07F.g.phd.1
CHEM: unknown DYE: unknown TIME: Thu Aug 21 14:11:27 2008
GGTCCMNGGGGGGGCTACCTAGNGTCCCTGAGTACGCTCCGNAAAATCCCCGGTTCAAGC
GCAGCGTCCGATACGTATCACTGCCATCGAAAGCGCGGGCTGCGGGCCTCCAAACAGCCG
AACAAATCAGTCTGAGCCCAACACCGAGGAGCGAAGGTGGCCGGCGATGGACCACTGCAGT
TCGCCCCGTGGTTCATGTCGAGCACGCCGATCAGCAGCGTGACGAACATCATGTTCCGGTAG
TTCTCGGACAGGCGGTTGTTGATCTTCTGCATGATCAGCGCCGGGTCGGTCTCGTCCTCG
GCGGTGGCGCGGATCAGGGTGC CGTGACCGCCATGAACAGCGCCGCGGCACGCCCTTG
TCCGACACGTCCCCGATCGCCAGGCATAGCCGGCCGTCGGGCAGCGTGAAGTAATCGTAG
AGGTCGCCGCCCACTTCCTTGGCCGGCAGCATCAGGCGTTCAAGTCGATCTGCTCGCGG
GTGATGGGCGACAGCGGCACCGGCAGCAGGCCGAGCTGGATCGCGCGGGCGATGTTACG
TCGCTCTCGAAGCGTTTCGCGCGCGGTGGTCTCGCGCATCAAGGGGGCGAAGTTCTCGCGC
AGCTTGGGGTTCATTGAACAAGAACGAAGGCGGCAAGCGGGCCCATTTCTGTCCTGGTGTT
TGTCGGGCAAGGCCGCGAATTTGGCCGGCAGGCCCATTCGGGGTCCAGTTCTGGTCGG
GGCAAAAAGCGGGCCTTATGGGTTAAGGTTTTAGCGGGCCCGTCTAAAAAGAGTCCAACTT
AAGTTAGCGTGGATTGGAAGTCATAACTCCCTCTATTATTGTTACCTAATTTCAATTCAA
TGGGGCCCGGTTAAAAACAGTTCCGGAATGGGAAAAACCGGGGGTTACCAAATTAATCGG
CTTGAAAAGAAATCCCCCTTTTGCAAAATGGGGGTTATAAAGAAAAAA
```

Figura 2.10.. Arquivo de saída CAP3 - Singlets

2. 4. 4. BLAST

A ferramenta Blast permite comparar uma sequência de DNA ou proteína com um banco de dados de sequências (proteínas e DNA). O resultado pode ser utilizado para inferir relacionamentos funcionais e evolucionários, assim como, ajudar na identificação de membros de famílias gênicas à medida que procura identificar a presença de uma sequência, de DNA ou proteína, suficientemente parecida com a pesquisada, previamente depositada em banco de dados público.

A ferramenta, adquirida gratuitamente ou executada *online*, ainda descarta os resultados não produtivos, e estende a vizinhança da região de homologia detectada até não mais conseguir, retornando as sequências com maior homologia.

Pode ser dividido em etapas, iniciando pela montagem da lista de palavras, segue procurando pelas palavras em cada sequência do banco, em seguida, para cada palavra encontrada na sequência é realizada uma extensão em ambas as direções, finalizando com o alinhamento das sequências. Desta forma, não objetiva atingir o alinhamento final, mas sim identificar sequências com nível de homologia significativo.

Para se avaliar um alinhamento é significativo ou se é uma mera coincidência de alguns poucos pares de bases que apresentaram alguma identidade entre as sequências, é necessário saber qual a possibilidade daquela similaridade ter ocorrido ao acaso. Desta forma devem ser observados alguns valores que são atribuídos pelos programas escolhidos para o pareamento entre as sequências, aferindo a similaridade dentro do segmento comparado.

O *e-value* ou equação associada ao valor do *score*, é um importante valor de análise do alinhamento, pois mede a possibilidade do evento de alinhamento ocorrer ao acaso. Quanto menor seu valor, menor a chance de tal comparação ter sido encontrada por pura coincidência, desta forma, o melhor alinhamento possível é alcançado com *e-value* igual a zero.

