Alinhamento de seqüências biológicas em ambiente distribuído

Edson José Savi Júnior

Orientador: Paulo de Tarso Mendes Luna

www.furb.br

Roteiro da Apresentação

- Introdução
  - Objetivo do Trabalho
- Fundamentação Teórica
- Desenvolvimento do Trabalho
  - Especificação
  - Implementação
  - Operacionalidade
  - Resultados e discussão
- Conclusão
  - Extensões

Introdução

- Alinhamento de seqüências permite inferir sobre as propriedades de determinado gene

- A comparação de seqüências biológicas necessita de computadores de alto desempenho
  - Alto poder de processamento
  - Uso de memória considerável

Objetivos

- Implementar rotinas numa ferramenta de alinhamento de seqüências biológicas open source para realizar o processamento em ambiente distribuído
- Armazenar o resultado do alinhamento entre seqüências em um banco de dados (arquivo simples)
- Demonstrar o grau de similaridade entre as seqüências na ferramenta

Bioinformática

- Foca-se na criação e no uso de ferramentas para a organização e a análise de dados biológicos através dos computadores.

- Fusão da informática, matemática e ciências biológicas

Alinhamento de seqüências biológicas

- Processo de comparação de seqüências biológicas
- Operação primitiva considerada mais importante na área de biologia computacional
- Consiste em encontrar trechos semelhantes nas seqüências de entrada
- Tipos de alinhamentos
  - Em pares
  - Alinhamentos múltiplos
  - Global (Needleman-Wunsch)
  - Local (Smith-Waterman)

Alinhamento em par

- Duas seqüências são combinadas aleatoriamente
- A qualidade da combinação é avaliada e pontuada
- Uma seqüência é movida sobre a outra e a combinação é pontuada novamente

Resultado

- Similaridade: melhor pontuação dentre os possíveis alinhamentos
- Identidade: presença do mesmo ácido nucléico ou aminoácido na mesma posição nas seqüências

Alinhamento de seqüências biológicas

Exemplo:

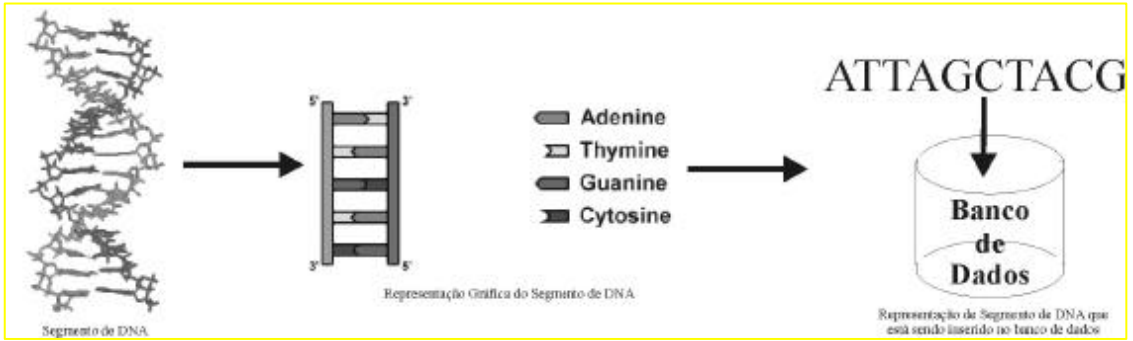
```
Seqüências AAAC e AGC

      alinhamento ótimo entre (AAA )C
                              ( AG )C

      alinhamento ótimo entre (AAAC)-
                              ( AG )C

      alinhamento ótimo entre (AAA )C
                              (AGC )-
```

Algoritmo de Smith-Waterman

- Alinhamento local de seqüências
- Utiliza a técnica de programação dinâmica
- Resultado é o alinhamento ótimo
- Dividido em três fases:
  - inicialização de matrizes (definido valores para *match*, *mismatch* e *gaps*)
  - preenchimento da matriz de alinhamento (uso da função MAX)
  - traceback
- Tempo e o espaço requerido é de O(mn), onde m e n são o comprimento das seqüências

Bancos de dados biológicos

- Armazenam informações de caráter biológico
- Banco de dados de arquivos simples ou relacionais
  - Comum o uso de banco de dados de arquivos simples



Sistemas Distribuídos

- Coleção de computadores independentes que se apresentam ao usuário como um sistema único e consistente

- Uma grande tarefa computacional pode ser dividida em pequenas tarefas que são distribuídas ao redor das estações, como se fosse um supercomputador massivamente paralelo.

JAligner

- Ferramenta open source
- Alinhamento de seqüências em par
- Uso do algoritmo de Smith-Waterman
- Utiliza o modelo de penalização de gaps
- Disponível em pacotes de ferramentas para bioinformática, como por exemplo: BioWeka, RDPquery, STRAP e EMBOSS

JPVM - Java parallel virtual machine

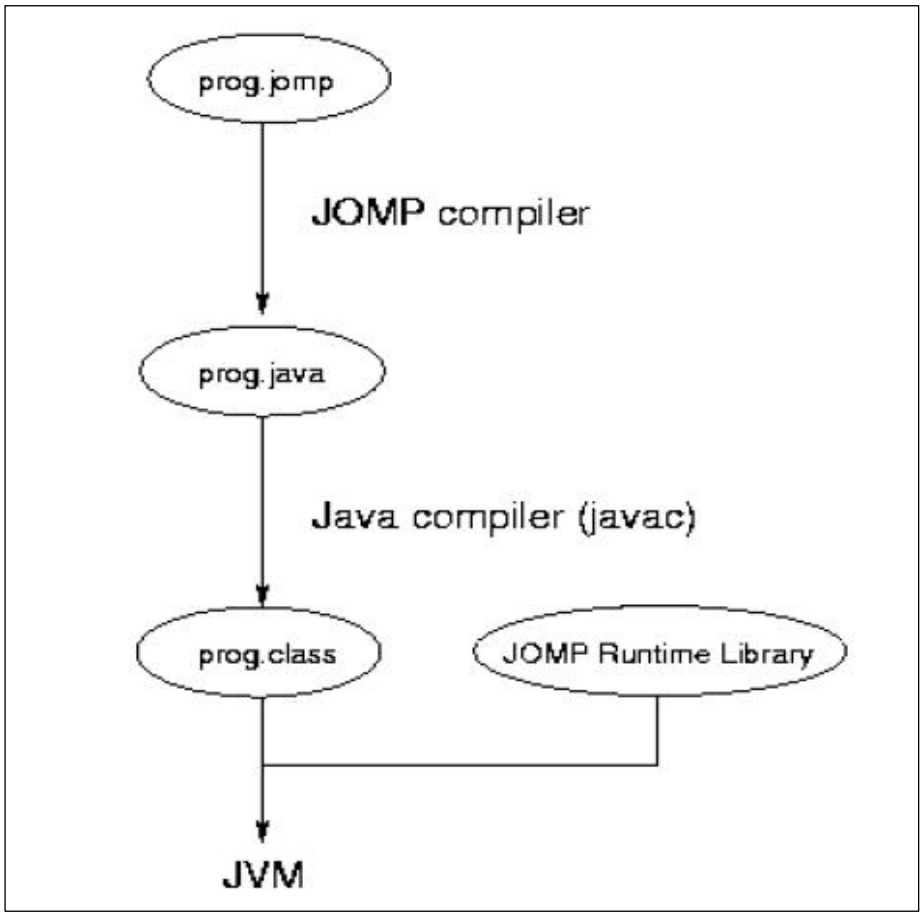
- Implementado inteiramente em Java
- Segue a estrutura básica do PVM
- Comunicação entre vários computadores através de troca de mensagens
- Alta portabilidade

