Ferramentas de Bioinformática:

Dos Cromatogramas ao Agrupamento

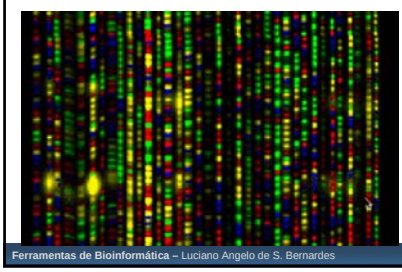
Estrutura

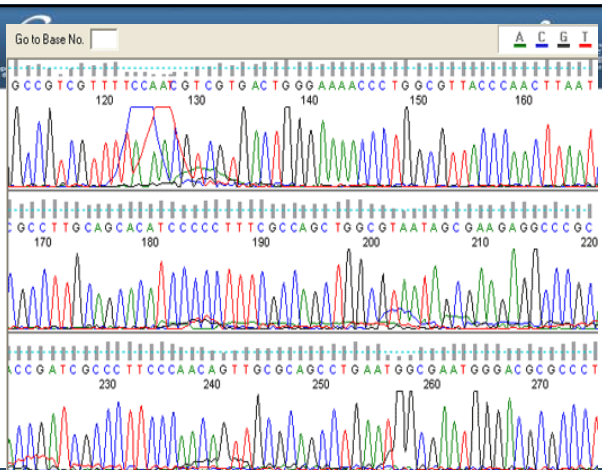
- Cromatogramas: máquinas seqüenciadores;
- Phred: sintaxe, diretórios e arquivos;
- Phd2fasta: sintaxe e arquivos;
- Cross_match: sintaxe e vector;
- Phrap: sintaxe e arquivos;
- Cap3: sintaxe e arquivos;
- Contig Viewer: visualização;
- Consed: visualização;
- Pipeline.

Seqüenciador

1986 - Sequenciador Automático de DNA (Leroy Hood)

Lloyd M. Smith et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (June 12, 1986).
T. Hunkapiller, R. J. Kaiser, B. F. Koop, L. Hood. Large-Scale and automated DNA sequence determination. *Science* **254**. 59-67 (October 4. 1991).





Phred

- **Lê dados de sinais de fluorescência vindos de sequenciadores automáticos de DNA**
  - SCF (standard chromatogram format);
  - ABI (applied biosystems);
  - ESD (MegaBACE) e
  - LI-COR.
- **Define as bases (base calling)**
- **Atribui valores de qualidade às bases**
  - valor baseado na estimação do erro calculado para cada base individualmente.
- **Gera arquivo de saída**
  - bases definidas e valores de qualidade.

Phred

RESEARCH

Base-Calling of Automated Sequencer Traces Using *Phred.* I. Accuracy Assessment

Brent Ewing,[1] LaDeana Hillier,[2] Michael C. Wendl,[2] and Phil Green[1,3]

[1]Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA; [2]Genome Sequencing Center, Washington University School of Medicine, Saint Louis, Missouri 63108 USA

8:175–185 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

Phred

RESEARCH

Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities

Brent Ewing and Phil Green[1]

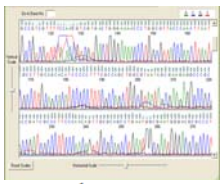
Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA

8:186–194 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

Phred

Linha de comando:

phred -id chromat_dir -pd phd_dir -sd seq_dir -qd qual_dir



```
>sequence1_name
ccccagctccggatgcaggcacaaacaacgtatggacaataggcactcttctaccaggtg
aatcaggaacaatagtcataaaagtgaaggtccctgaacagatagacctctcttttacag
aaacaggcagtgtaatagggagaaggtgttgtaatggtcaccaaagacctatccacagagc
agaagccatattcgcttatgaatagcgttaccctatcctgtacagaactggctcctaaaa
gtgccttcgctaccacaaaagtgagtggcgtgcccggaacgagcctgggtattacagaaa
gcggcagtggcctctactcttctgacgagaacttggacttacagacaaagaacagaacta
tcatccttaagaagaccacgcatgcggaatacaaacccacttcttttaacttctc
```