Há várias modalidades de BLAST, que podem ser usados para buscas em diferentes bases de dados de sequências, entre elas, o BLASTN⁴, que busca homologia

⁴ http://www.incogen.com/public_documents/vibe/details/blastn.html

entre sequências de nucleotídeos, e o BLASTX⁵, que busca sequências de nucleotídeos e proteínas.

Para os alinhamentos, foram utilizados o BlastX e o BlastN, com um valor limitante de *e-value* $< 10^{-5}$, selecionado pelos pesquisadores do projeto como um valor significativo para se adquirir alinhamentos de alta qualidade. Assim, as sequências *contigs* e *singlets* foram comparadas com todas as sequências de um banco de dados disponível pelo NCBI com as sequências não redundantes do Genbank.

O BlastX (FIGURA 2.11.) traduz a sequência de nucleotídeos para proteína, isto é, compara uma sequência de nucleotídeos, com um banco de dados de proteínas e classifica as sequências com similaridades em ordem da maior para a menor similaridade de acordo com o *e-value*.

⁵ http://www.incogen.com/public_documents/vibe/details/blastx.html

Sequences producing significant alignments:				Score (bits)	E Value
ref YP_350336.1	chaperone protein HscA	[Pseudomonas fluorescens...	432	e-119	
ref YP_262041.1	chaperone protein HscA	[Pseudomonas fluorescens...	426	e-117	
ref YP_273571.1	chaperone protein HscA	[Pseudomonas syringae pv...	404	e-111	
ref YP_234330.1	chaperone protein HscA	[Pseudomonas syringae pv...	404	e-111	
ref NP_791253.1	chaperone protein HscA	[Pseudomonas syringae pv...	402	e-110	
ref NP_743007.1	chaperone protein HscA	[Pseudomonas putida KT24...	402	e-110	
ref YP_001266222.1	chaperone protein HscA	[Pseudomonas putida F...	402	e-110	
ref YP_001667135.1	Fe-S protein assembly chaperone HscA	[Pseudo...	399	e-109	
dbj BAD01053.1	heat shock protein HscA	[Pseudomonas putida]	399	e-109	
ref YP_606731.1	chaperone protein HscA	[Pseudomonas entomophila...	399	e-109	
ref ZP_01641058.1	Fe-S protein assembly chaperone HscA	[Pseudom...	397	e-109	
ref YP_001188987.1	chaperone protein HscA	[Pseudomonas mendocin...	395	e-108	
gb AAC79497.2	heat shock protein 66-KDa	[Pseudomonas aeruginosa]	393	e-107	
ref YP_789322.1	chaperone protein HscA	[Pseudomonas aeruginosa ...	385	e-105	
ref NP_252499.1	chaperone protein HscA	[Pseudomonas aeruginosa ...	385	e-105	
ref ZP_00973574.1	COG0443: Molecular chaperone	[Pseudomonas aer...	384	e-105	
ref YP_001346687.1	Fe-S protein assembly chaperone HscA	[Pseudo...	383	e-104	
ref ZP_00418128.1	Fe-S protein assembly chaperone HscA	[Azotoba...	382	e-104	
sp O69221 HSCA_AZОВI	Chaperone protein hscA homolog	>gi 3046319 ...	380	e-104	
ref YP_001173521.1	chaperone protein HscA	[Pseudomonas stutzeri...	380	e-103	

>ref|YP_350336.1| chaperone protein HscA [Pseudomonas fluorescens PfO-1]
sp|Q3K7A9|HSCA_PSEPF Chaperone protein hscA homolog
gb|ABA76345.1| Fe-S protein assembly chaperone HscA [Pseudomonas fluorescens PfO-1]
Length = 621

Score = 432 bits (1110), Expect = e-119
Identities = 241/306 (78%), Positives = 251/306 (82%), Gaps = 8/306 (2%)
Frame = -2

Query: 957 AWGGDT*ITHSRVDH-----HQRFFSDWDPGAQLILFQPPAPPKSP-D*SCVVEVSY 802
A GGD+ + DH +D DPGAQ L Q K S VEV+Y
Sbjct: 229 ATGGDSALGGDDFDHAIAGWIIIESASLSADLDPGAQRSLLQAACAAKEALTDSDSVEVAY 288

Figura 2.11. Arquivo de saída BlastX

O BlastN compara uma sequência de nucleotídeos de entrada com todas as sequências de nucleotídeos do banco de dados (FIGURA 2.12.).

