Montagem de Seqüências Curtas ou Muito Curtas

Candidata: Maria Angélica Lopes de Souza
Orientador: Zanoni Dias

22 de Outubro de 2008

Roteiro

Motivação

- Bioinformática
- Seqüenciar um genoma requer fragmentação e montagem
- Montagem esbarra na presença de regiões repetidas
- Seqüências curtas ou muito curtas não são bem exploradas
- Clusterização de ESTs (*Expressed Sequence Tags*)

Objetivos

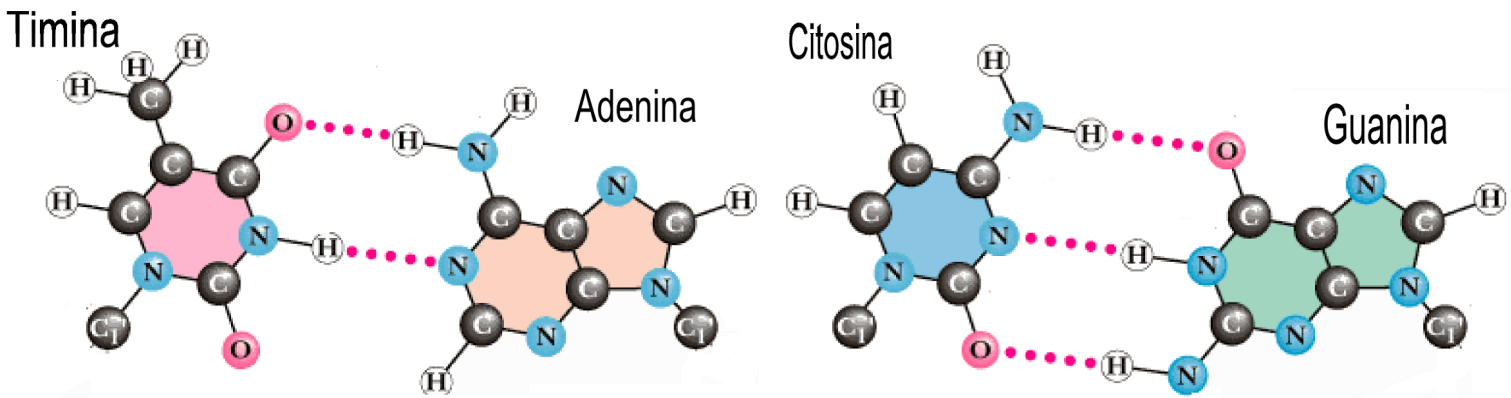
- Estudo de técnicas de montagem e clusterização
- Identificar deficiências
- Desenvolver e validar metodologias de montagem e clusterização