.seq

```
>sequence001
TGGGGNAGNNNNNAGGGGNNNGAGGAACTGAGCCCCTCGAGAGCCTTTCG
GAGACACTATAGAACATGTTTGTACAAAAAAGCAGGCTGGTACCGGTCCG
GAATTCCCGGGATCGGAAACACTTAATATATATTCAGTCTAAATGCCAGC
GAACAGCGTGAGCTGGCCAATCGGATGGAAAGGAAGCAGATGAAGGAGTT
CATGACGGTATACTCGAACATCGTCCAGCGCTGCTTCGAAGACTGCGTCA
ATGATTTCACAACGAAATCCCTCATCAGCCGCGAACAAGGCTGCGTCAAC
CGCTGCTTCGACAAGTTCATGAAGGGCTCCGAGCGCATCAACCAGCGCTT
CCAGGAGCAGAACGCTCAGATGATGCAGTCGGGCCAACTACCTGGGAAAT
AATATCATTTGAAAGAAACGGCCATTTAGACTTGGGGTTTGAAGCAGTCA
TTGTGAAGGGGGTTCTGATGGGACGATTGTCATCGCCTTTTGGTTCCCTT
ATTGCTCGTCAGGGAACTCTTTTGTTGTCAACGGCCTTTCTGCAATGATT
GTGGATTTCAATTTCCATGAAACAAGAATTTTCATTGGCAGCTAGCTGGG
```


.qual

```
>sequence001
6 6 9 9 4 0 4 4 0 0 0 0 0 4 7 9 9 4 0 0 0 4 8 6 0 0
6 6 6 6 13 13 7 7 6 6 6 6 6 8 9 8 8 9 9 6 7 6 6 8 10
12 17 24 20 24 9 7 7 11 23 29 24 19 13 8 6 6 8 16 21
25 28 30 33 37 40 37 35 29 29 29 29 29 35 37 42 42
43 42 37 40 37 37 37 37 45 45 45 51 40 40 40 40 40
40 56 56 51 45 45 45 45 45 46 44 27 22 22 22 27 31
46 45 45 45 51 51 51 56 51 51 45 45 45 45 40 40 42
56 56 56 56 45 40 40 40 40 40 40 40 40 40 40 40 40
51 51 51 56 56 44 56 40 37 35 35 35 35 35 39 51 56
56 51 51 51 51 56 56 56 56 56 56 56 56 44 42 42
46 56 51 51 51 51 51 56 56 56 56 56 56 56 51 51 51 51
51 51 51 43 45 51 45 45 40 51 51 51 51 51 51 51 51
```


.phd

```
BEGIN_SEQUENCE sequence001

BEGIN_COMMENT

CHROMAT_FILE: sequence001
ABI_THUMBPRINT:
0670522202760000040122011622 00
PHRED_VERSION: 0.020425.c
CALL_METHOD: phred
QUALITY_LEVELS: 99
TIME: Wed Jun 27 16:22:27 2007
TRACE_ARRAY_MIN_INDEX: 0
TRACE_ARRAY_MAX_INDEX: 9281
TRIM: 69 613 0.0500
TRACE_PEAK_AREA_RATIO: 0.0090
CHEM: term
DYE: big

END_COMMENT

BEGIN_DNA
t 6 9
g 6 11
g 9 13
g 9 15
```

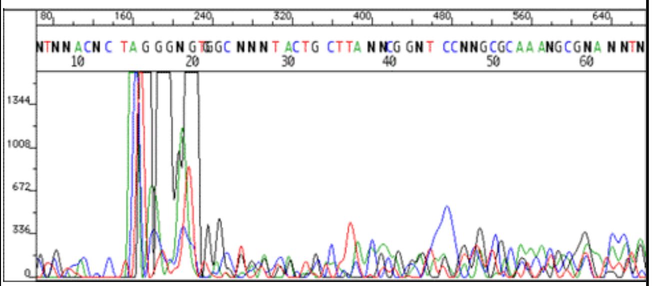
```
a 45 1424
a 45 1437
t 51 1447
a 51 1459
t 51 1471
a 56 1483
t 51 1495
a 51 1506
t 45 1518
t 45 1532
```

```
c 20 9021
c 25 9035
a 27 9047
t 15 9059
t 11 9072
c 15 9086
a 13 9094
t 12 9110
c 9 9124
a 9 9131
t 9 9149
c 10 9165
a 10 9171
t 10 9185
g 12 9201
a 7 9215
g 7 9222
g 7 9239
a 9 9253
t 9 9266
g 7 9280
END_DNA

END_SEQUENCE
```