2. 5. DESENVOLVIMENTO DO SISTEMA WEB

Para a disponibilização dos resultados, optou-se pela *Internet*, considerada uma excelente ferramenta para facilitar a comunicação de pessoas, empresas e instituições, disponibilizando informações a diferentes pessoas espalhadas por todo o mundo, 24 horas por dia, 7 dias por semana. A facilidade de utilizar esse meio de comunicação possibilita a exposição de conteúdos diversos, a usuários específicos. Segundo Tim Bernes Lee (concebeu a *web* em 1990 no modelo utilizado até hoje), muitas pesquisas foram feitas na tentativa de solucionar o problema de reorganização da massa de conteúdo da *web*, e juntamente com um grande grupo de pesquisadores dos Estados Unidos, o W3C (*World Wide Web Consortium*) propõem a construção de uma camada de tecnologia sem alterar a interface, com uma representação estrutural e semântica automatizada das informações existentes, e tais informações foram levadas em consideração para a finalização da interface gráfica.

Desta forma, para o sistema *web*, buscou-se um modelo que permita a compreensão e o gerenciamento de todas as formas de conteúdo, com a valorização semântica dos conteúdos e de agentes coletores de conteúdos para processar as informações e os resultados dos softwares utilizados.

A página estruturada foi desenvolvida com a utilização de elementos gráficos com base em conceitos ergonômicos de forma a transmitir, de forma clara e objetiva, a mensagem desejada. Para Vassos (1998, p. 146) o caráter da *web* é determinado pelo estilo de escrita, seja ele formal ou informal, com ou sem o uso de jargões, pela fonte

usada (casual ou conservadora), por fatores como a cor do texto e do fundo, e ainda, pelo uso de elementos adicionais tais como animação. Portanto, é fundamental na construção de uma página observar esses critérios e sua harmonização.

O sistema *web*, apresentado no Capítulo seguinte, foi desenvolvido e encontra-se disponível no endereço: <http://bioinfo1.fmrp.usp.br/~ccintra/>

2. 6. CONSIDERAÇÕES FINAIS

Todas as linguagens de programação, assim como as ferramentas de bioinformática, foram selecionadas para que o *pipeline* e a visualização via *web* pudessem ser desenvolvidos de forma prática, dinâmica e segura. O desenvolvimento de um sistema *web* foi realizado após o levantamento dos requisitos do usuário, a análise das tarefas estabelecendo os requisitos do sistema, *design* da interface, entre outras técnicas, aplicadas na busca pela melhor apresentação dos resultados aos pesquisadores participantes do projeto. Todos os dados provenientes dos programas apresentados, foram disponibilizados na *web*, para os pesquisadores do projeto da FAPESP, buscando sempre a máxima otimização da página a ser apresentada ao usuário com o intuito de possibilitar a análise dos resultados de forma prática, rápida e segura.

Capítulo 3 – Resultados

3. 1. CONSIDERAÇÕES INICIAIS

Para a visualização dos resultados, foi projetada uma interface simples e eficiente, possibilitando ao usuário visualizar os resultados, assim como, comparar o resultado de cada programa.

Os cromatogramas foram obtidos após o sequenciamento realizado no Departamento de Biotecnologia da Universidade de Ribeirão Preto - UNAERP.

