# Reconstrução Filogenética em Ambiente Distribuído

Felipe Fernandes Albrecht
Jomi Fred Hübner

Centro de Ciências Exatas e Naturais
Departamento de Sistemas e Computação
Universidade Regional de Blumenau

2006/2

# Sumário

# Conceitos básicos

## O que é filogenética

Filogenética é o estudo das relações evolucionárias entre os seres vivos.

## Teoria da evolução

- todos os seres vivos possuem um ancestral comum;
- todos seres vivos sofrem mutações, podendo ser benéficas ou maléficas;
- os mais aptos ao ambiente sobrevivem e geram descendentes e o menos aptos são extintos.

# Métodos para análise filogenética

Analisando as diferenças entre os seres vivos, é possível inferir uma filogenia entre ele, ou seja, as relações evolutivas.
As análises podem:

- morfológicas: utilizando características visíveis: tamanho, meio de reprodução, órgãos, estrutura óssea...
- molecular: utilizando características presentes no DNA, RNA e Proteínas.

# Informações moleculares

## DNA e RNA

Os Ácidos Desoxirribonucléico e Ribonucléico carregam o material genético, ou hereditário, dos seres vivos e contém as informações para a construção das proteínas.

## Proteínas

São os blocos que constituem os seres vivos. Catalizadoras de reações químicas, constituintes de organelas citoplasmáticas, anticorpos, toxinas, homónomos. [Alberts]

# Seqüencia protéica

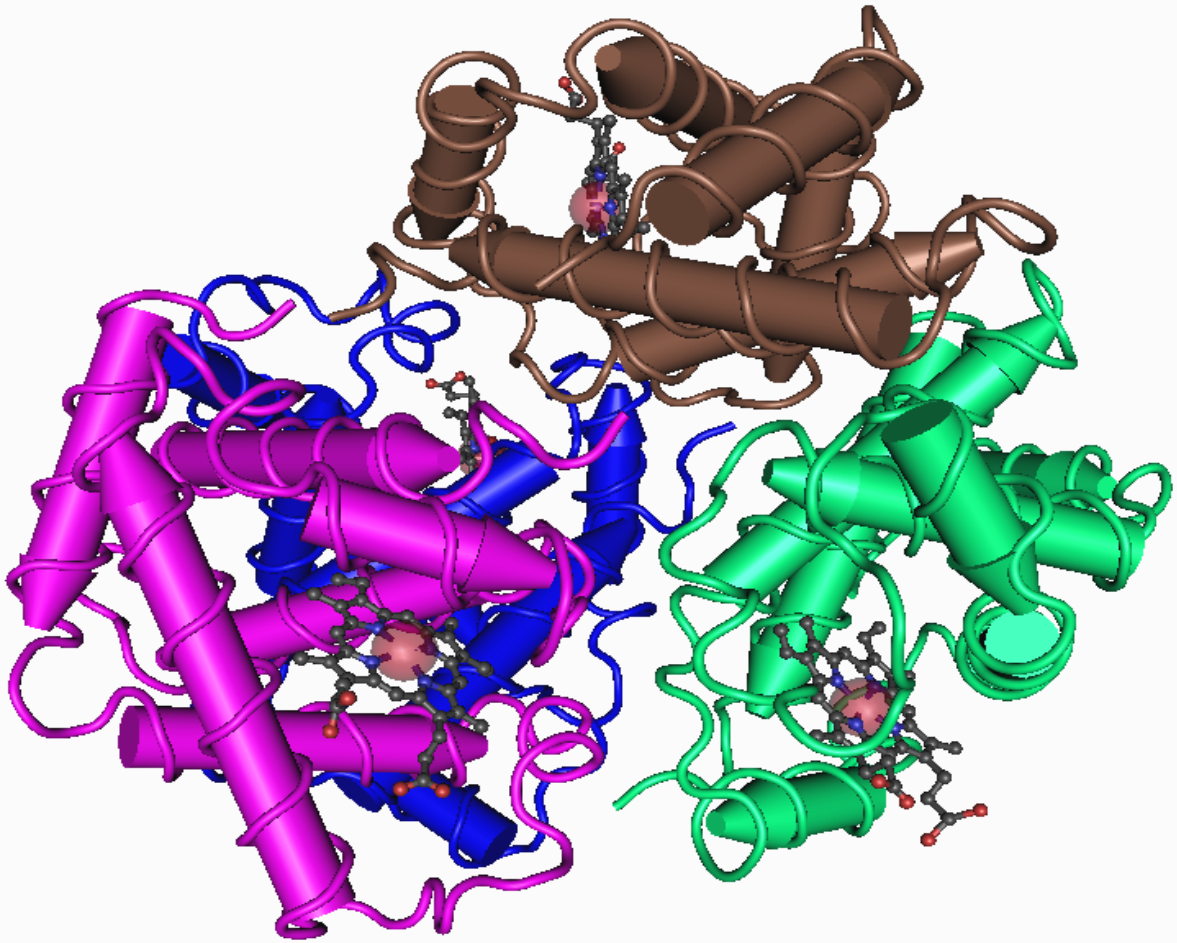
**Example**

mvlspadktn vkaawgkvga hageygaeal ermflsfptt ktyfphfdls hgsaqvkghg kkvadaltna vahvddmpna lsalsdlhah klrvdpvnfk llshcllvtl aahlpaeftp avhasldkfl asvstvltsk yr

**Aminoácidos**

São os constituintes das proteínas.
Existe 20 aminoácidos e são: hidrofóbicos e com carga: positiva, negativa.
Os aminoácidos interferem na forma das proteínas.