JOMP

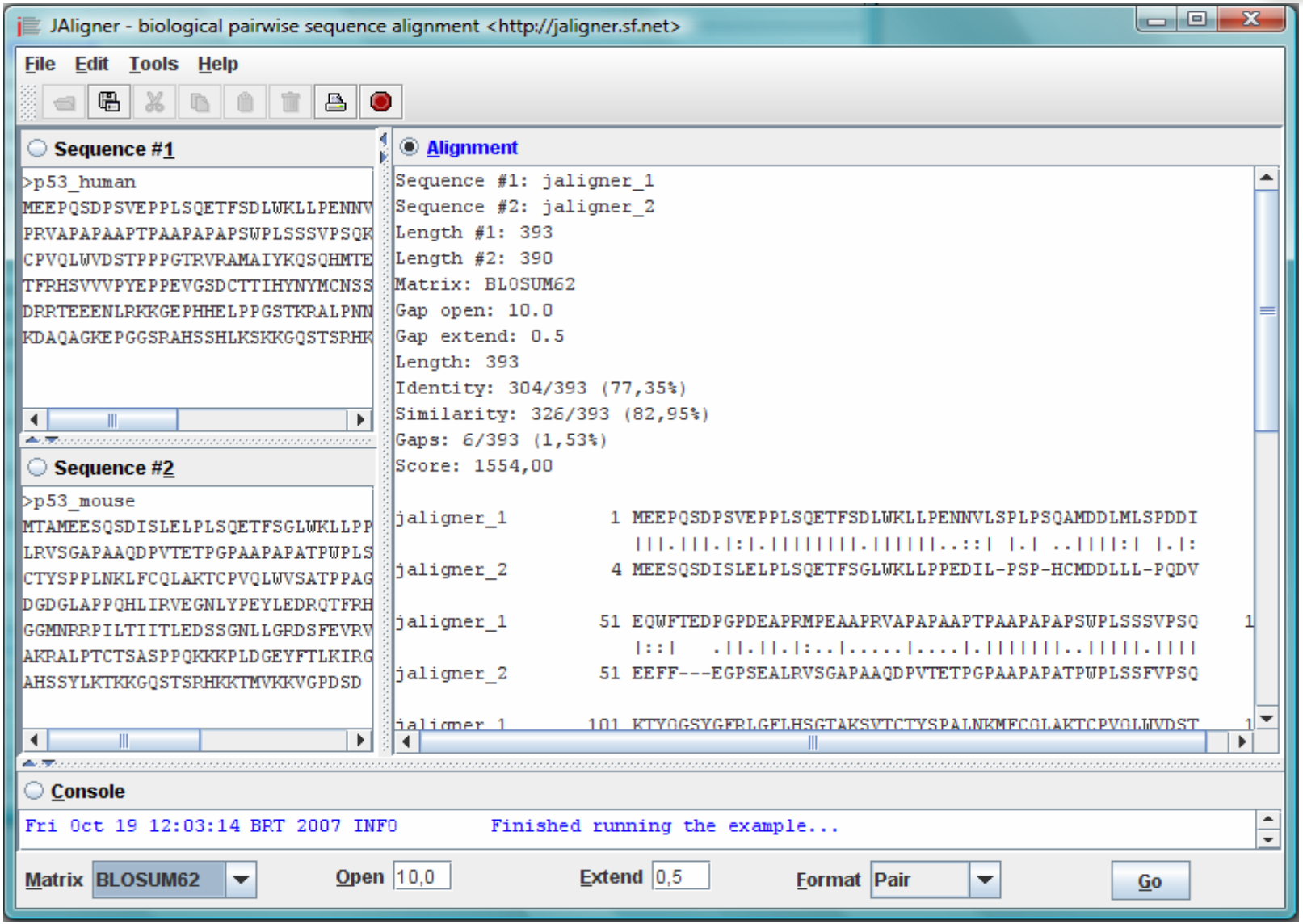
- Utiliza o modelo de paralelização de tarefas em tempo de execução
- Uso de diretivas para paralelizar programas
- Consiste em um conjunto de diretivas e bibliotecas usadas em tempo de execução
- Cria um time de threads ao encontrar uma região paralela
- Quantidade de threads pode ser definida previamente