As amostras foram encaminhadas ao Laboratório de Bioinformática do Departamento de Genética da Faculdade de Medicina de Ribeirão Preto. Para o arranjo das sequências obtidas, utilizaram-se programas de bioinformática, como o PHRED, CROSSMATCH e CAP3. Para os alinhamentos, utilizou-se as ferramentas BlastX, para sequências de nucleotídeos e de proteínas, e BlastN, para sequências de nucleotídeos, sequências essas, disponibilizados pelo NCBI, utilizando um valor limitante de *e-value* $< 10^{-5}$ selecionado pelos pesquisadores do projeto como um valor significativo para se adquirir sequências de alta similaridade.

3.2. RESULTADOS DO CAP3

Após o Cap3, obteve-se 124 sequências com mais de um *read* que formaram sequências *consensos* chamadas *contigs*, e 656 sequências que não conseguiram se alinhar a outra sequência para formar um *contig*, chamadas *singlets*. Com esses resultados, podemos observar a FIGURA 3.1. com a estatística do resultado do Cap3.

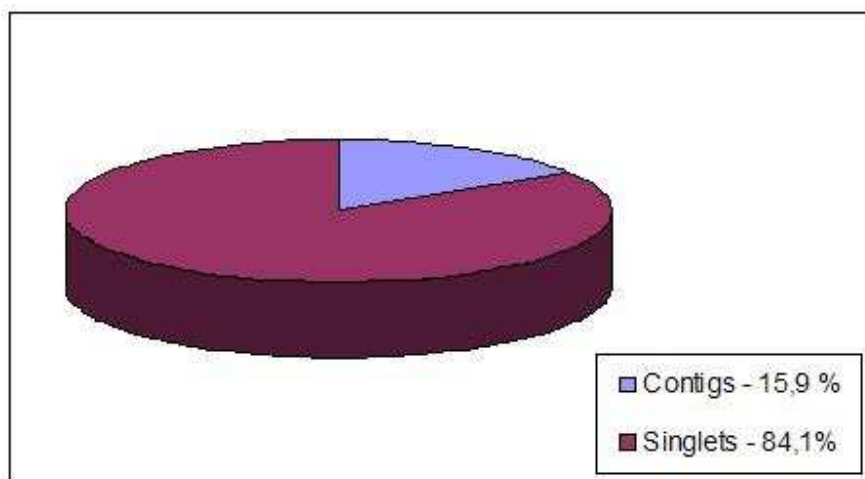


Figura 3.1. Estatística Cap3

3. 3. SISTEMA WEB

A página inicial (FIGURA 3.2.) do sistema *web* apresenta ao usuário um breve resumo sobre o levantamento dos dados e sobre os programas de bioinformática envolvidos no projeto. Botões de fácil acesso ao resumo do projeto, aos programas utilizados no projeto, aos resultados obtidos por cada ferramenta de bioinformática e ainda, ao grupo dos dois laboratórios envolvidos no projeto, estão disponibilizados do lado esquerdo do *site*. Pode-se observar ainda, um título inicial “Biblioteca digital de glândula de peçonha da aranha *Parawixia bistriata*” com a imagem da aranha *Parawixia bistriata* e uma sequência binária ao fundo, sendo que estes conjuntos de informações estão presentes em todas as páginas do sistema *web*, bem como os logos inferiores do Departamento de Bioinformática do lado esquerdo e da Universidade de Ribeirão Preto (UNAERP) do lado direito.



FIGURA 3.2. Página inicial do site

Para proporcionar ao visitante uma informação mais clara sobre o objetivo do projeto intitulado “Desenvolvimento de um sistema *web* para visualização e análise *in silico* de biblioteca de cDNA da peçonha da aranha *Parawixia bistriata*”, disponibilizou-se um resumo do projeto (FIGURA 3.3) contendo uma breve introdução, as ferramentas de bioinformática utilizadas, bem como as palavras chave para o projeto. Ainda apresenta ao usuário, botões de fácil acesso para voltar à página inicial ou seguir para as demais páginas disponibilizadas.