# Estrutura tri-dimensional



# Sumário

# O que é?

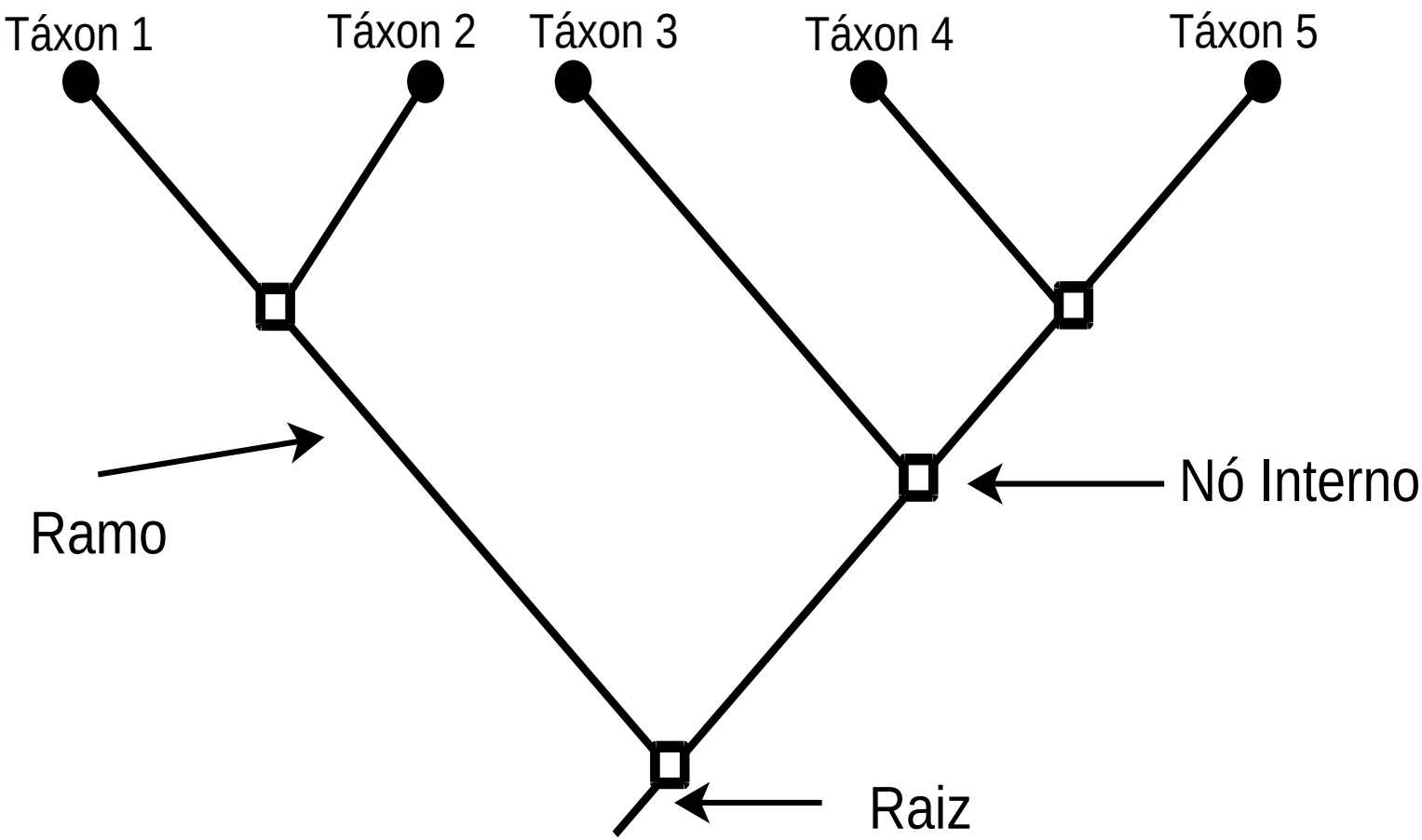
## Filogenética molecular

A filogenética molecular infere filogenias utilizando informações moleculares, como o DNA, RNA e Proteínas.
As unidades que serão inferidas, os táxons, podem ser seqüências ou trechos de proteínas, genes, sequências do genoma ou famílias de genes ou proteínas, entre outros.