Genômica

- Estudo do genoma
- DNA $\longrightarrow$ dupla hélice
- Complemento reverso

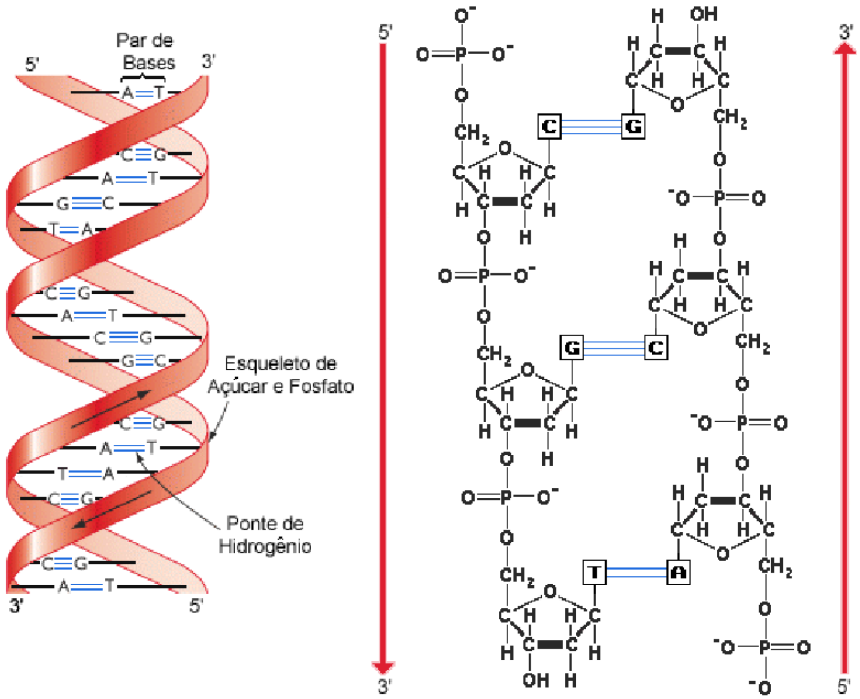
Genômica

- Estudo do genoma
- DNA → dupla hélice
- Complemento reverso



Genômica

- Estudo do genoma
- DNA → dupla hélice
- Complemento reverso

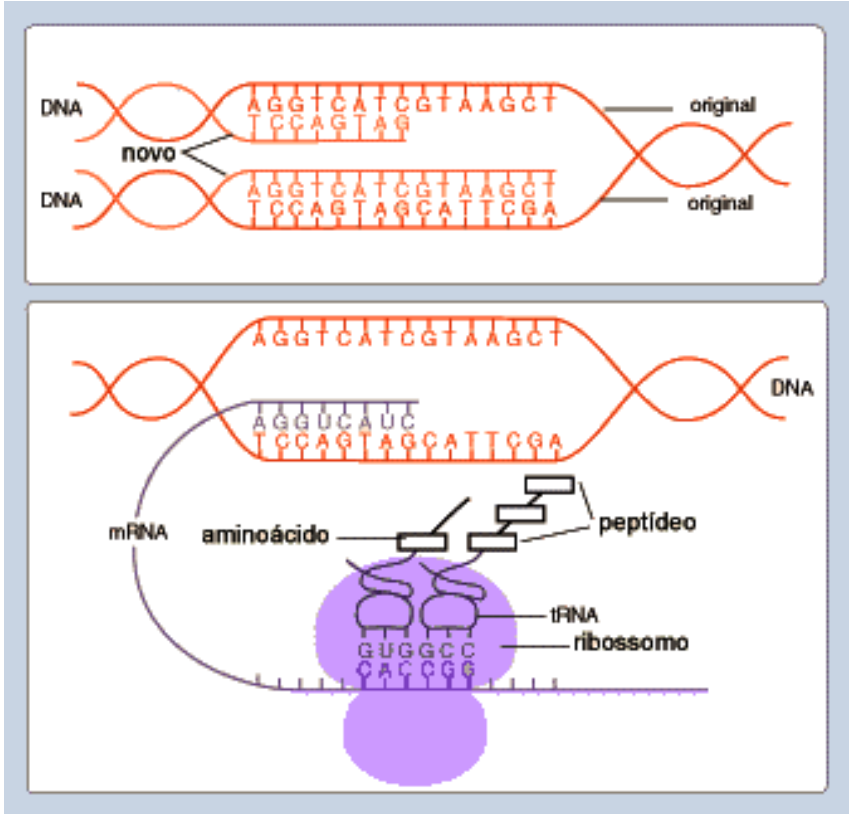


Genômica

- DNA $\longrightarrow$ transcrição $\longrightarrow$ RNA (mRNA)
- RNA tem fita simples, uracila no lugar de timina
- RNA $\longrightarrow$ tradução $\longrightarrow$ proteínas

Genômica

- DNA → transcrição → RNA (mRNA)
- RNA tem fita simples, uracila no lugar de timina
- RNA → tradução → proteínas



Genômica

- DNA → transcrição → RNA (mRNA)
- RNA tem fita simples, uracila no lugar de timina
- RNA → tradução → proteínas

Primeira base	Segunda base U	Segunda base C	Segunda base A	Segunda base G	Terceira base
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Projetos de Seqüenciamento

- Projetos Genomas → visam a obtenção de genomas completos
- O CNPq aliado ao MCT criou em 2000 a Rede Genoma Nacional
- Entrez Genome → mais de 2000 vírus, 925 bactérias, 1630 eucariotos (em 12 de setembro de 2008)

Projetos ESTs (Expressed Sequence Tags)

- Visam obter rapidamente uma boa aproximação do índice gênico de um organismo
- ESTs são obtidos pelo seqüenciamento de cDNA (gerado pela à partir do mRNA)
- Produção de bibliotecas
- dbEST $\longrightarrow$ estavam disponíveis mais de 54.495.893 seqüências públicas de ESTs de mais de 1.594 organismos diferentes (em 12 de setembro de 2008)

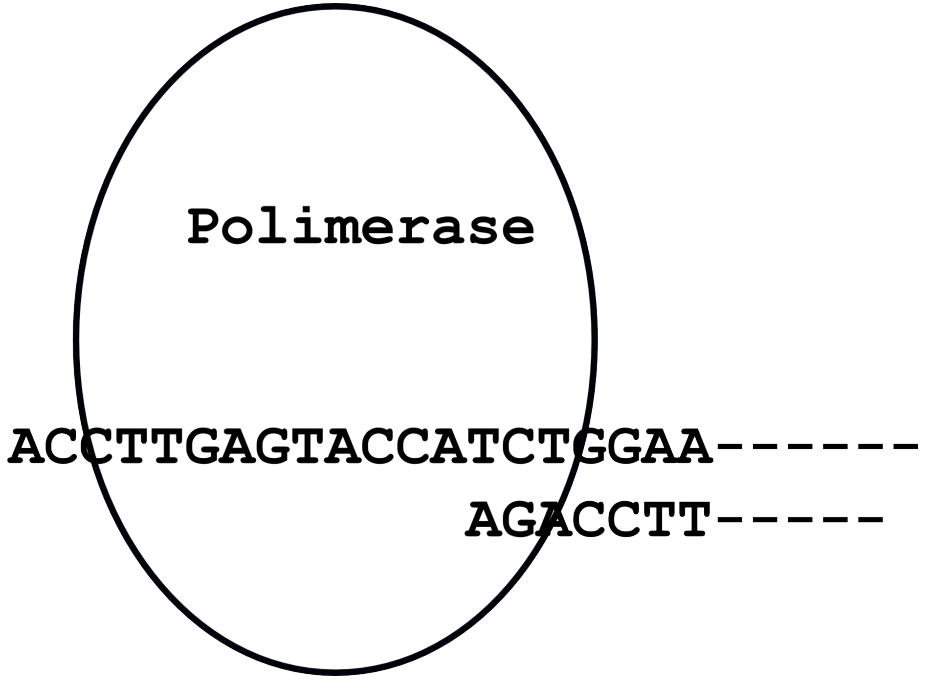
Projetos ESTs (Expressed Sequence Tags)