JOMP

- Fluxograma de uso



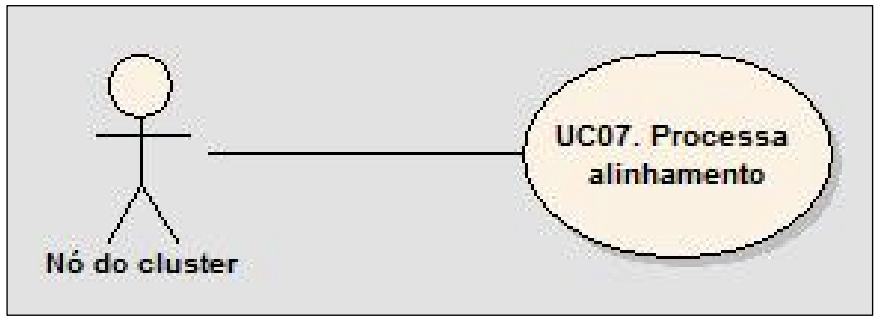
JAligner


JAligner - biological pairwise sequence alignment <http://jaligner.sf.net>
File Edit Tools Help
Sequence #1
>p53_human
MEEPQSDPSVEPPLSQETFSDLWKLLPENNV
PRVAPAPAAPTPAAPAPAPSWPLSSSVPSQK
CPVQLWVDSTPPPGTRVRAMAIYKQSQHMTE
TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSS
DRRTEEENLRKKGEPHHELPPGSTKRALPNN
KDAQAGKEPGGSRAHSSHLKSKKGQSTSRHK
Sequence #2
>p53_mouse
MTAMEESQSDISLELPLSQETFSGLWKLLPP
LRVSGAPAAQDPVTETPGPAAPAPATPWPLS
CTYSPPLNKLFCQLAKTCPVQLWVSATPPAG
DGDGLAPPQHLIRVEGNLYPEYLEDRQTFRH
GGMNRRPILTIITLEDSSGNLLGRDSFEVRV
AKPRALPTCTSASPPQKKKPLDGEYFTLKIRG
AHSSYLKTKKGQSTSRHKKTMVKKVGPDSD
Alignment
Sequence #1: jaligner_1
Sequence #2: jaligner_2
Length #1: 393
Length #2: 390
Matrix: BLOSUM62
Gap open: 10.0
Gap extend: 0.5
Length: 393
Identity: 304/393 (77,35%)
Similarity: 326/393 (82,95%)
Gaps: 6/393 (1,53%)
Score: 1554,00
jaligner_1 1 MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDI
jaligner_2 4 MEESQSDISLELPLSQETFSGLWKLLPPEDIL-PSP-HCMDDLLL-PQDV
jaligner_1 51 EQWFTEDPGPDEAPPMPEAAPRVAPAPAAPTPAAPAPAPSWPLSSSVPSQ 1
jaligner_2 51 EEFF---EGPSEALRVSGAPAAQDPVTETPGPAAPAPATPWPLSSFVPSQ
jaligner_1 101 KTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDST 1
Console
Fri Oct 19 12:03:14 BRT 2007 INFO Finished running the example...
Matrix BLOSUM62
Open 10,0
Extend 0,5
Format Pair
Go

Requisitos principais

- Carregar seqüências biológicas, armazenando-as em uma fila para processamento
- Realizar o alinhamento das seqüências biológicas com as demais seqüências armazenadas na fila, em ambiente distribuído
- Apresentar o grau de similaridade, identidade e pontuação, resultante do alinhamento entre as seqüências biológicas
- Armazenar o resultado dos alinhamentos em um banco de dados de arquivos simples

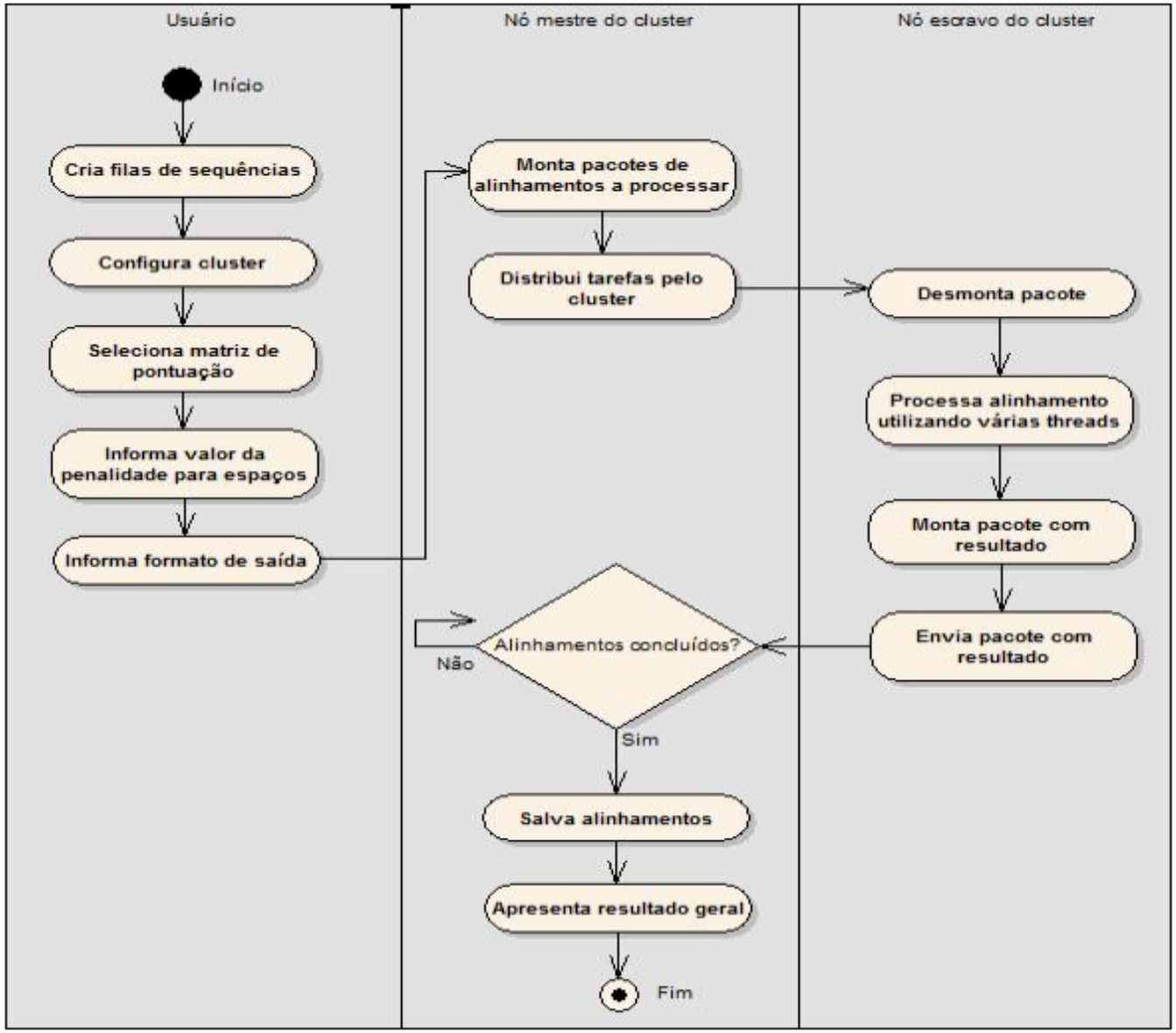
Caso de uso – Processa alinhamento



UC07: Processa alinhamento	
Resumo	Nó do cluster realiza processamento do alinhamento de seqüências biológicas.
Seqüência de ações	1. Aguarda o recebimento do pacote contendo seqüências biológicas para alinhar. 2. Realiza o alinhamento através do algoritmo de Smith-Waterman, dividindo o processamento conforme o número de *threads* pré-definidos na ferramenta JAligner. 3. Monta pacote contendo resultado do alinhamento das seqüências. 4. Envia resultado para nó do cluster que disparou o processamento do alinhamento.