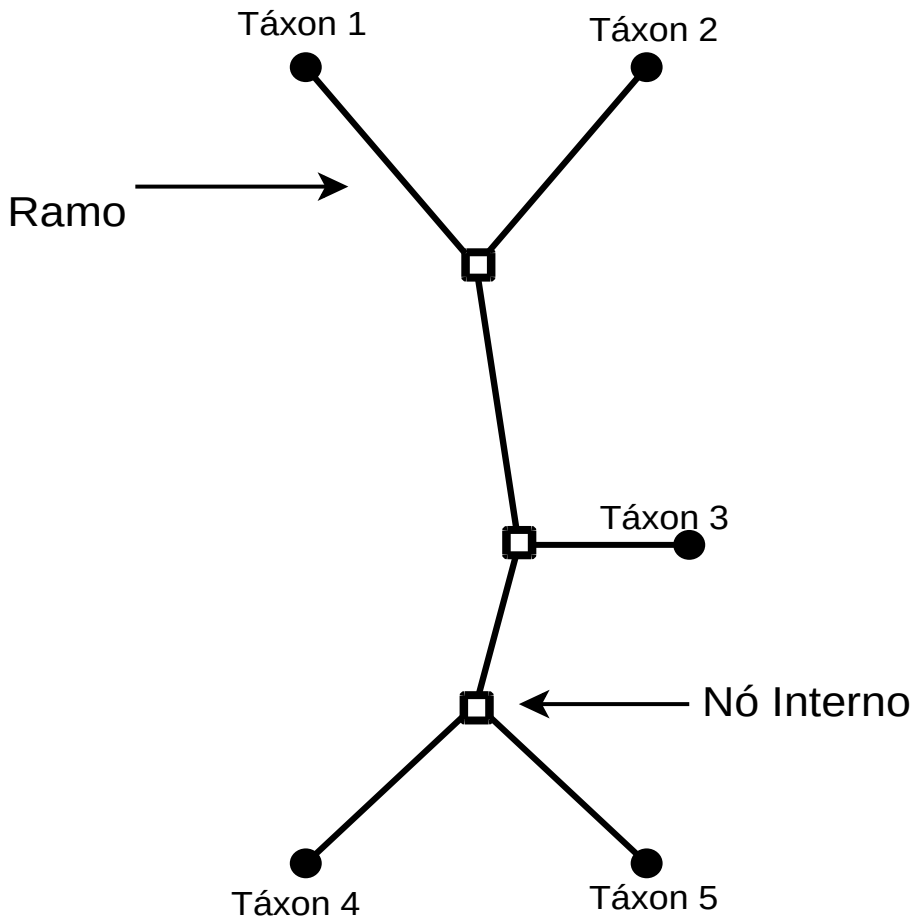
## Inferência filogenética

A inferência exibe as relações parentescas entre os táxons na forma de uma árvore.

# Árvores filogênicas com raiz



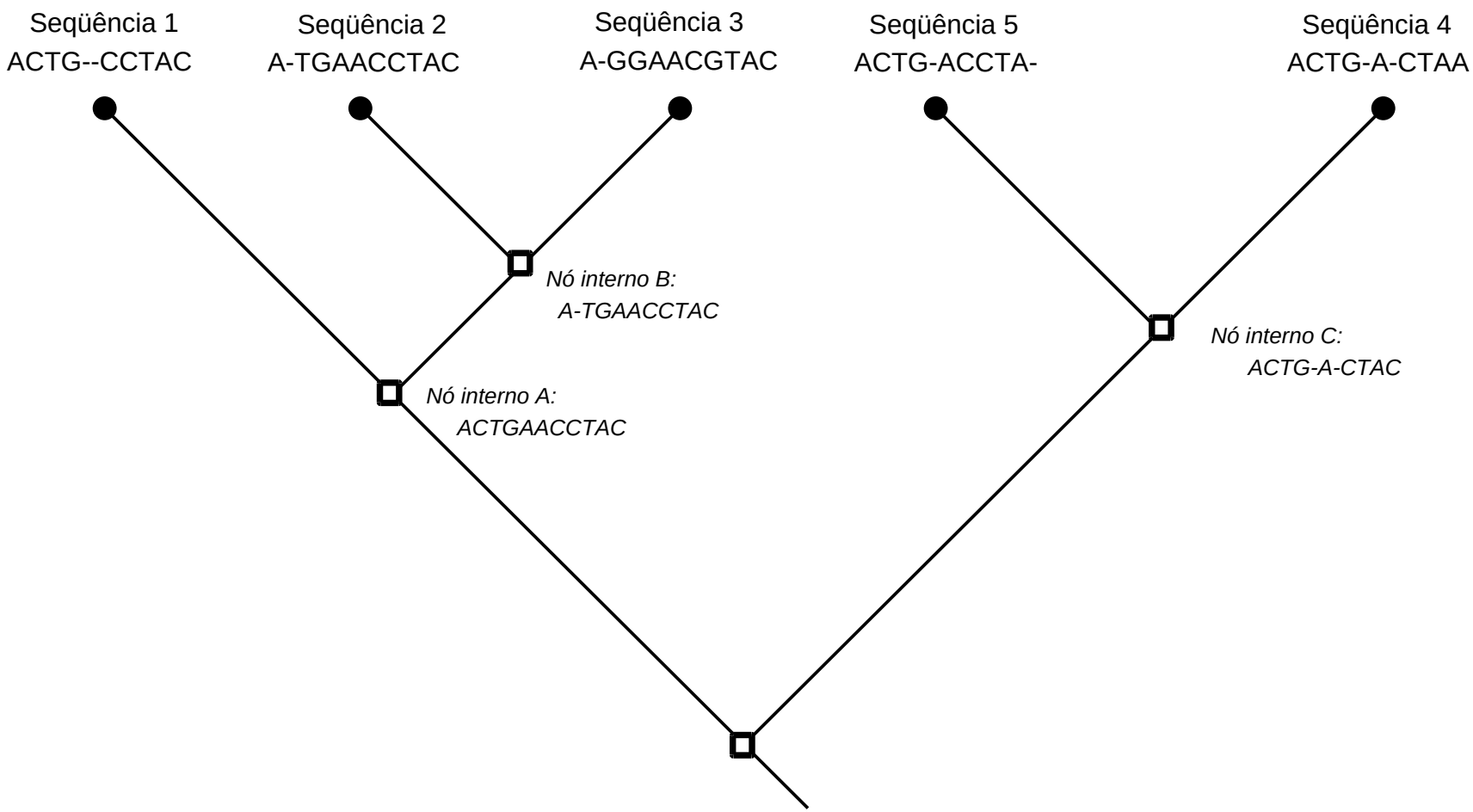
# Árvores filogenéticas sem raiz



# Técnicas para inferência filogenética

- Taxonomia Cladística;
- Taxonomia Numérica.