- Projetos de Seqüenciamento apoiados pela FAPESP: SUCEST, FORESTs, Projeto Genomas Agronômicos e Ambientais, Projeto Genoma do Câncer Humano
- Projeto Genoma Café
  - Tarefa de alto custo $\rightarrow$ utilização de ESTs
  - Concluído em 2004, resultou em um banco de dados de mais de 200 mil seqüências de cDNA.
  - Identificação de mais de 35 mil genes
  - 30% das seqüências de sua base de dados nunca foram descritas em bases de dados mundiais.

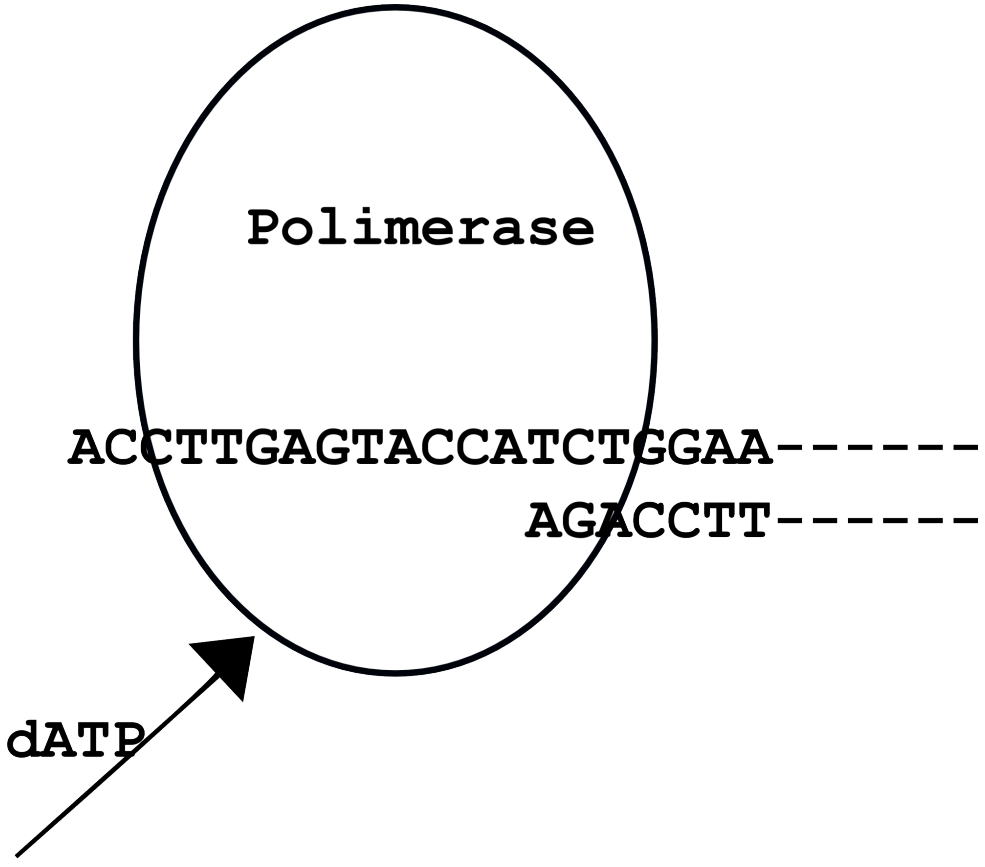
Estratégias de Seqüenciamento

- Tecnologias atuais de seqüenciamento limitadas em 1000 bases
- Fragmentação: digestão (enzimas de restrição), *shotgun* (fragmentos únicos que são clonados)
- Clonagem
- Seqüenciamento por terminação de cadeia (Sanger)
- Pirosseqüenciamento (454):
  - Síntese de DNA em tempo real
  - Liberação de pirofosfato $\rightarrow$ luz
  - Produz *reads* curtos (100-200 bases)