Biblioteca digital de glândula de peçonha de aranha *Parawixia bistriata*

Home	Resumo	Programas Utilizados	Resultados	Grupo
------	--------	----------------------	------------	-------

Biblioteca digital de glândula de peçonha de aranha *Parawixia bistriata*

Animais peçonhentos de diferentes filos desenvolveram poderosas peçonhas, arsenais químicos com substâncias capazes de atordoar, paralisar ou matar outros organismos (MCCORMIC & MEINWALD, 1993). Muitos componentes das peçonhas funcionam como neurotoxinas, tendo notável especificidade e afinidade de ligação por receptores ou canais iônicos neuronais (USHERWOOD, 1994).

As peçonhas de artrópodos são ricas fontes de neurotoxinas, verdadeiras ferramentas moleculares com ação seletiva e específica de grande relevância clínico-científica. A aranha *Parawixia bistriata*, cuja peçonha e compostos protéico-peptídicos são objetos desse estudo, é facilmente encontrada na região de Ribeirão Preto garantindo assim, a formação de uma biblioteca de peçonha e tecido glandular considerável na busca pela identificação de novos genes e possíveis novos produtos.

Assim, esse projeto, apoiado pela FAPESP, prevê a prospecção de genes de interesse em glândula de peçonha da aranha *Parawixia bistriata*, pelo sequenciamento de clones de uma biblioteca de cDNA, identificação e caracterização funcional das ESTs encontrados. Para revelar o conhecimento contido nos dados e ainda, buscar similaridades em bancos de dados, ferramentas de Bioinformática serão utilizadas na análise da biblioteca de cDNA de glândula de peçonha da aranha *Parawixia bistriata*, automatizando a obtenção, distribuição e análise de dados genéticos, além de combiná-los com modelos matemáticos.

A apresentação dos resultados de cada fase da análise será feita através de uma interface gráfica eliminando, assim, a necessidade dos pesquisadores envolvidos do uso de comandos de execução, essenciais no manuseio das ferramentas de Bioinformática, seja na entrada dos dados, nas opções de processamento ou nas opções de saída.

O resultado final será um sistema de seleção dos resultados por menu, com o uso de gráficos, botões e técnicas de scrolling, possibilitando a exibição de informações simultâneas em diferentes janelas.

Palavras chave: bioinformática, peçonha, ESTs, banco de genes;

Site melhor visualizado na resolução 800 X 600
Web designer: camila_infobio@yahoo.com.br

3.3. Resumo disponível no site

A FIGURA 3.4 apresenta uma interface ao usuário que disponibiliza um *link* para o *site* de cada programa utilizado no projeto bem como ao direcionar o cursor sobre o nome do programa, pode-se seguir o *pipeline* utilizado. Disponibiliza ao usuário, botões de fácil acesso para as demais páginas do sistema *web*.

Biblioteca digital de glândula de
peçonha de aranha *Parawixia bistriata*

Home Resumo **Programas Utilizados** Resultados Grupo

Programas Utilizados

arranjo de sequência

Phred Fasta Crossmatch

Cap3

alinhamento de sequência

NCBI

BlastX BlastN

Aproxime o mouse para acompanhar o pipeline.

Clique nos programas para ser direcionado aos sites.

UNAERP

Site melhor visualizado na resolução 800 X 600
Web designer: camifa_infobio@yahoo.com.br

FIGURA 3.4. Programas utilizados

Ao acessar a interface dos Resultados (FIGURA 3.5.), o usuário pode visualizar o *pipeline* do projeto e ainda ter disponível o acesso aos resultados das ferramentas utilizadas, seguindo o *pipeline*, iniciando-se pelo PHERD, seguindo pelo CROSSMATCH e Cap3, onde pode visualizar a estatística antes de acessar o resultado das *contigs* ou dos *singlets*. Em seguida, o usuário pode acessar o resultado do BlastX ou do BlastN.