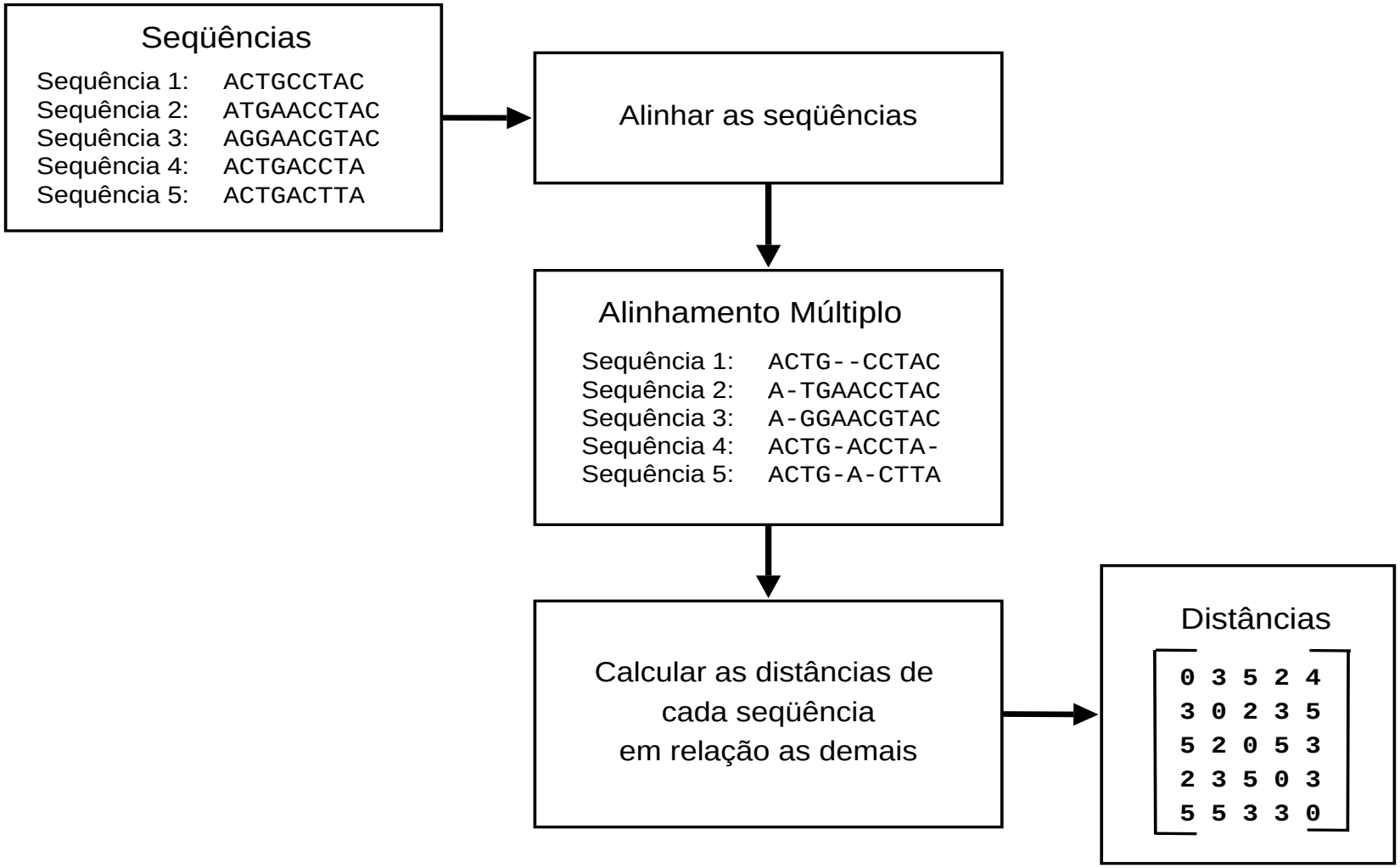
# Filogenética molecular cladística



# Filogenética molecular numérica

Ao invés de analisar as diferenças pontuais entre os táxons, é calculada uma distância que representa o grau de similaridade entre eles.

# Criação das distâncias



# Métodos para inferência filogenética numérica

- unweighted Pair-Group Method using Arithmetic averages (UPGMA);
- neighbor-joining (NJ);
- least-squares (LS).

# Least Squares

## Equação para obter o valor do Least Squares

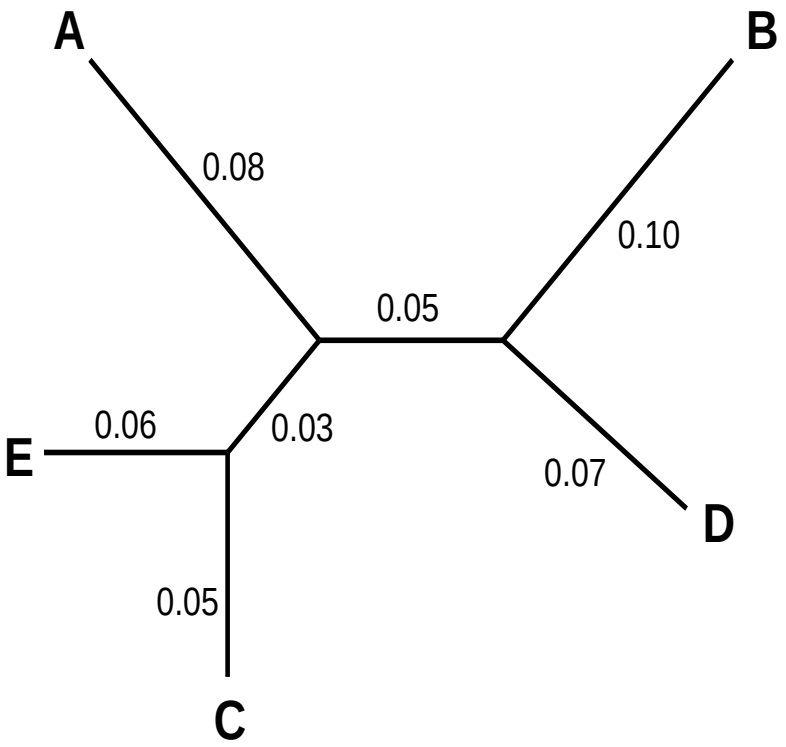
$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(D_{ij} - d_{ij})^2 \quad (1)$$

## Objetivo

Deve-se minimizar o valor de $Q$.

Ou seja: fazer com que as distâncias dos táxons na árvore representem o mais fielmente as distâncias da matriz.