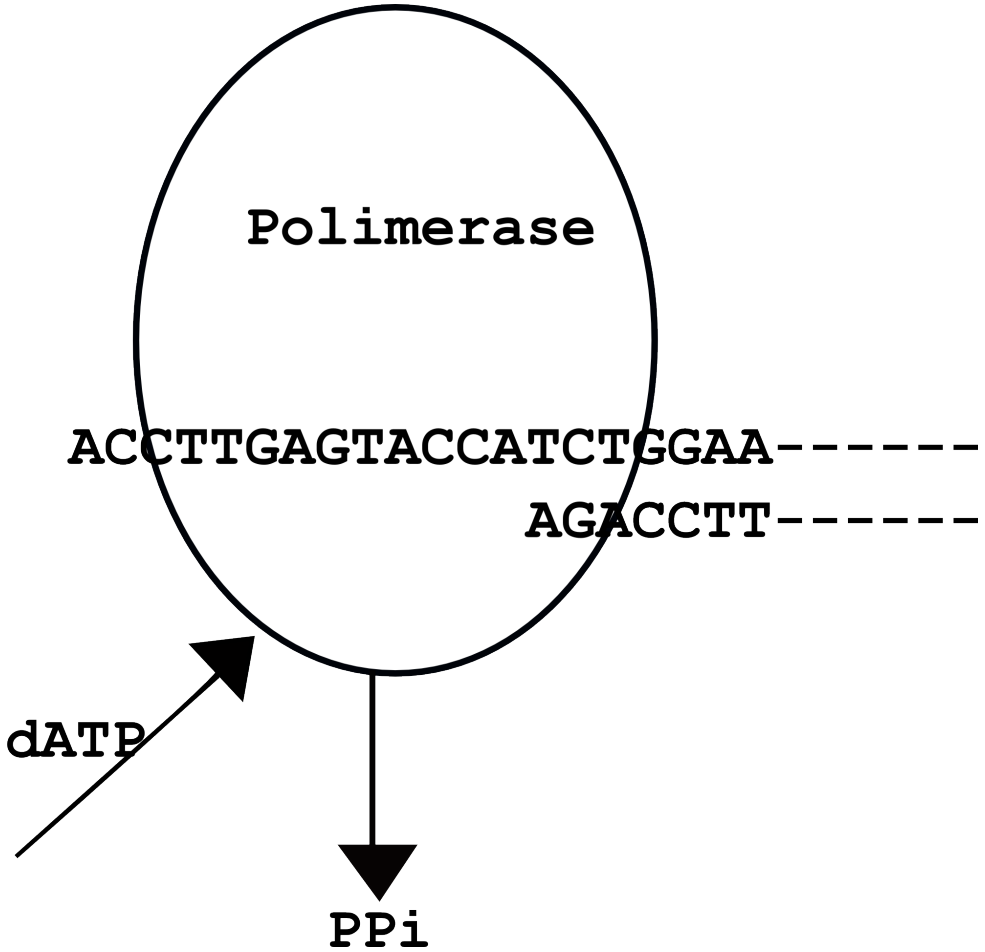
Pirosseqüenciamento



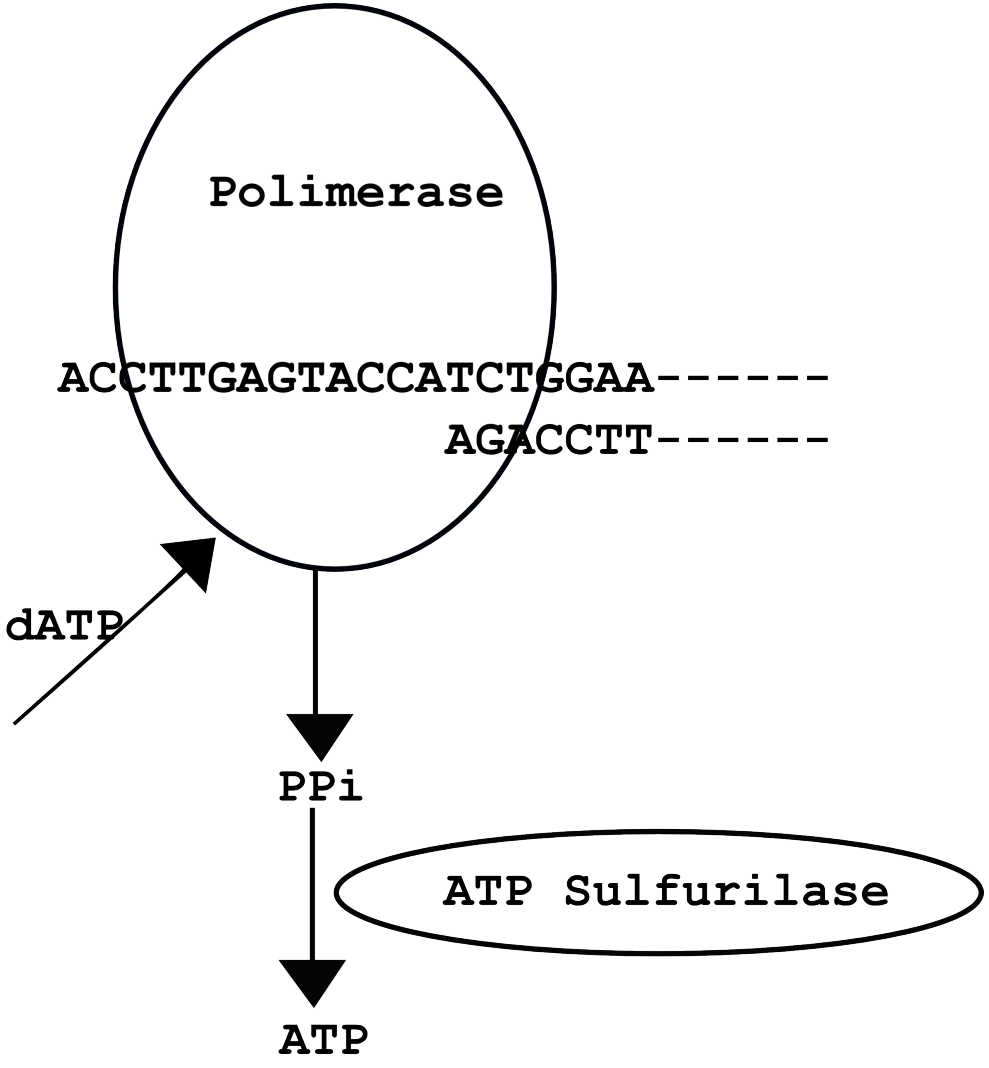
Pirosseqüenciamento



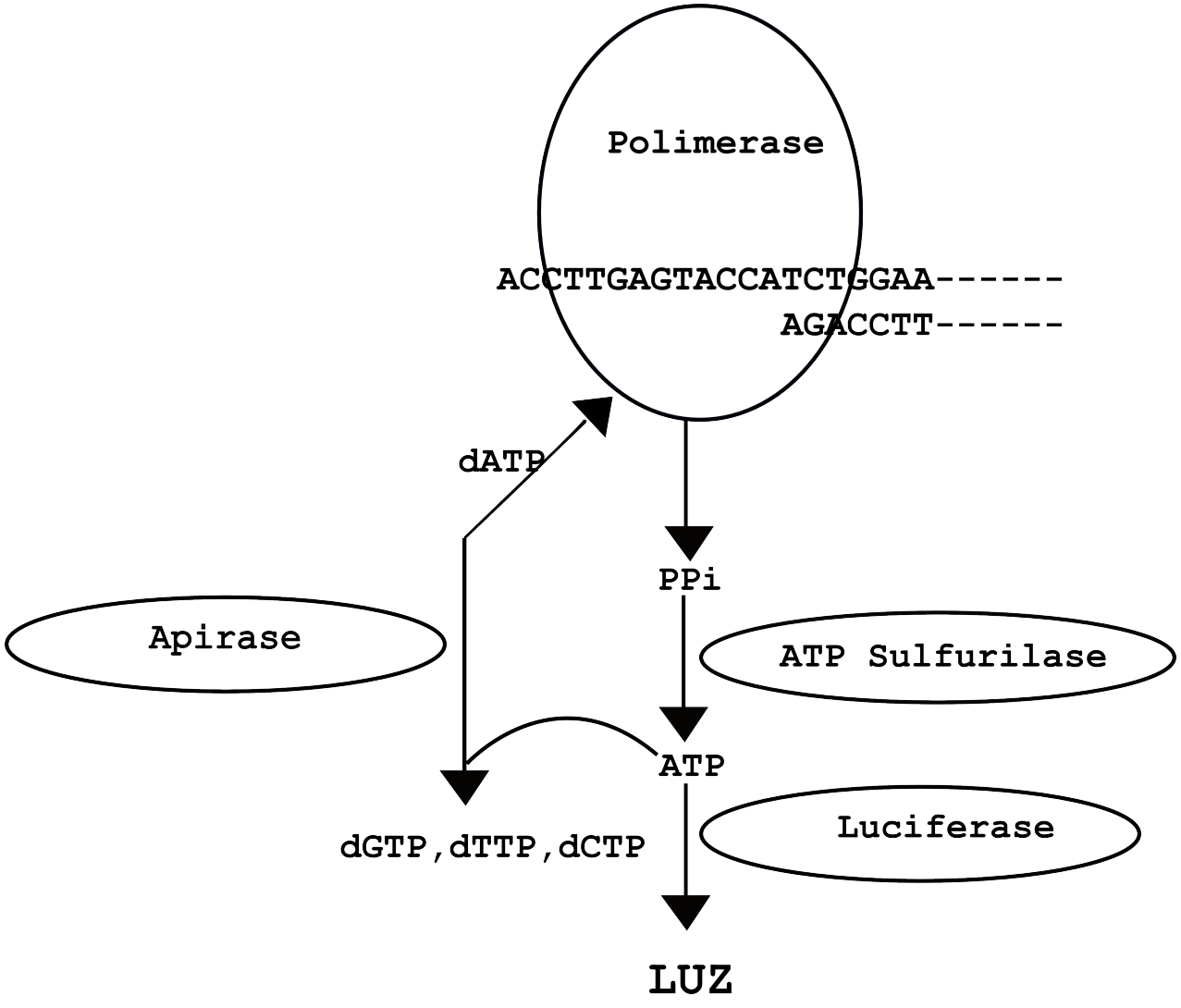
Pirosseqüenciamento



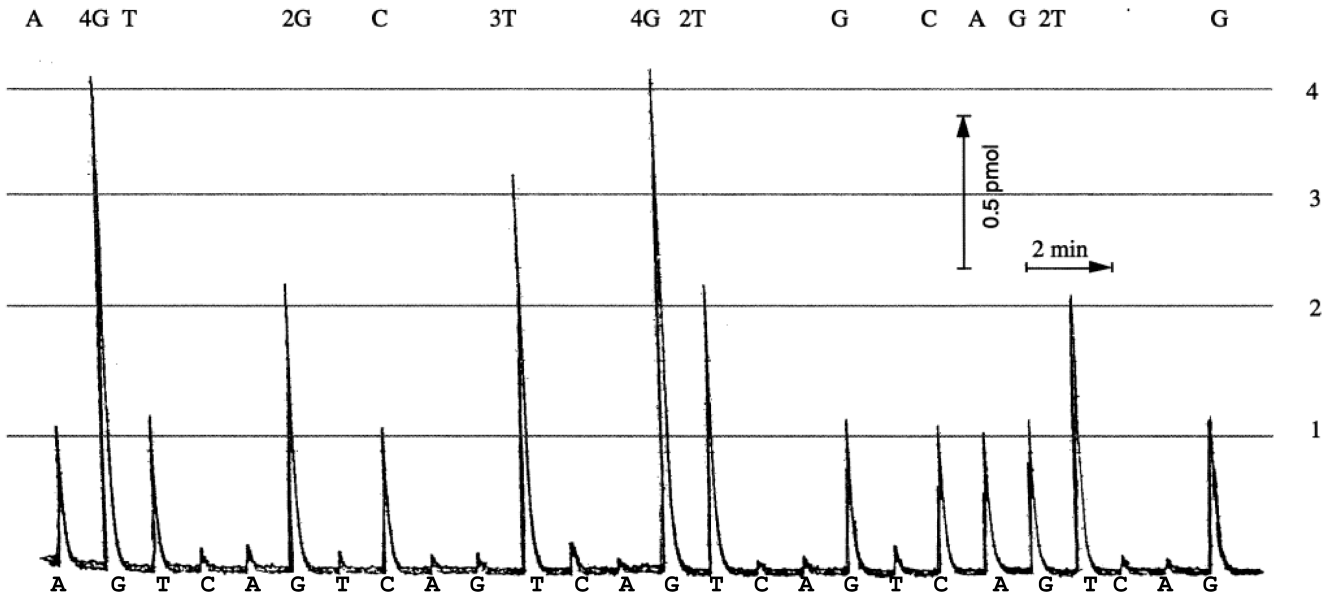
Pirosseqüenciamento



Pirosseqüenciamento



Pirosseqüenciamento



Estratégias de Seqüenciamento

- Illumina:
  - Seqüenciamento por síntese