Cromatogramas

Regiões de baixa qualidade



Fasta

Regiões de baixa qualidade

```
>Placa136_G01_13.ab1  769   87  452 ABI
CCGCGACAAGCAGTCTGTGCGGCGTGCCATAATACATGCAAGGTTATAAC
GCTGAATGAGGATCTTGCTCTACTACATGAGTTGCGAACCAGTGAGTAAC
GCGTAGGTAACCTGCCTGGTAGCGGGGGATAACTATTGGAAACGATAGCT
AATACCGCATAAGAGTAGATGTTGCATGACATTTGCTTAAAAGGTGCAAT
CACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGA
ATCTTCGGCAATGGACGGAAGTCTGACCGAGCAACGCCGCGTGAGTGAAG
AAGGTTTTCGGATCGTAAAGCTCTGTTGTAAGAGAAGAACGAGTGTGAGA
GTGGAAAGTTCACACTGTGACGGTATCTTACCAGAAAGGGACGGCTAACT
ACGTGCCAGCAGCCGCGGTAATACGTAGGTCCCGAGCGTTGTCCGGATTT
ATTGGGCGTAAAGCGAGCGCAGGCGGTTAGATAAGTCTGAAGTTTAAGGC
TGTGGCTTAACCATAGTACGCTTTGGAAACTGTTTAACTTGAGTGCAAGA
GGGGAGAGTGGAATTCATGTGTACCGGTGAAATGCGTAAATTTATGGAGG
GACACCGNTGGCGAAAGCGGCTCTCGGGCTGGTAACTGACGCTGAGGCTC
CAAAACGTGGGGAGCAAAA
```


Fasta

Regiões de baixa qualidade

```
>Placa136_G01_13.ab1  769   87  452 ABI
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 17 11 11 11 11 27 27 48 57 57
57 57 57 57 57 48 48 35 35 35 43 43 57 57 57 57 68 68 68 50 50 50
50 50 50 68 68 57 57 57 57 50 50 57 57 57 57 59 54 57 57 57 57 57
57 50 50 50 50 50 57 57 57 57 57 57 57 57 68 68 68 68 68 57  50 50
50 68 68 57 57 57 57 57 57 57 57 57 57 57 57 48 48 44 47 44 48 46
59 59 68 68 59 59 59 59 59 57 57 68 68 57 57 59 59 57 57 57 48 57
57 59 59 48 48 43 43 43 46 46 48 59 57 57 39 39 35 39 37 37 32 39
39 47 40 41 42 47 44 44 44 37 37 41 41 41 41 41 40 37 37 37 47 47
48 48 41 41 41 41 41 41 41 37 31 32 37 37 37 41 47 59 59 47 47 59
47 43 44 41 41 44 59 44 43 43 41 41 36 37 41 37 37 37 37 32 31 41
41 41 37 37 32 32 37 41 47 21 17 8 8 8 19 21 41 37 30 32 33 32 28
22 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```


Cálculo da qualidade

Fórmula para calcular os valores atribuídos às bases:

$q = -10 * \log_{10}(p)$

p = probabilidade estimada do erro para uma determinada base;
q = valor da qualidade;

Exemplo:

q = 20 ⟶ p = 10^{-2} (1 erro em 100 bases)

q = 40 ⟶ p = 10^{-4} (1 erro em 10.000 bases)

Phd2fasta

Linha de comando:

phd2fasta -id phd_dir -os seqs_fasta -oq seqs_fasta.qual