Especificação

Diagrama de atividades



Técnicas e ferramentas utilizadas

- Ferramentas
  - Eclipse
  - Netbeans
  - Enterprise Architect
- Técnicas de computação paralela
  - JPVM
  - JOMP

Implementações no JAligner

- Fila de seqüências
  - Carregar seqüência
  - Salvar fila de seqüência
  - Carregar fila de seqüência

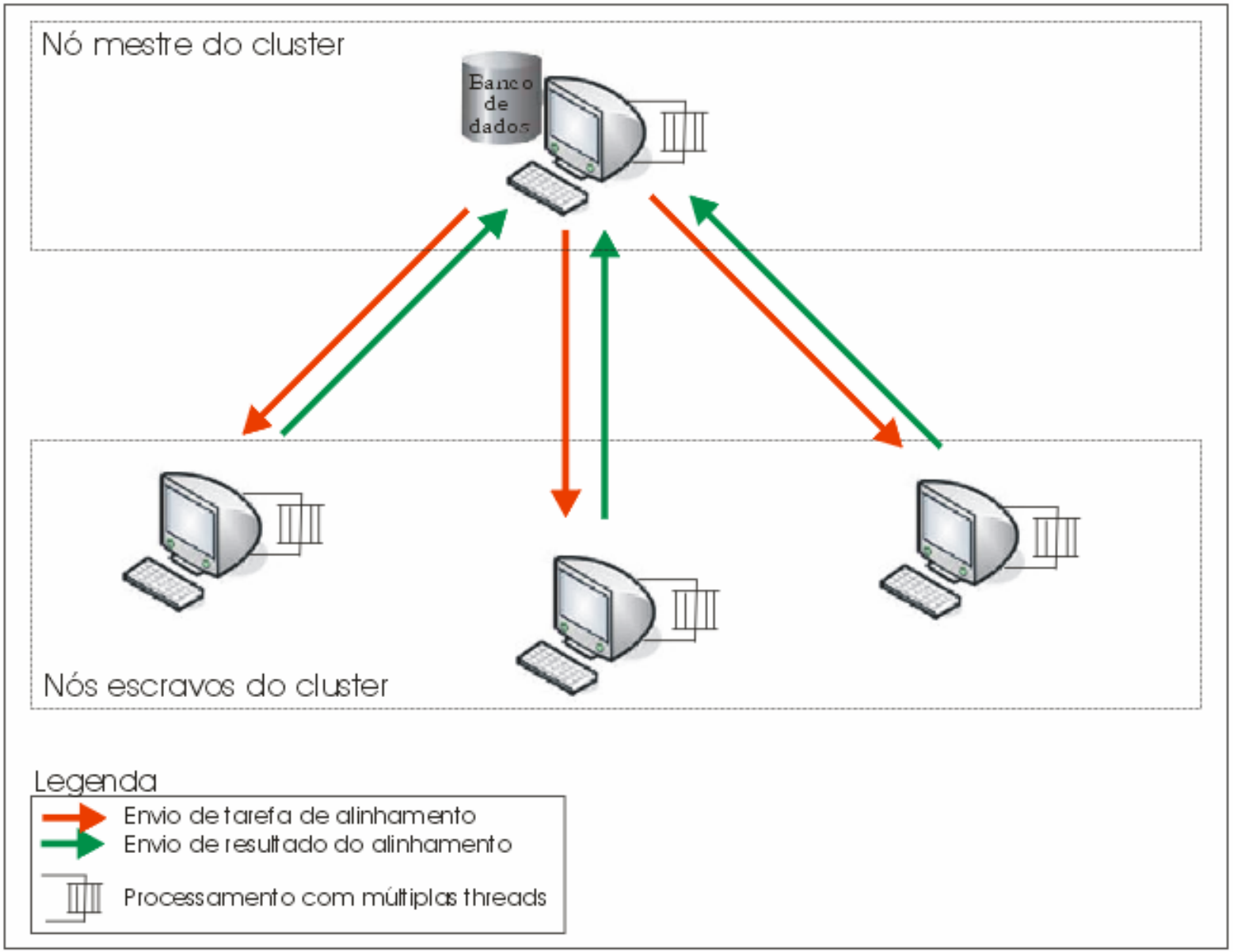
- Distribuição das seqüências a alinhar pelos computadores que formam o cluster através do JPVM.

- Processamento paralelo da tarefa de alinhamento, em cada computador do cluster, através do JOMP.

- Armazenamento dos alinhamentos gerados em um banco de dados de arquivos simples.