  - Produz *reads* muito curtos
- SOLID:
  - Analisa genoma do início ao fim
  - Taxa de erro é zero
  - Produz *reads* muito curtos

Estratégias de Seqüenciamento

Seqüenciamento por:	Tamanho (em bases)	Bases/Execução	Tempo de execução
Sanger	1000	100 Kilobases	3,5 horas
454 em 2008	200-300	100 Megabases	7,5 horas
454 em 2009	350-400	400-600 Megabases	7,5 horas
Illumina em 2008	50	3 Gigabases	2-3 dias
Illumina em 2009	75	15 Gigabases	7 dias
SOLID em 2008	35	6 Gigabases	6 dias
SOLID em 2009	50	20 Gigabases	3,5 dias

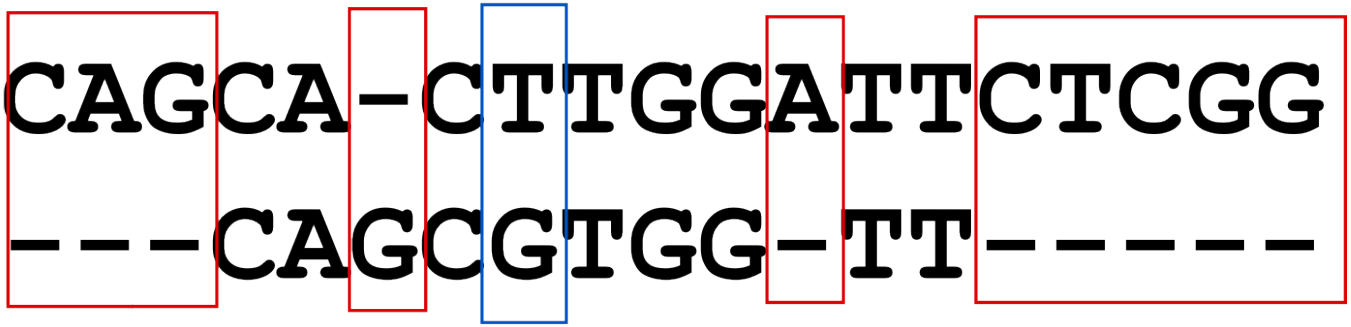
Estratégias de Montagem

- É preciso utilizar estratégia de montagem que atinja alta cobertura da seqüência original
- Genoma completo
- Alinhamento - Similaridade