```
BEGIN_SEQUENCE sequence001

BEGIN_COMMENT

CHROMAT_FILE: sequence001
ABI_THUMBPRINT:
067052220276000004012201162200
PHRED_VERSION: 0.020425.c
CALL_METHOD: phred
QUALITY_LEVELS: 99
TIME: Wed Jun 27 16:22:27 2007
TRACE_ARRAY_MIN_INDEX: 0
TRACE_ARRAY_MAX_INDEX: 9281
TRIM: 69 613 0.0500
TRACE_PEAK_AREA_RATIO: 0.0090
CHEM: term
DYE: big

END_COMMENT

BEGIN_DNA
t 6 9
g 6 11
g 9 13
g 9 15
```

→

```
>sequence001
TGGGGNAGNNNNNAGGGGNNNGAGGAACTGAGCCCCTCGAGAGCCTTTCG
GAGACACTATAGAACATGTTTGTACAAAAAAGCAGGCTGGTACCGGTCCG
GAATTCCCGGGATCGGAAACACTTAATATATATTCAGTCTAAATGCCAGC
>sequence002
GAACAGCGTGAGCTGGCCAATCGGATGGAAAGGAAGCAGATGAAGGAGTT
CATGACGGTATACTCGAACATCGTCCAGCGCTGCTTCGAAGACTGCGTCA
ATGATTTCACAACGAAATCCCTCATCAGCCGCGAACAAGGCTGCGTCAAC
CGCTGCTTCGACAAGTTCATGAAGGGCTCCGAGCGCATCAACCAGCGCTT
>sequence003
CCAGGAGCAGAACGCTCAGATGATGCAGTCGGGCCAACTACCTGGGAAAT
AATATCATTTGAAAGAAACGGCCATTTAGACTTGGGGTTTGAAGCAGTCA
TTGTGAAGGGGGTTCTGATGGGACGATTGTCATCGCCTTTTGGTTCCCTT
ATTGCTCGTCAGGGAACTCTTTTGTTGTCAACGGCCTTTCTGCAATGATT

…
```

→

```
>sequence001
6 6 6 6 13 13 7 7 6 6 6 6 6 8 9 8 8 9 9 6 7 6 6 8
12 17 24 20 24 9 7 7 11 23 29 24 19 13 8 6 6 8 16
25 28 30 33 37 40 37 35 29 29 29 29 29 35 37 42 42
>sequence002
6 6 9 9 4 0 4 4 0 0 0 0 0 4 7 9 9 4 0 0 0 4 8 6 0
12 17 24 20 24 9 7 7 11 23 29 24 19 13 8 6 6 8 16
46 56 51 51 51 51 51 56 56 56 56 56 56 51 51 51 51
51 51 51 43 45 51 45 45 40 51 51 51 51 51 51 51 51
>sequence003
4 7 9 9 4 0 0 0 4 8 6 0 6 6 6 6 13 13 7 7 6 6 6 6
12 17 24 20 24 9 7 7 11 23 29 24 19 13 8 6 6 8 16
25 28 30 33 37 40 37 35 29 29 29 29 29 35 37 42 42
43 42 37 40 37 37 37 37 45 45 45 51 40 40 40 40 40
40 56 56 51 45 45 45 45 45 46 44 27 22 22 22 27 31

…
```


Exclusão de vetor

```
>Vetor de clonagem
CCGCGACAAGCAGTCTGTGCGGCGTGCCATAATACATGCAAGGTTATAACGCTGAA
TGAGGATCTTGCTCTACTACATGAGTTGCGAACCAGTGAGTAACGCGTAGGTAACC
TGCCTGGTAGCGGGGGATAACTATTGGAAACGATAGCTAATACCGCATAAGAGTAG
ATGTTGCATGACATTTGCTTAAAAGGTGCAATTGCACCACTACCAGATGGACCTGC
GTTGTATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCGACGATACATAGCCGA
CCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGA
GGCAGCAGTAGGGAATCTTCGGCAATGGACGGAAGTCTGACCGAGCAACGCCGCGT
GAGTGAAGAAGGTTTTCGGATCGTAAAGCTCTGTTGTAAGAGAAGAACGAGTGTGA
GAGTGGAAAGTTCACACTGTGACGGTATCTTACCAGAAAGGGACGGCTAACTACGT
GCCAGCAGCCGCGGTAATACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTA
AAGCGAGCGCAGGCGGTTAGATAAGTCTGAAGTTTAAGGCTGTGGCTTAACCATAG
TACGCTTTGGAAACTGTTTAACTTGAGTGCAAGAGGGGAGAGTGGAATTCATGTGT
ACCGGTGAAATGCGTAAATTTATGGAGGGACACCGNTGGCGAAAGCGGCTCTCGGG
CTGGTAACTGACGCTGAGGCTCCAAAACGTGGGGAGCAAAA
```

Cross_match

Linha de comando:

Cross_match seqs_fasta vetor.seq –minmatch 10 –minscore 20 –screen > screen.out