- Visualização da similaridade, identidade e pontuação dos alinhamentos gerados

Implementação

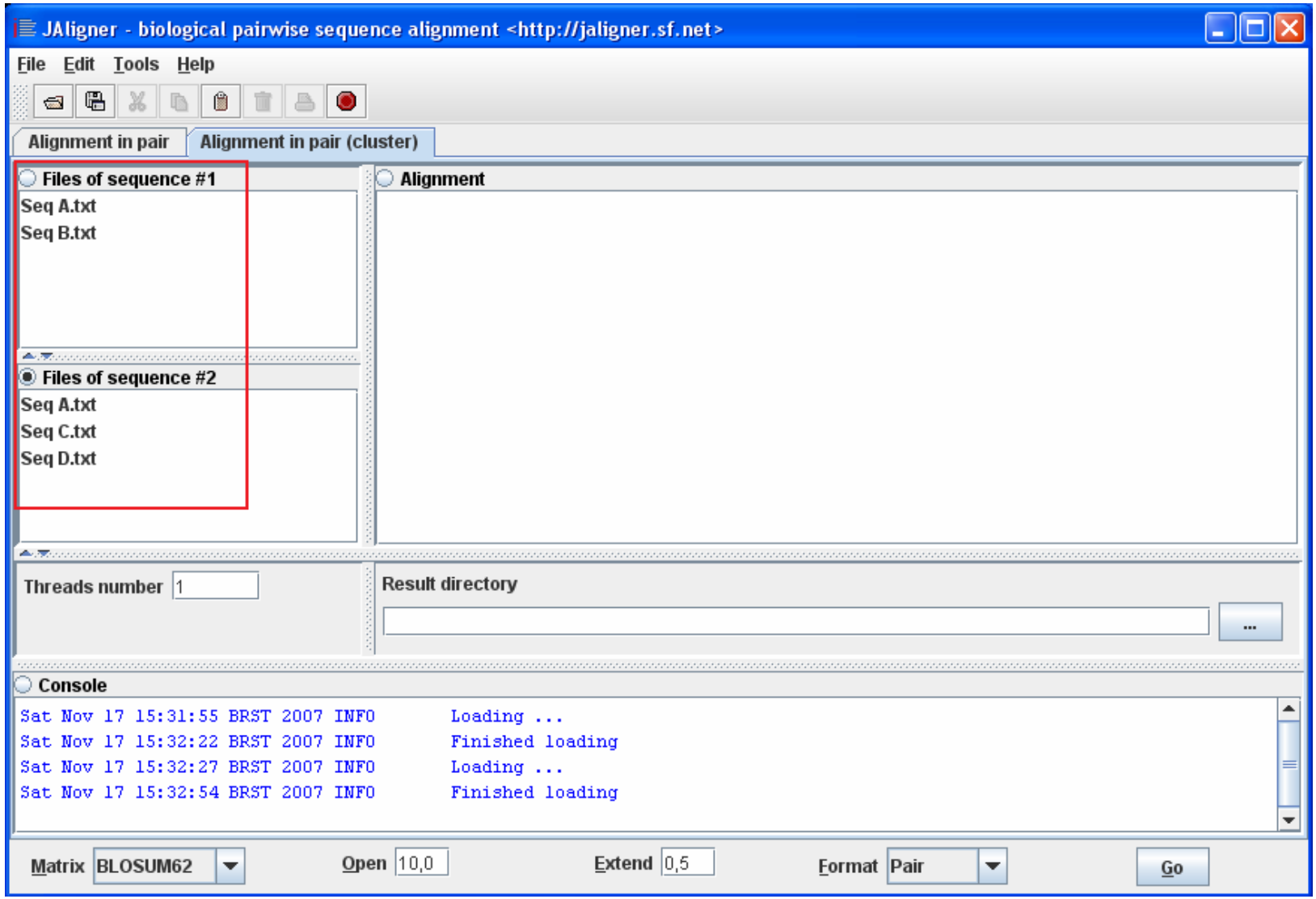


Operacionalidade

- Configuração das variáveis de ambiente do sistema operacional
  - java
  - classpath
- Configuração do cluster
  - jpvmDaemon
  - jpvmConsole
- Adicionar computadores que fazem parte da máquina virtual paralela
  - Uso do comando *add*, no jpvmConsole