Estratégias de Montagem

- É preciso utilizar estratégia de montagem que atinja alta cobertura da seqüência original
- Genoma completo
- Alinhamento - Similaridade

```
CAGCACTTGGATTCTCGG
CAGCGTGGTT
```



Estratégias de Montagem

- É preciso utilizar estratégia de montagem que atinja alta cobertura da seqüência original
- Genoma completo
- Alinhamento - Similaridade
- Sobreposição-Layout-Consenso

Estratégias de Montagem

- É preciso utilizar estratégia de montagem que atinja alta cobertura da seqüência original
- Genoma completo
- Alinhamento - Similaridade
- Sobreposição-Layout-Consenso

ACTGATGGCCTAATACGATAG

CGATAGTCTGAGACAGAGTCA

ACTGATGGCCTAATACGATAG
CGATAGTCTGAGACAGAGTCA

Estratégias de Montagem

- É preciso utilizar estratégia de montagem que atinja alta cobertura da seqüência original
- Genoma completo
- Alinhamento - Similaridade
- Sobreposição-Layout-Consenso
- Formação de *contigs/singletons*

Estratégias de Montagem

- É preciso utilizar estratégia de montagem que atinja alta cobertura da seqüência original
- Genoma completo
- Alinhamento - Similaridade
- Sobreposição-Layout-Consenso
- Formação de *contigs/singletons*

Problemas

- Erros nos *reads*
- Regiões repetidas

Estratégias de Montagem

Problemas

- Erros nos *reads*
- Regiões repetidas

Estratégias de Montagem

Problemas

- Erros nos *reads*
- Regiões repetidas

Estratégias de Montagem

Problemas

- Erros nos *reads*
- Regiões repetidas

Estratégias de Montagem

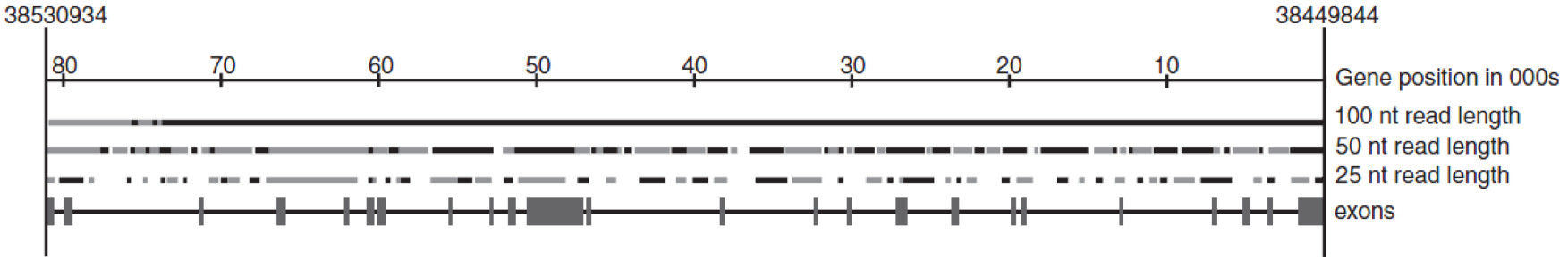
Problemas

- Erros nos *reads*
- Regiões repetidas
- Mais complexo trabalhar com *reads* curtos ou muito curtos.

Estratégias de Montagem

Problemas

- Erros nos *reads*
- Regiões repetidas
- Mais complexo trabalhar com *reads* curtos ou muito curtos.



Trecho do cromossomo 17 humano

Estratégias de Montagem

Problemas

- Erros nos *reads*
- Regiões repetidas
- Mais complexo trabalhar com *reads* curtos ou muito curtos.
- Não exploram corretamente ESTs

Clusterização de ESTs

- Clusterização: agrupamento de seqüências semelhantes em um grupo chamado *cluster*
- Cálculo de similaridade
- É preciso saber quais genes ocorrem em um organismo e também com que freqüência ocorrem (nem todos os genes ocorrem com a mesma freqüência)
- Obtenção da seqüência consenso, formada pela análise das seqüências que formam o *cluster* e que é aceita como a seqüência com maior probabilidade de ser a do gene existente no organismo
- ESTs apresentam particularidades como SNP, domínios conservados e famílias multigênicas.