```
>Qualidade
17 11 11 11 11 27 27 48 57 57 57 57 57 57 57 48 48 35 35 35 43 43 57 57 57 57 68
>Seqüência 1
CCATAATACATGCAAGGGCGACGAGCAGACTGTCCGTCGTGCCACTTATAACGCTGAATGA
  ||||||||||||||
TAATAATACATGCAAGGTGATAATTAATTGAAAGATGCAACTGCAGCCTGCGTTGTATTAG
>Vector
```

```
>Seqüência 1
CCXXXXXXXXXXXXXXGCGACGAGCAGACTGTCCGTCGTGCCACTTATAACGCTGAATGA
```


Vetor marcado

```
>sequence001
CCXXXXXXXXXXXXXXGCGACGAGCAGACTGTCCGTCGTGCCACTTATA
GAGACACTATAGAACATGTTTGTACAAAAAAGCAGGCTGGTACCGGTCCG
GAATTCCCGGGATCGGAAACACTTAATATATATTCAGTCTAAATGCCAGC
GAACAGCGTGAGCTGGCCAATCGGATGGAAAGGAAGCAGATGAAGGAGTT
CATGACGGTATACTCGAACATCGTCCAGCGCTGCTTCGAAGACTGCGTCA
ATGATTTCACAACGAAATCCCTCATCAGCCGCGAACAAGGCTGCGTCAAC
CGCTGCTTCGACAAGTTCATGAAGGGCTCCGAGCGCATCAACCAGCGCTT
CCAGGAGCAGAACGCTCAGATGATGCAGTCGGGCCAACTACCTGGGAAAT
AATATCATTTGAAAGAAACGGCCATTTAGACTTGGGGTTTGAAGCAGTCA
TTGTGAAGGGGGTTCTGATGGGACGATTGTCATCGCCTTTTGGTTCCCTT
ATTGCTCGTCAGGGAACTCTTTTGTTGTCAACGGCCTTTCTGCAATGATT
GTGGATTTCAATTTCCATGAAACAAGAATTTTCATTGGCAGCTAGCTGGG
```


Montadores

Assemblers

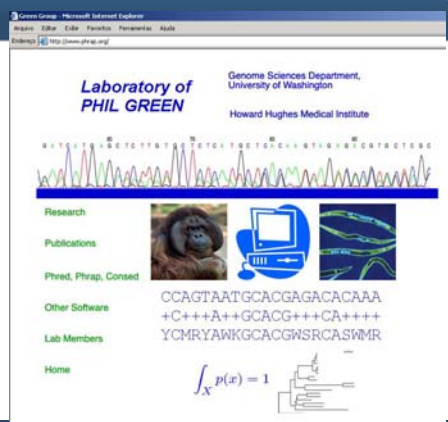
Algoritmo

- Verifica o alinhamentos de todas as seqüências;
- Escolhe as duas seqüências que tem maior sobreposição;
- Mescla as seqüências escolhidas;
- Repete os dois passos anteriores até que reste apenas um;
- O resultado é uma solução subótima para o problema.

Phrap

(Phragment Assembly Program)

- Usa “reads” completas;
- Combina informações de usuário e outras internas;
- Contigs → a partir segmentos de mais alta qualidade ao invés de criar um consenso;
- Fornece extensas informações sobre a montagens;
- Manipula grande quantidade de dados;
- Não faz mascaramento de repetições;
- Dados aberrantes são reconhecidos e tratados;
- Contigs mais longos.