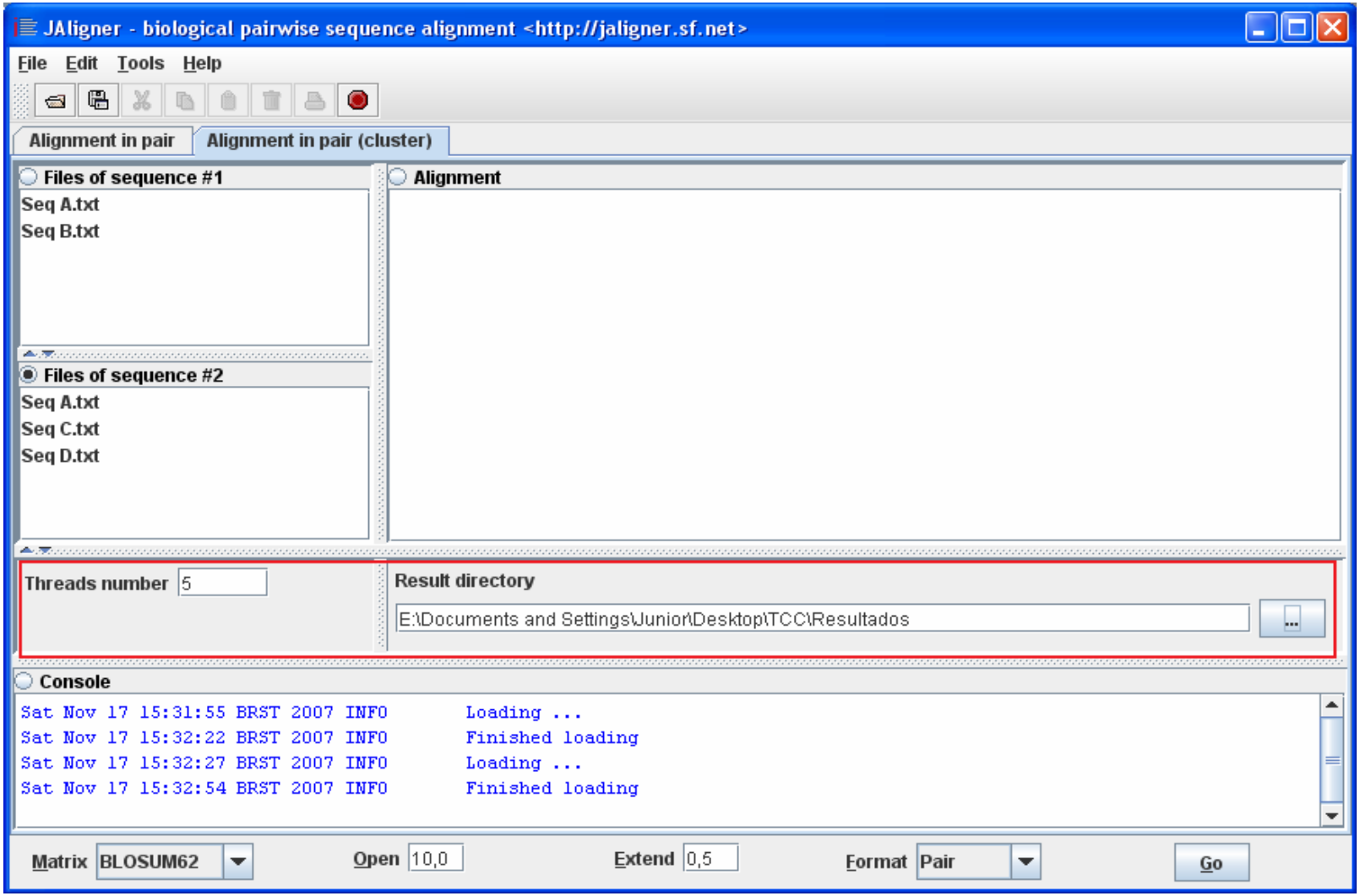
Operacionalidade

- Criando, salvando e carregando fila de seqüências



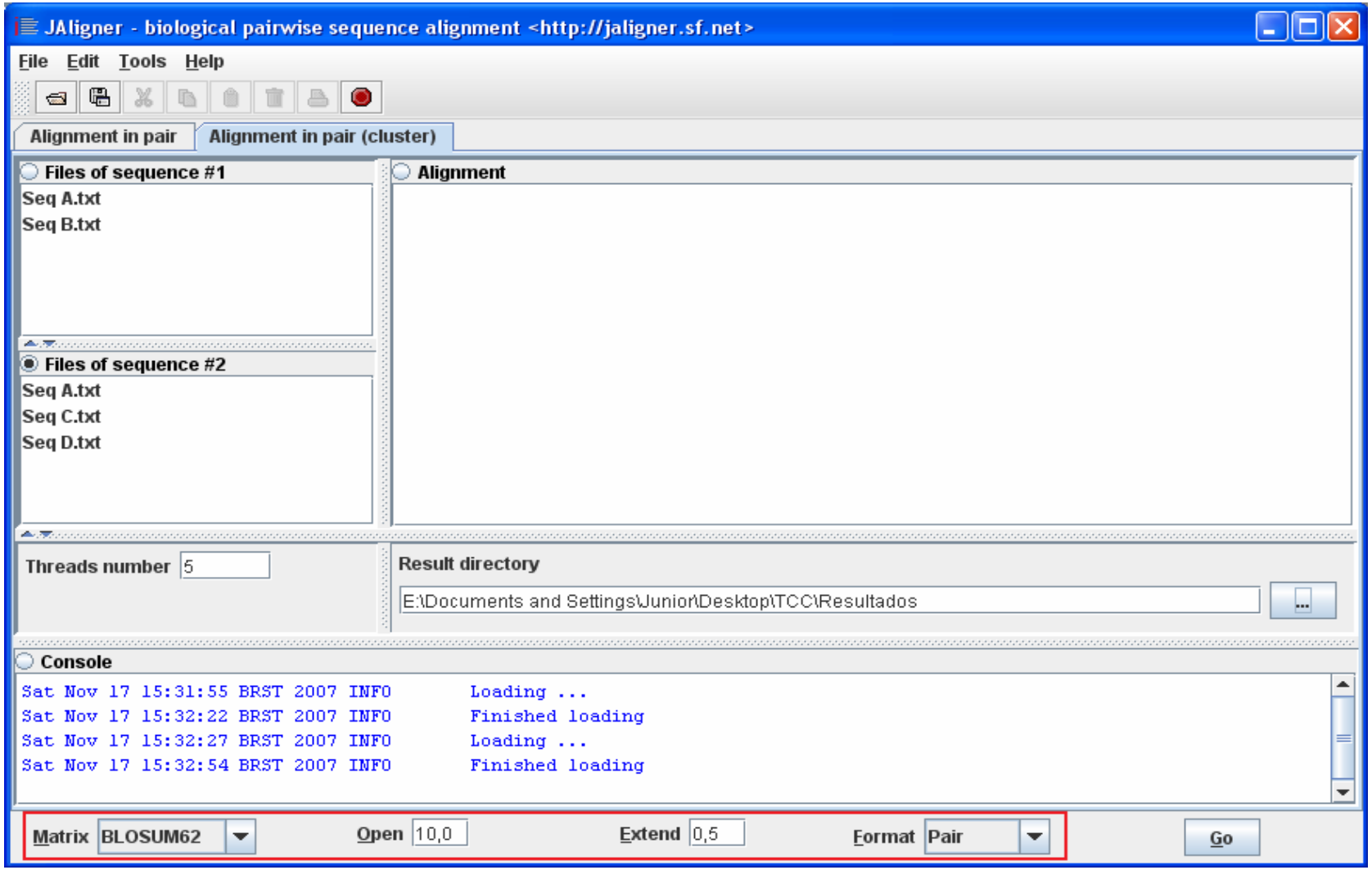
Operacionalidade

- Informando quantidade de Threads e diretório onde serão salvos os alinhamentos



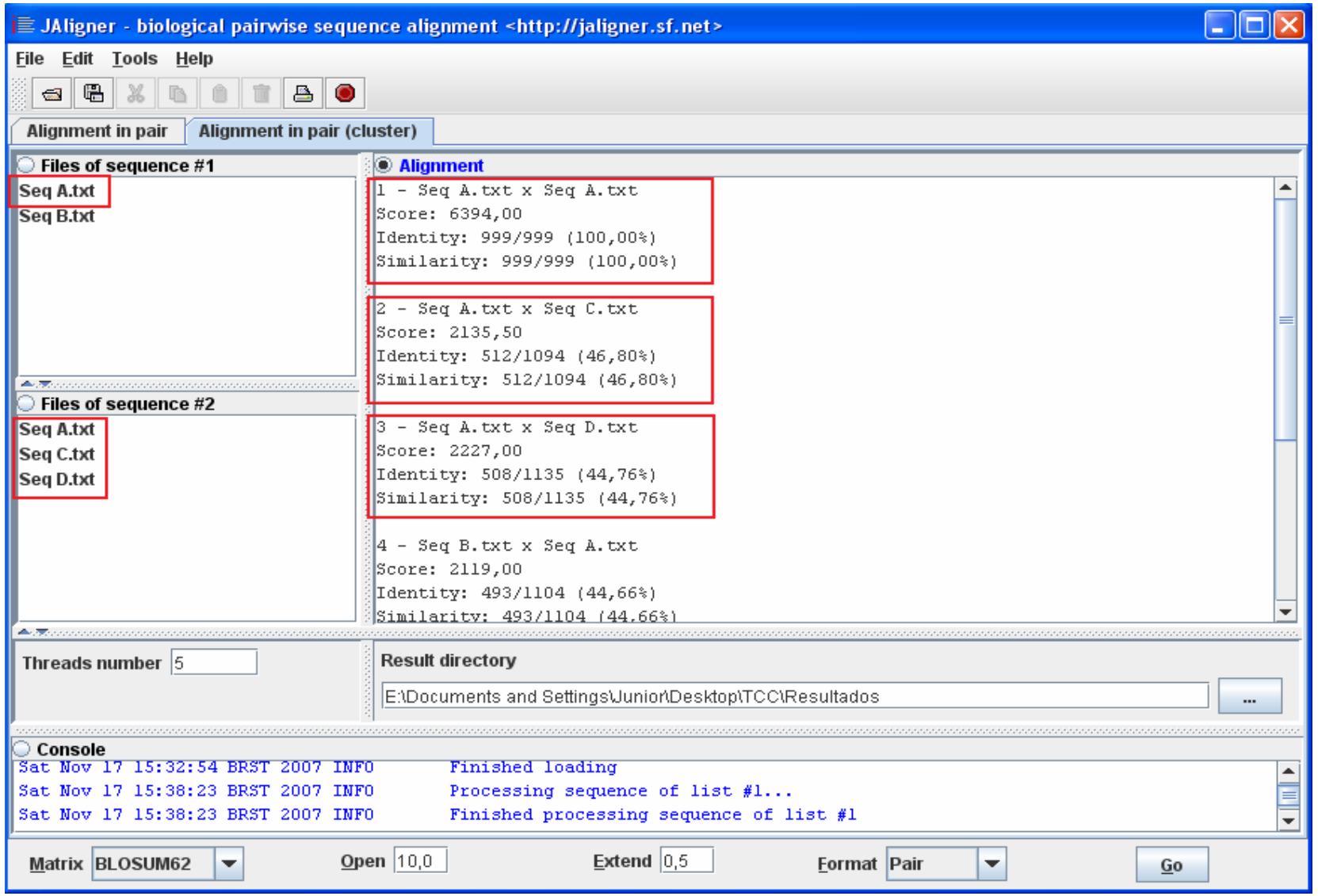
Operacionalidade

- Informando matriz de pontuação, valor para os gaps abertos e estendidos e o formato de saída do alinhamento