# Exemplo de árvore e as distâncias na qual foi inferida




| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | 0.23 | 0.16 | 0.20 | 0.17 |
| B | 0.23 | 0 | 0.23 | 0.17 | 0.24 |
| C | 0.16 | 0.23 | 0 | 0.15 | 0.11 |
| D | 0.20 | 0.17 | 0.15 | 0 | 0.21 |
| E | 0.17 | 0.24 | 0.11 | 0.21 | 0 |

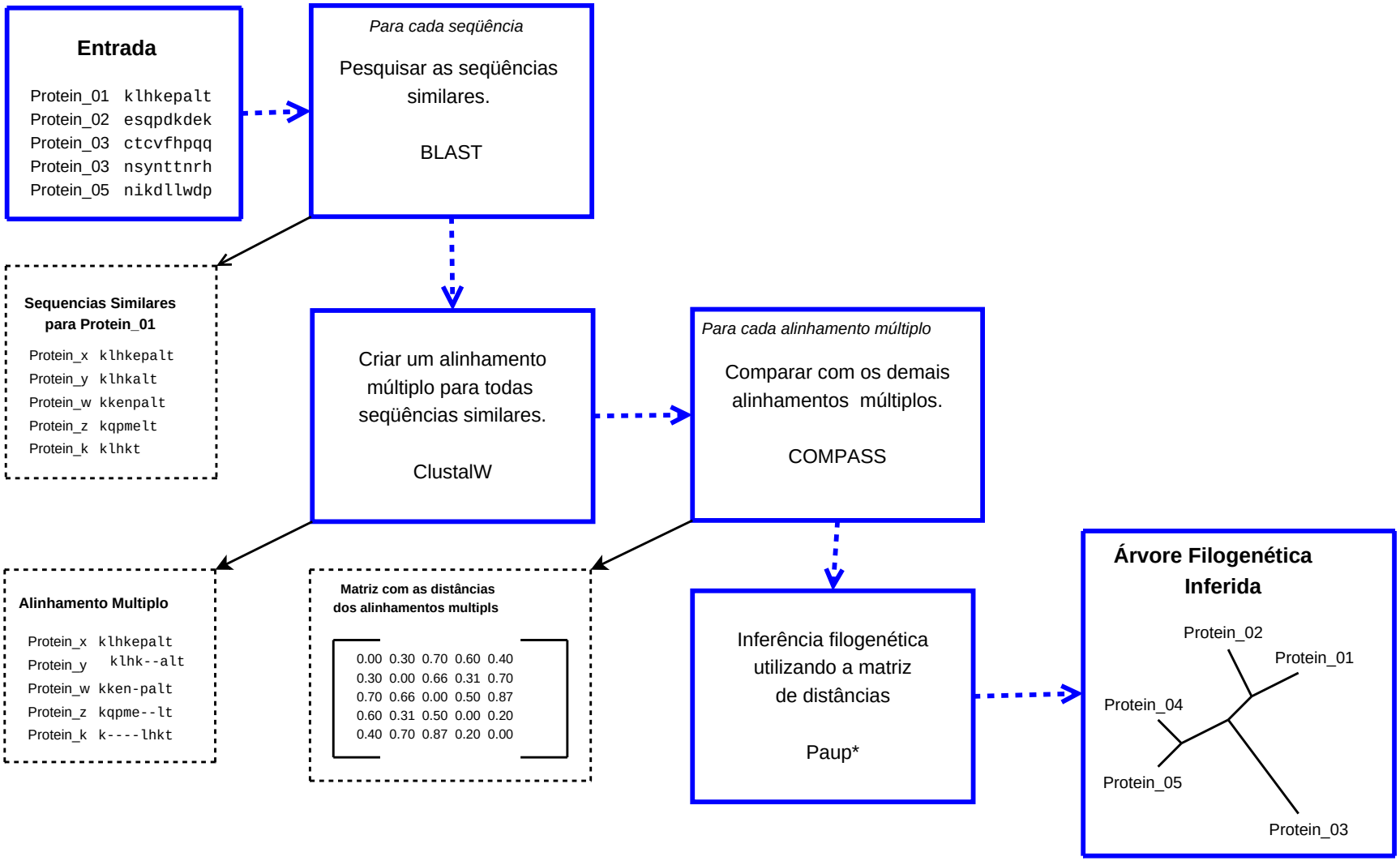
# Sumário

# Filogenética de homologias distantes

## Bases para inferência de filogenias distantes de proteínas

- o que define a função de uma proteínas é a sua estrutura e não a sua seqüência;
- estruturas terciárias são mais conservadas do que as seqüências das proteínas;
- através da estrutura da família da proteína é capaz de inferir suas relações filogenéticas.

# Passos da execução do workflow



# Sumário

# Objetivos do trabalho

## Objetivo específico

- disponibilizar uma ferramenta para a inferência filogenética em um ambiente distribuído.

## Objetivos gerais

- propor um algoritmo para inferência de árvores filogenética em ambiente distribuído;
- implementar o algoritmo num software de reconstrução de árvores filogenéticas do pacote PHYLIP;
- substituir o software PAUP* no workflow de filogenias distantes.

# Sumário

# Tempos de execução do workflow

| Trecho | Tempo em segundos | Descrição |
|---|---|---|
| BLAST | 4675 | Pesquisa por seqüências similares |
| Curl | 2202 | Obtenção das seqüências similares |
| ClusalW | 4811 | Alinhamento das seqüências |
| Compass | 918 | Comparação dos alinhamentos |
| PAUP* | 0 | Inferência da árvore filogenética |
| Consense | 3 | Consenso dar árvores inferidas |
| Total | 12614 | Tempo total de execução |

O tempo de execução com as seqüências originais são 210 minutos.