Biblioteca digital de glândula de peçonha de aranha *Parawixia bistriata*

Home Resumo **Programas Utilizados** Resultados Grupo

Phred
Fasta
CrossMatch
Cap3
BlastX
BlastN

```

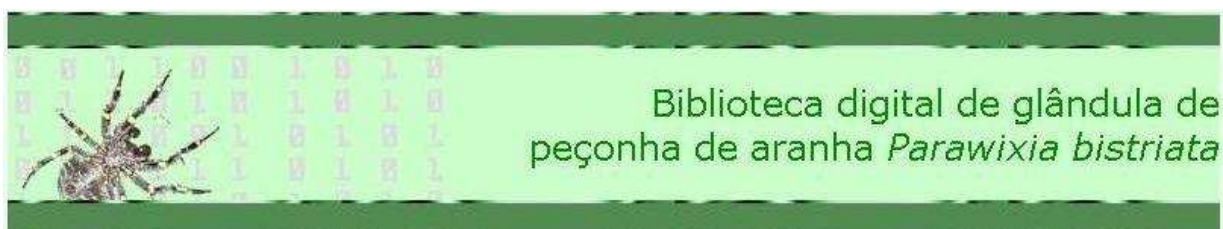
    graph TD
      A["Executa o PHRED  
Atribui valor de qualidade às bases  
Gera os arquivos *.phd 1"] --> B["Converte os arquivos *.phd para fasta  
Executa o ph2Fasta"]
      B --> C["Identifica e mascara os vetores  
Executa o CROSS_MATCH"]
      C --> D["Montagem do genoma  
CAP3  
base de valores de qualidade"]
      D --> E["Contigs + Singlets"]
      E --> F["NCBI GenBank"]
      F --> G["BLASTX ----- BLASTN"]
  
```

Site melhor visualizado na resolução 800 X 600
Web designer: camila_infobio@yahoo.com.br

3.5. Página de RESULTADOS, disponibilizada no site

Ao acessar o resultado de cada programa, de forma geral, o usuário pode visualizar todas as placas, separadas por arquivos, e ao selecionar algum arquivo, visualiza o resultado referente à ferramenta em questão. Por exemplo, após selecionar na página resultado a ferramenta PHRED, será apresentado a FIGURA 3.6. na qual

está apresentado todos os arquivos .phd, ao selecionar algum arquivo é apresentado o resultado da ferramenta PHRED para o arquivo .phd selecionado.



[PB01001A01F.g.phd.1](#)
[PB01001A02F.g.phd.1](#)
[PB01001A03F.g.phd.1](#)
[PB01001A04F.g.phd.1](#)
[PB01001A05F.g.phd.1](#)
[PB01001A06F.g.phd.1](#)
[PB01001A07F.g.phd.1](#)
[PB01001A08F.g.phd.1](#)
[PB01001A09F.g.phd.1](#)
[PB01001A10F.g.phd.1](#)
[PB01001A11F.g.phd.1](#)
[PB01001A12F.g.phd.1](#)
[PB01001B01F.g.phd.1](#)
[PB01001B02F.g.phd.1](#)
[PB01001B03F.g.phd.1](#)
[PB01001B04F.g.phd.1](#)
[PB01001B05F.g.phd.1](#)
[PB01001B06F.g.phd.1](#)
[PB01001B07F.g.phd.1](#)
[PB01001B08F.g.phd.1](#)
[PB01001B09F.g.phd.1](#)
[PB01001B10F.g.phd.1](#)
[PB01001B11F.g.phd.1](#)
[PB01001B12F.g.phd.1](#)
[PB01001C01F.g.phd.1](#)

FIGURA 3.6. Visualização dos resultados

O usuário pode visualizar os integrantes dos dois departamentos envolvidos no projeto ao acessar a interface Grupo (FIGURA 3.7.) que apresenta os integrantes do Grupo de Bioinformática do Departamento de Genética do lado esquerdo e os integrantes do Departamento de Biotecnologia do lado direito, sendo o primeiro da

Faculdade de Medicina de Ribeirão Preto – USP e o segundo da Universidade de Ribeirão Preto – UNAERP.