Phrap

Linha de comando:

phrap seqs_fasta.screen -new_ace > phrap.out

- Programa de “montagem” das seqüências
  - [arquivo].ace
  - [arquivo].contigs
  - [arquivo].contigs.qual
  - [arquivo].log
  - [arquivo].problems
  - [arquivo].problems.qual
  - [arquivo].singlets
  - Phrap.out

CAP3

- Usa partes de alta qualidade das “reads”, para conter o erro no final das seqüências consenso;
- Reconhece informações do usuário;
- Fornece extensas informações;
- Restrições de “forward-reverse”;
- Produção de ordem parcial de contigs por restrição, tornando mais fácil e rápido sua construção;
- Menos erros internos no consenso;
- Contigs mais curtos, com menos erros

CAP3

CAP3: A DNA Sequence Assembly Program

Xiaoqiu Huang[1,2] and Anup Madan[3]

[1]Department of Computer Science, Michigan Technological University, Houghton, Michigan 49931 USA; [3]Department of Molecular Biotechnology, University of Washington, School of Medicine, Seattle, Washington 98195 USA

We describe the third generation of the CAP sequence assembly program. The CAP3 program includes a number of improvements and new features. The program has a capability to clip 5′ and 3′ low-quality regions of reads. It uses base quality values in computation of overlaps between reads, construction of multiple sequence alignments of reads, and generation of consensus sequences. The program also uses forward–reverse constraints to correct assembly errors and link contigs. Results of CAP3 on four BAC data sets are presented. The performance of CAP3 was compared with that of PHRAP on a number of BAC data sets. PHRAP often produces longer contigs than CAP3 whereas CAP3 often produces fewer errors in consensus sequences than PHRAP. It is easier to construct scaffolds with CAP3 than with PHRAP on low-pass data with forward–reverse constraints.