# Otimizações no workflow

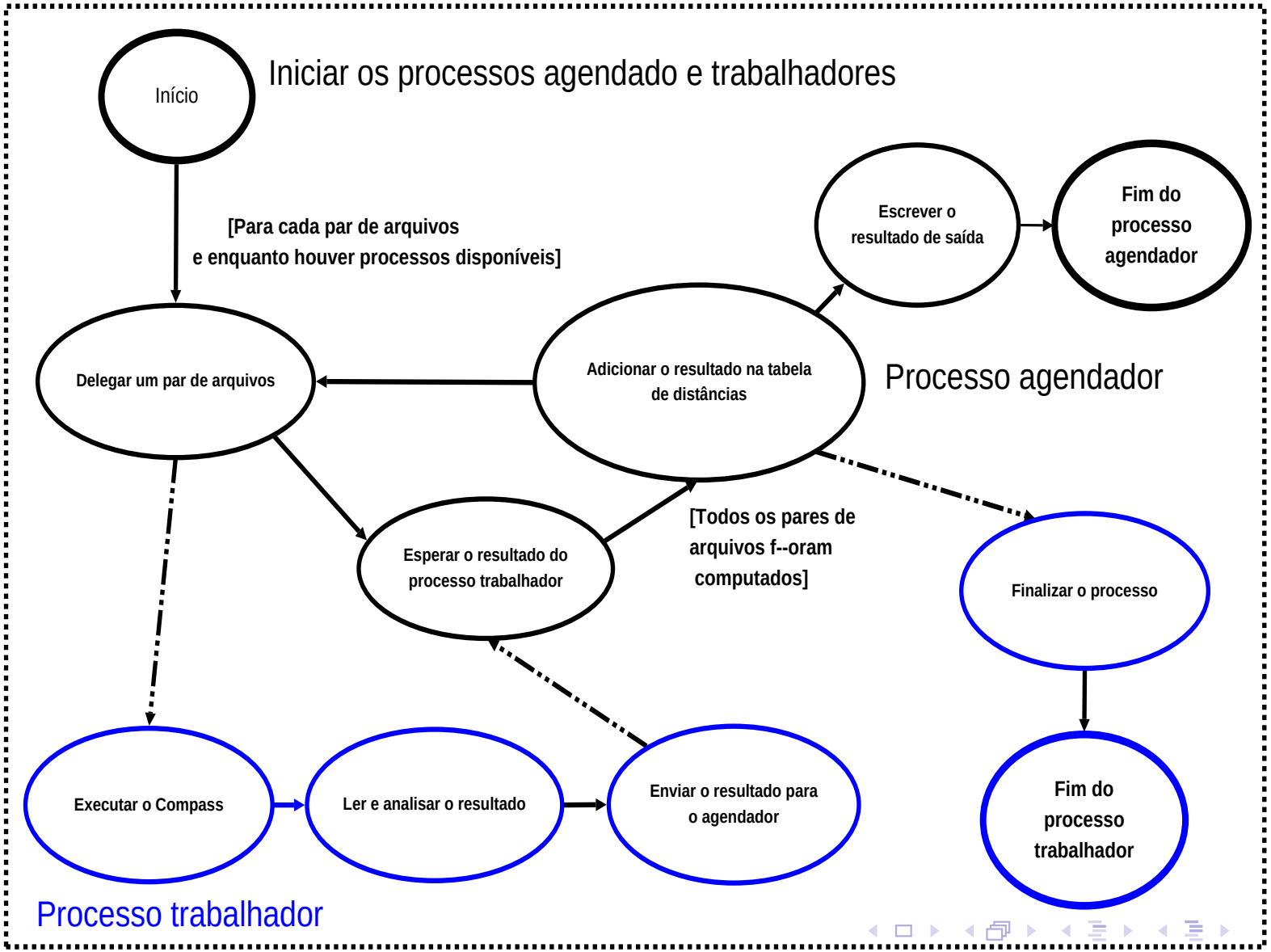
## Problema

Observado os tempos de execução de cada fase do *workflow*, verificou-se que o problema não esta na inferência filogenética através das distâncias, mas sim, no cálculo destas distâncias.

## Otimizações

- substituição do *curl* pelo *fastacmd*;
- substituição do *BLAST* pelo *mpiBlast*;
- implementação de um agendador para execuções múltiplas do *compass* em ambiente distribuído.

# Agendador para execuções múltiplas



# Resultados das otimizações

O *workflow* com as otimizações sendo executado num cluster de 5 computadores, o tempo de execução total foi reduzido de 210 para 110 minutos.
Um ganho aproximado de 52%.