Biblioteca digital de glândula de peçonha de aranha *Parawixia bistriata*

Home Resumo Programas Utilizados Resultados **Grupo**

Grupo de Bioinformática
Departamento de Genética
Faculdade de Medicina de Ribeirão Preto
- USP -

Profª. Dra. Silvana Giuliatti
Saulo França Amui - Doutorando
André Luis da Silva Breve - Mestrando
Daniel Macedo de Melo Jorge - Mestrando
Gabriela Félix dos Santos - Mestrando
Rodrigo Martins Brandão - Mestrando
Thiago Yukio Kikuchi Oliveira - Mestrando
Camila Santana Justo Cintra Sampaio - Graduanda

Departamento de Biotecnologia
Faculdade de Ribeirão Preto
- UNAERP -

Dra Sonia M. Zingaretti Di Mauro
Dra. Ana Lucia Fachin- pesquisador
Dr. René de Oliveira Beleboni- pesquisador
Dr. Mozart Marins - pesquisador
Dra. Suzelei de Castro França- pesquisador
Patricia Roberto, MSc
Vanessa Colnagui Fernandes IC
Luciana Sampaio Amâncio IC

Site melhor visualizado na resolução 800 X 600
Web designer: camila_infobio@yahoo.com.br

FIGURA 3.7. Visualização dos integrantes do Grupo de Bioinformática e do Departamento de Biotecnologia

3. 4. CONSIDERAÇÕES FINAIS

Todos os dados apresentados, referentes aos resultados obtidos, estão disponibilizados na *web*, para os pesquisados do projeto da FAPESP, em sistema *web*, após o levantamento dos requisitos do usuário, da análise das tarefas estabelecendo os requisitos do sistema, *design* da interface, entre outras, buscando sempre a máxima otimização da página a ser apresentada ao usuário com o intuito de possibilitar a análise dos resultados de forma prática, rápida e segura.

Capítulo 4 – Conclusões

Com o objetivo de aplicar ferramentas de bioinformática para o arranjo e alinhamento de sequências de cDNA da aranha *Parawixia bistrata*, assim como o desenvolvimento de um sistema *web* para visualização dos dados resultantes do arranjo e alinhamento pelos pesquisadores colaboradores do projeto, todos os cromatogramas obtidos em bancadas foram analisados pelo *pipeline* desenvolvido, ou seja, desde o *software* PHRED, passando pelo CROSSMATCH e CAP3, até os alinhamentos realizados através do *software* Blast. O sistema *web*, também proposto, foi finalizado e o sistema pode ser acessado através do endereço: <http://bioinfo1.fmrp.usp.br/~ccintra>.

Os resultados, tanto do sequenciamento, quanto do alinhamento, estão sendo analisados pelos pesquisadores do Departamento de Biotecnologia da Universidade de Ribeirão Preto – UNAERP.

Por fim, quanto ao *pipeline* utilizado, há apenas a disponibilização dos resultados. A execução de cada algoritmo do *pipeline* através de linhas de comando foi objetivo por preferência dos pesquisadores envolvidos no projeto, os quais não viram necessidade de um sistema que executasse o *pipeline online*. Quanto ao sistema, uma nova versão deverá apresentar um banco de dados, para armazenamento dos dados e resultados das análises. Dessa forma, formulários para submissão dos dados deverão ser criados, resultando numa atualização do sistema *web*.

Referências bibliográficas

ALTSCHUL, S.F.; GISH, WARREN; MILLER, WEBB; MYERS, EUGENE W. (1990). Lipman, David J.; Basic Local Alignment Search Tool, Journal of Molecular Biology **215**, 403-410.

EWING, BRENT; HILLIER, LADEANA; WENDL, MICHAEL C.; GREEN, PHIL. (1997). Base Calling of Automated Sequencer Traces Using Phred - Error probabilities, Genome Res 1998; 8:186-94.

Huang, X.; Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. Genome Research, 9: 868-877.

VASSOS, T. (1998). MARKETING ESTRATÉGICO NA INTERNET. TRAD. E REV. TÉCNICA ARÃO SAPIRO. SÃO PAULO, EDITORA MAKRON.

WALL, L.; CRISTIANSEN, T.; ORWANT, J. (1987). Programming Perl, editora O'Reilly. 3ª edição.