CAP3

Linha de comando:

cap3 seqs_fasta.screen > cap3.out

- Programa de “montagem” das seqüências
  - [arquivo].cap.ace
  - [arquivo].cap.contigs
  - [arquivo].cap.contigs.qual
  - [arquivo].cap.info
  - [arquivo].cap.singlets
  - Cap3.out

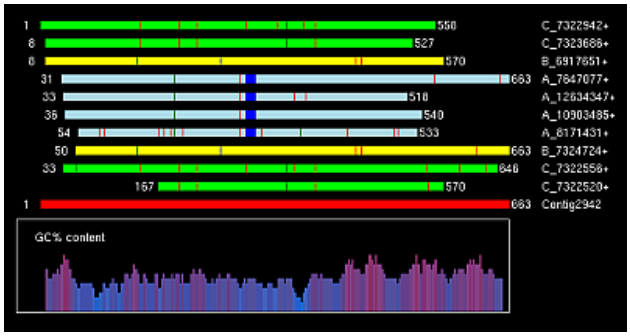
Visualizadores

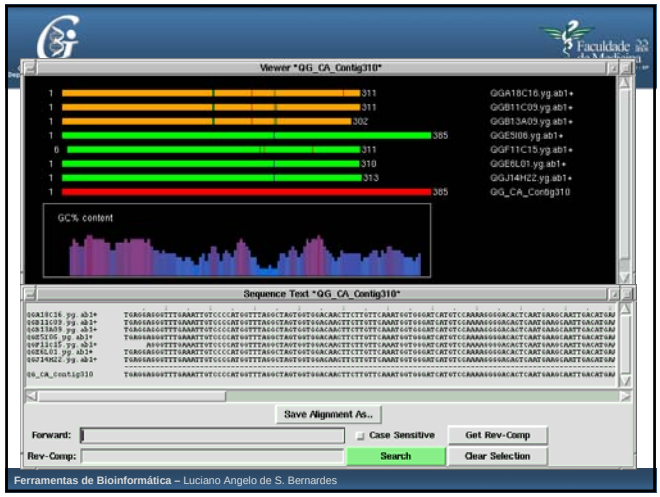
- Permitem visão geral da distribuição das seqüências;
- Permitem algumas interações com as visualizações;
- Permitem comparações e buscas;
- Permitem certas modificações;

Contig Viewer

- Desenvolvido em linguagem Python;
- Sistemas Operacionais:
  - Windows
  - MAC OS X
  - Linux/UNIX
- Cap3

Contig Viewer





CONSED

- Visualização, edição e finalização das montagens;
- Facilita a checagem da exatidão das montagens;
- Permite visualização dos valores de qualidade e cromatogramas ;
- Estrutura de diretórios: chromat_dir, edit_dir e phd_dir;
- Phrap;

Consed

RESEARCH

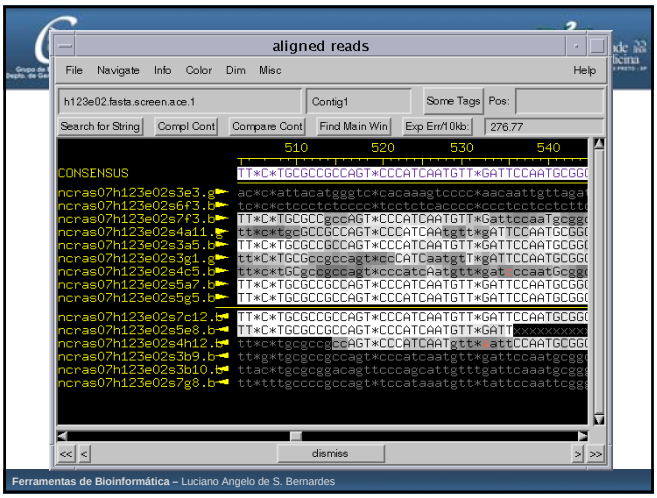
Consed: A Graphical Tool for Sequence Finishing

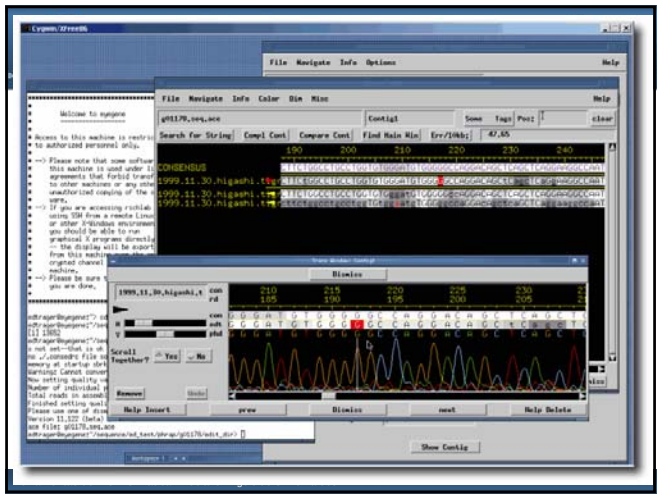
David Gordon,[2] Chris Abajian,[1] and Phil Green[2]

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA

8:195–202 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

GENOME RESEARCH ◆ 195

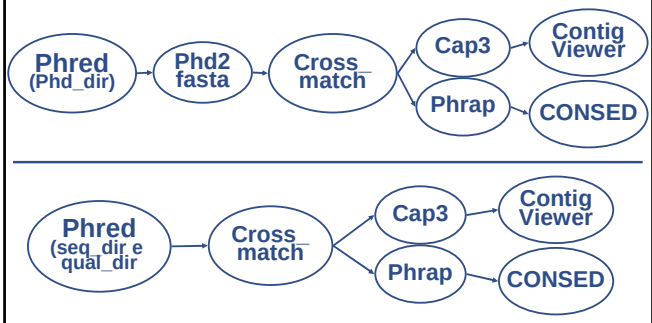




Pipeline

- Um conjunto de processos encadeados através das suas saídas padrão, de forma que a cada uma dessas é utilizada como entrada do processo seguinte.

Pipeline



PhredPhrap

- Script em linguagem perl;
- PhredPhrap roda:
  - phred;
  - phd2fasta;
  - cross_match;
  - phrap;
- Execução:
  - **phredphrap**

eGene