# Sumário

# Algoritmo alternativo para inferência filogenética

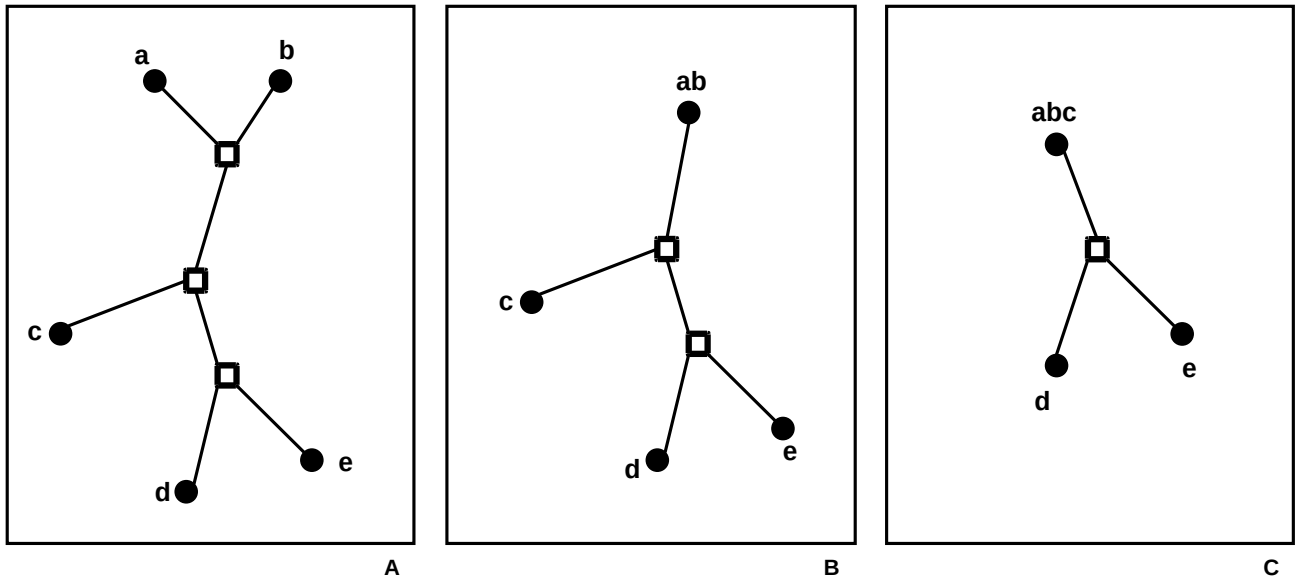
## Processo de execução do algoritmo

1. Iniciar com três táxons e crie uma árvore com eles.
2. Adicionar um novo táxon em cada possível posição da árvore.
3. Escolher a melhor árvore (menor *least square*).
4. Otimizar as distâncias dos ramos em busca de um melhor *least square*.
5. Continuar enquanto houver táxons não adicionados.

## Equação para calculo do comprimento dos ramos

$$
\begin{aligned}
v_a &= (D_{ab} + D_{ac} - D_{bc})/2 \\
v_b &= (D_{ab} + D_{bc} - D_{ac})/2 \\
v_c &= (D_{ac} + D_{bc} - D_{ab})/2
\end{aligned} \tag{2}
$$

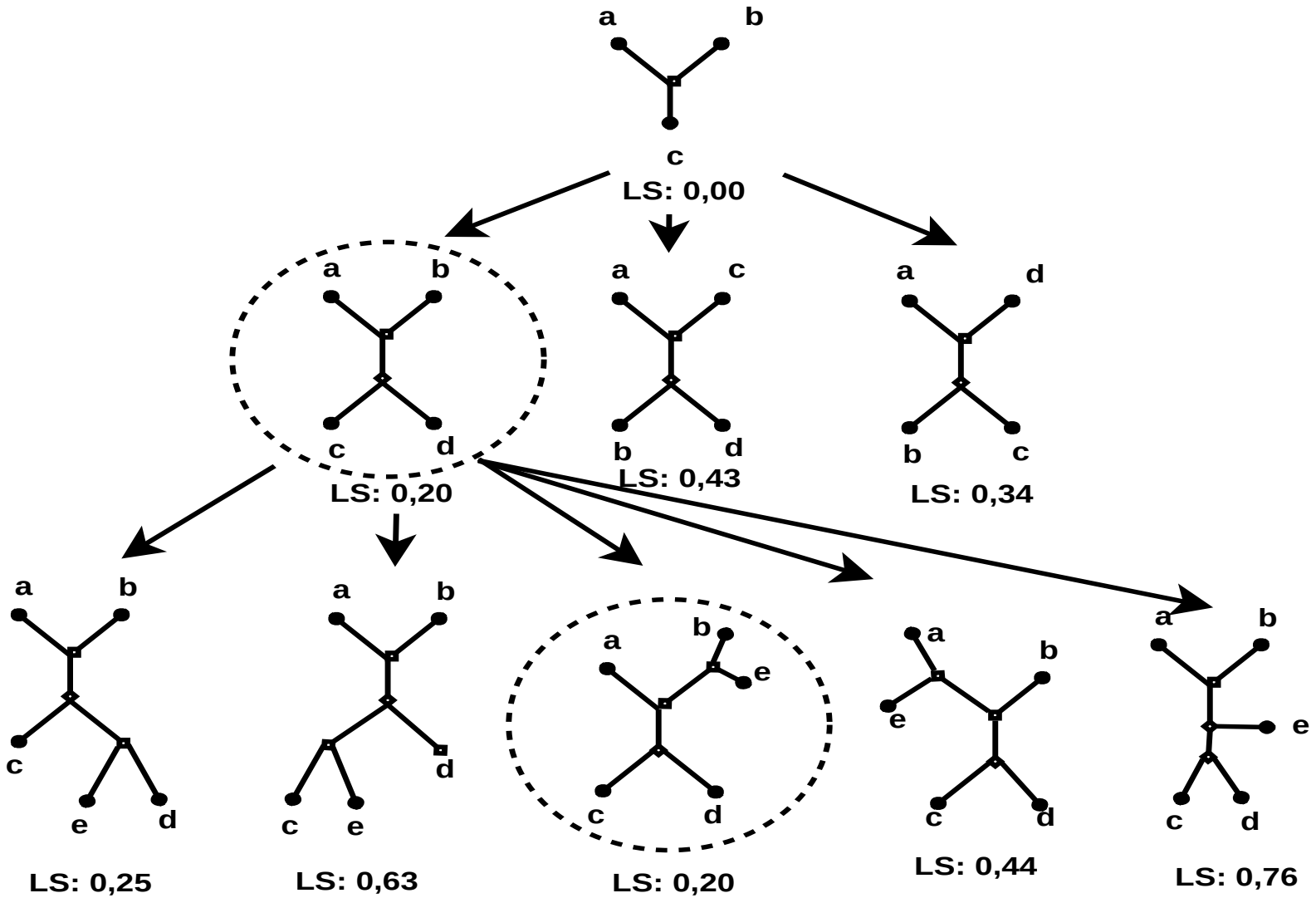
# Minimizando os nós da árvore




## Equação para minimizar a quantidade de nós

$$D_{kl} = \frac{w_{il}(D_{il} - v_i) + w_{jl}(D_{jl} - v_j)}{w_{il} + w_{jl}} \quad (3)$$

# Busca de árvores



# Heurísticas

- selecionar os táxons mais próximos para iniciar a árvore;
- calcular o desvio padrão do valor *least square* das árvores e eliminar as que estão aquém do limite definido;
- otimizar não exaustivamente as distâncias dos ramos.