Clusterização de ESTs

Problemas

- ESTs possuem problemas com seqüências redundantes, regiões de baixa qualidade e quimeras
- Métodos tradicionais de montagem não exploram clusterização de ESTs

Clusterização de ESTs

Problemas

- ESTs possuem problemas com seqüências redundantes, regiões de baixa qualidade e quimeras
- Métodos tradicionais de montagem não exploram clusterização de ESTs

1 CTGCTTTAAGGGTCGTTAATTGACGACTCTTGATATTTACTTAGTTTGAGTT
2 GAGCACTGCTTTAAGGGTCGTTAATTGACGACTCTTGATATTTACTTAGTTT
3 GAAAAGGATCTTTCTGATTCTCGAAGAATGAGGGGCAAGGGGATTGATCGA
4 CGTTAATTGACGACTCTTGATATTTACTTAGTTTGAGTTATGGACGA
5 CAAGTAGCTTTGGTAATCTTCTCAGTACAACCGACCCACCGTTTCAATCTTTGTA

Clusterização de ESTs

Problemas

- ESTs possuem problemas com seqüências redundantes, regiões de baixa qualidade e quimeras
- Métodos tradicionais de montagem não exploram clusterização de ESTs

```
1 CTGCTTTAAGGGTCGTTAATTGACGACTCTTGATATTTACTTAGTTTGAGTT
2 GAGCACTGCTTTAAGGGTCGTTAATTGACGACTCTTGATATTTACTTAGTTT
3 GAAAAGGATCTTTCTGATTCTCGAAGAATGAGGGGCAAGGGGATTGATCGA
4 CGTTAATTGACGACTCTTGATATTTACTTAGTTTGAGTTATGGACGA
5 CAAGTAGCTTTGGTAATCTTCTCAGTACAACCGACCCACCGTTTCAATCTTTGTA

  Cluster
C GAGCACTGCTTTAAGGGTCGTTAATTGACGACTCTTGATATTTACTAAGTTTGAGTTATGGACGA
2 GAGCACTGCTTTAAGGGTCGTTAATTGACGACTCTTGATATTTACTTAGTTT
1       CTGCTTTAAGGGTCGTTAATTGACGACTCTTGATATTTACTTAGTTTGAGTT
4                       CGTTAATTGACGACTCTTGATATTTACTTAGTTTGAGTTATGGACGA
```

Clusterização de ESTs

Problemas

- ESTs possuem problemas com seqüências redundantes, regiões de baixa qualidade e quimeras
- Métodos tradicionais de montagem não exploram clusterização de ESTs

```
1 CTGCTTTAAGGGTCGTTAATTGACGACTCTTGATATTTACTTAGTTTGAGTT
2 GAGCACTGCTTTAAGGGTCGTTAATTGACGACTCTTGATATTTACTTAGTTT
3 GAAAAGGATCTTTCTGATTCTCGAAGAATGAGGGGCAAGGGGATTGATCGA
4 CGTTAATTGACGACTCTTGATATTTACTTAGTTTGAGTTATGGACGA
5 CAAGTAGCTTTGGTAATCTTCTCAGTACAACCGACCCACCGTTTCAATCTTTGTA

Cluster-Singleton
3 GAAAAGGATCTTTCTGATTCTCGAAGAATGAGGGGCAAGGGGATTGATCGA
Cluster-Singleton
5 CAAGTAGCTTTGGTAATCTTCTCAGTACAACCGACCCACCGTTTCAATCTTTGTA
```

Programas para Montagem e Clusterização

- Phrap
- CAP3
- Suporte para *reads* curtos:
  - Celera Assembler
  - AMOS
  - MIRA
  - VCAKE
  - VELVET
  - SSAKE
  - SHARGS
  - EULER-SR

Materiais e Métodos

- ESTs: O Projeto FINEP (0886/2007) gerará mais de um milhão de seqüências curtas de planta *Coffea arabica L. cv.*
- Parceria com Embrapa Recursos Genéticos e Biotecnologia (Cenargen) e projeto universal CNPq (486303/2006).
- Linguagem de programação Python $\rightarrow$ biblioteca Biopython.
- Testes com alguns dos softwares existentes
- Identificação de deficiências

Materiais e Métodos

- Construção de metodogia:
  - Preparação das seqüências $\longrightarrow$ pré-processamento
  - Desenvolvimento de ferramentas que supram as deficiências encontradas
  - Validação da metodologia desenvolvida
  - Comparação com alguns dos softwares existentes
  - Avaliação qualitativa, por biólogos
  - Avaliação quantitativa: consistência interna, consistência externa, discrepância de bases, etc.

Cronograma

	2008				2009												2010	
	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	X	X	X	X														
2	X																	
3		X																
4	X	X	X	X														
5					X	X	X											
6							X	X	X									

1. Disciplinas obrigatórias do programa de mestrado.
2. Escrita do projeto de mestrado.
3. Exame de qualificação.

Cronograma

	2008				2009												2010	
	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	X	X	X	X														
2	X																	
3		X																
4	X	X	X	X														
5					X	X	X											
6							X	X	X									

4. Estudo da linguagem de programação Python e da biblioteca BioPython, específica para bioinformática.
5. Testes de algoritmos existentes.
6. Identificação de problemas.

Cronograma

	2008				2009												2010	
	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
7									X	X	X							
8											X	X	X					
9													X	X	X			
10											X	X	X	X	X	X		
11																	X	
12																		X

7. Construção da metodologia.
8. Implementação.
9. Testes.

Cronograma

	2008				2009												2010	
	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
7									X	X	X							
8											X	X	X					
9													X	X	X			
10											X	X	X	X	X	X		
11																	X	
12																		X

10 Escrita da dissertação.

11 Revisão final do texto da dissertação.

12 Defesa da dissertação.