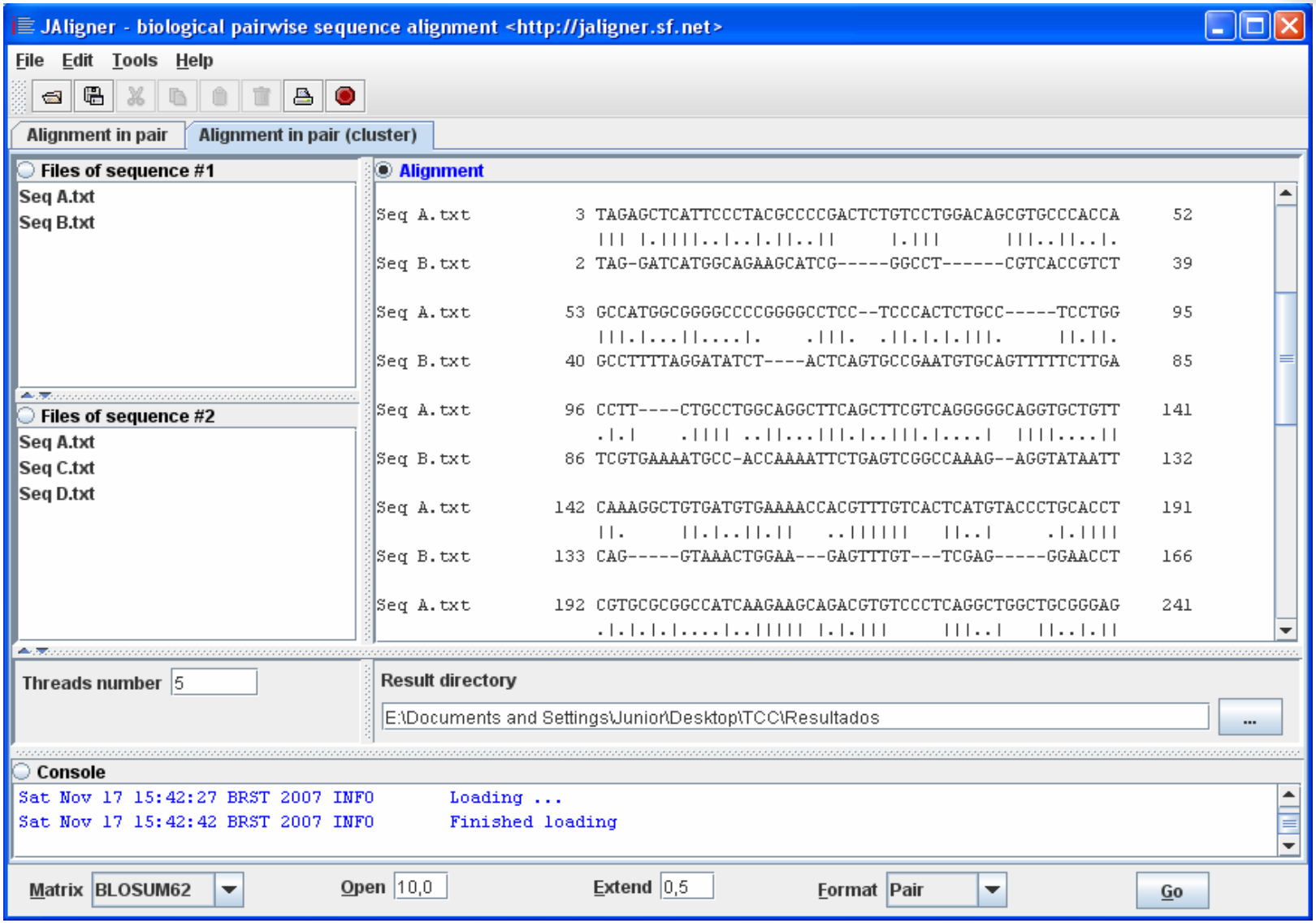
Operacionalidade

- Visualizando resultados



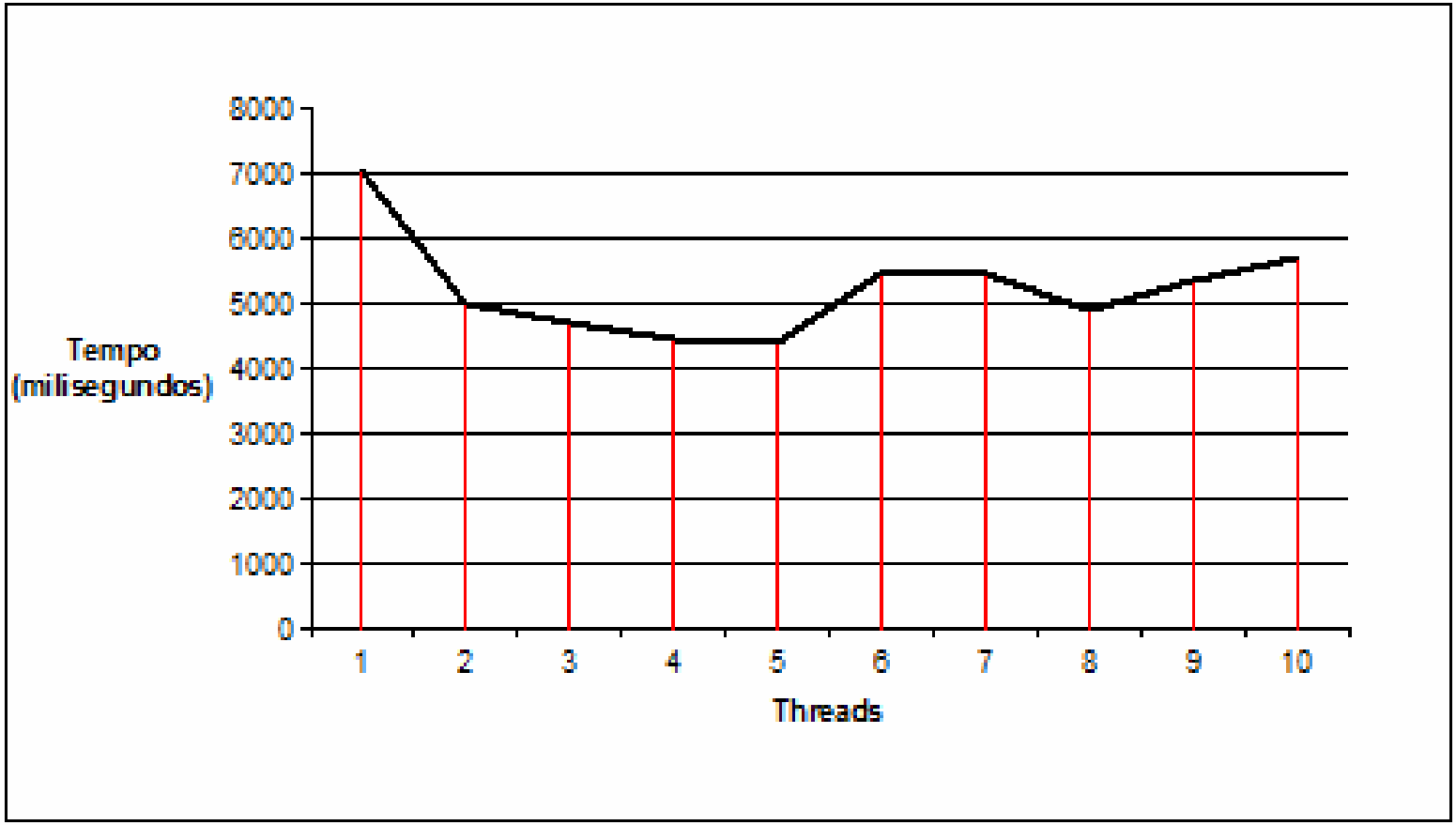
Operacionalidade

- Visualizando um alinhamento realizado



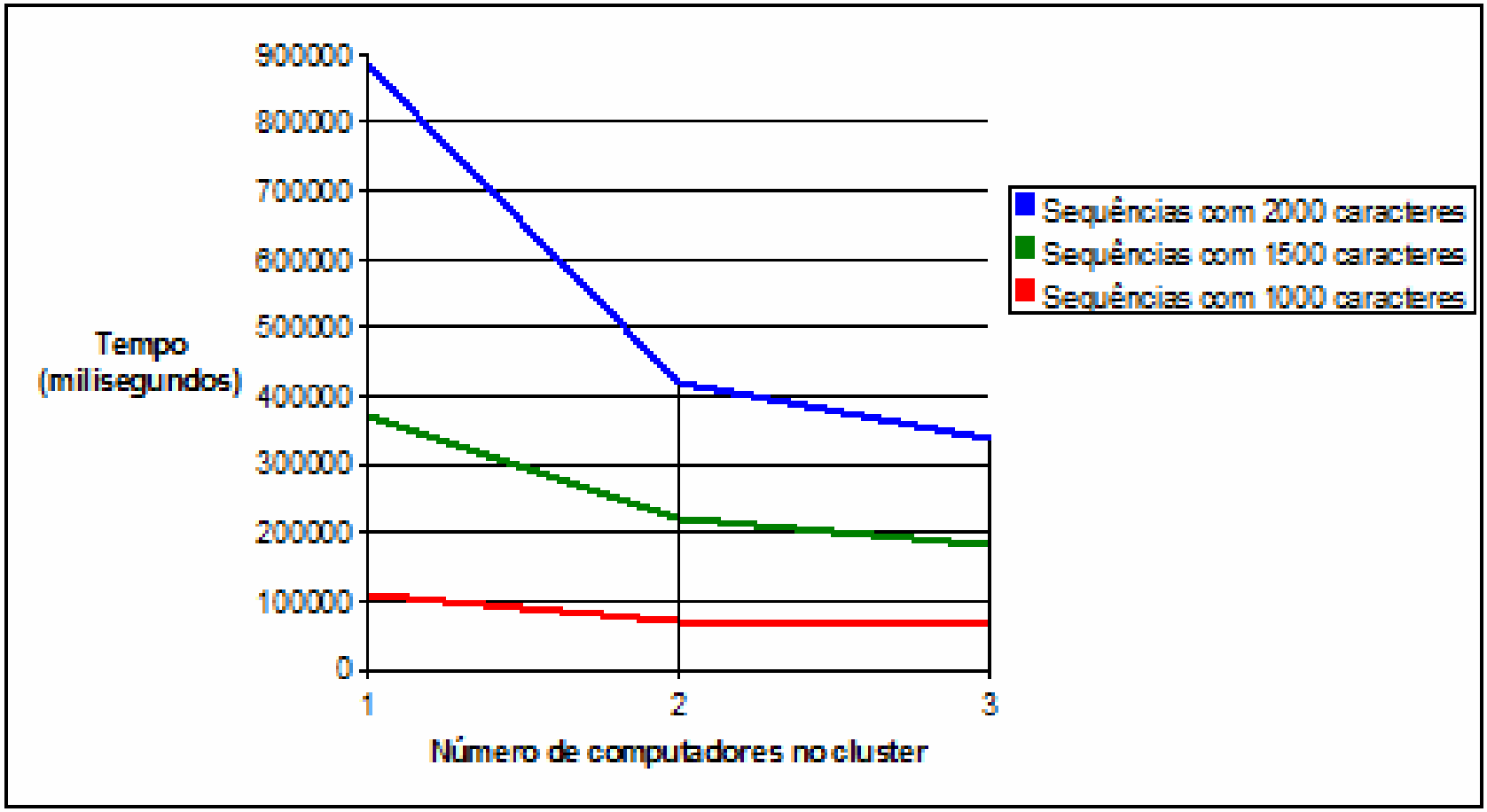
Resultados

- Resultado dos testes com o uso de várias threads.



Resultados

- Resultado dos testes do alinhamento em cluster, com até três computadores.



Resultados e discussão

- Possibilidade de criar filas de seqüências biológicas
- Ganho de performance com a paralelização do processamento do alinhamento em computadores com mais de um processador
- Diminuição do tempo de processamento ao realizar o alinhamento em ambiente distribuído

Resultados e discussão

Comparação entre JAligner, Blast e Fasta

- Blast e Fasta possuem melhor desempenho que JAligner, porém:
  - Blast e Fasta não garantem o alinhamento ótimo
  - Blast e Fasta utilizam heurísticas

Limitações JAligner

  - Uso elevado de memória RAM
  - Aguarda respostas dos nós do cluster por tempo indeterminado

Conclusão

- Uso das técnicas de programação paralela JPVM e JOMP possibilitou o processamento dos alinhamentos em ambiente distribuído

- Possibilidade de utilizar computadores, antes ociosos, de diversas arquiteturas para formar a máquina virtual paralela

Extensões

- Realização de alinhamentos múltiplos
- Meios de gerenciamento do cluster
- Uso de Grid computacional
- Apresentação dos resultados dos alinhamentos de forma gráfica