# Papéis no algoritmo paralelizado

- o processo agendador: inicia as iterações e recebe informações das árvores geradas e as selecionam;
- os processos trabalhadores: constroem as árvores, calcula o *least squares*, eliminam algumas árvores e envia para o agendador.

# Funcionamento do algoritmo paralelizado

1. Agendador envia uma coleção dos trios de táxons mais próximos.
2. Os trabalhadores criam árvores com os trios recebidos.
3. Para cada táxon não adiciona na árvore:
   1. Os processos trabalhadores adicionam o táxon a cada possível posição de cada árvore e calculam o *least squares*.
   2. Calculam a média e o desvio padrão dos *least squares* e eliminam algumas árvores.
   3. As árvores restantes são enviadas ao agendador e que verifica as árvores repetidas e novamente faz uma eliminação.
   4. O agendador transmite quais árvores que devem ser eliminadas.
4. O agendador solicita a melhor árvore ao seu processo gerador.
5. Processo trabalhador envia a árvore ao agendador e este a retorna ao usuário.

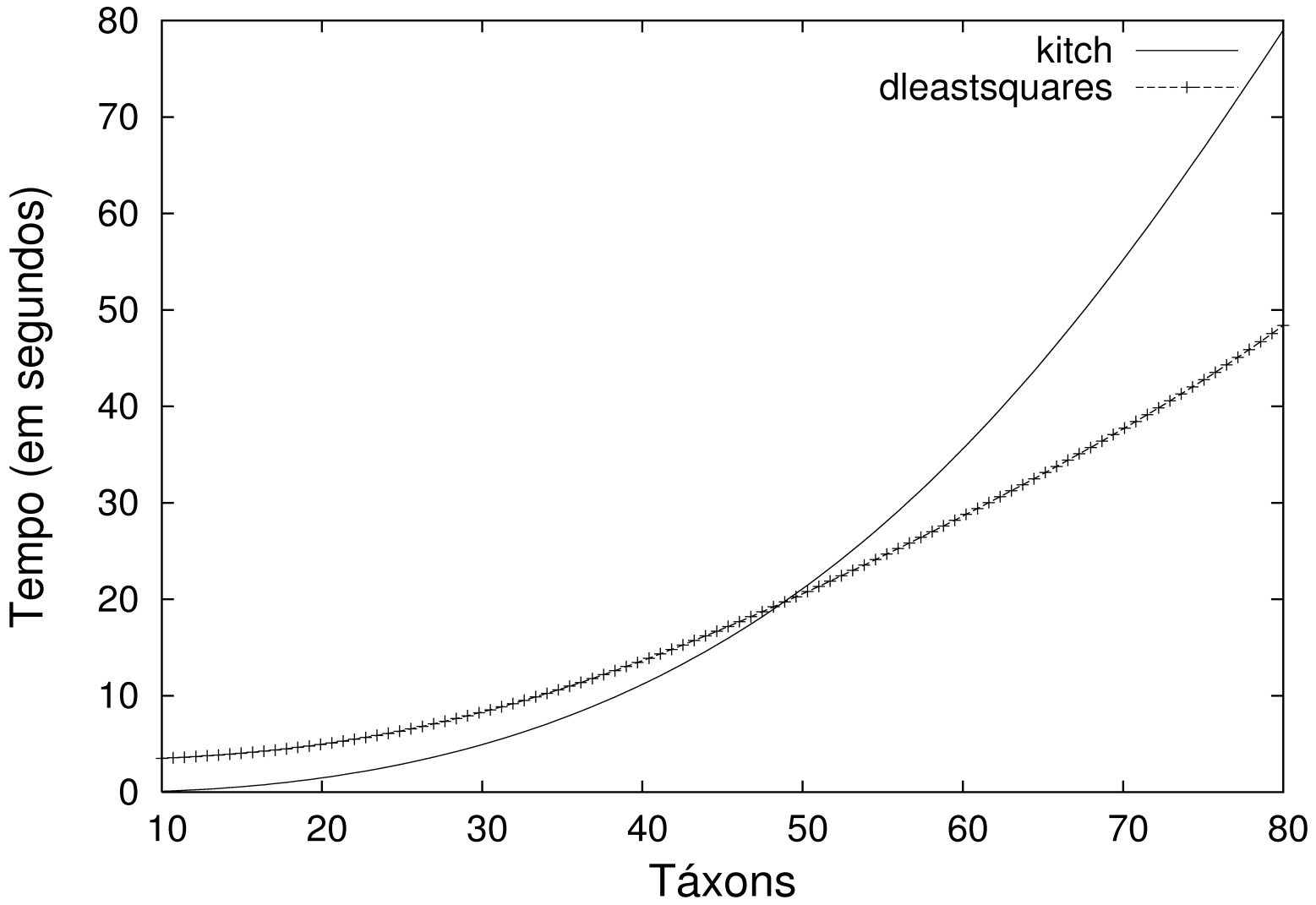
# Implementação do algoritmo paralelizado

Para implementação foi utilizado a linguagem C com o padrão MPI, utilizando o LAM como sua implementação.
O MPI fornece um meio de troca de mensagem entre os processos.

# Funcionalidade

Inferência filogenética em ambiente distribuído utilizando o padrão MPI.
Permite especificar a quantidade de trios inicias gerados, a quantidade mínima e máxima de árvores em cada processo trabalhador e a o resultado é compatível com o pacote de inferência filogenética PHYLIP.

# Resultados


kitch
dleastsquares
Tempo (em segundos)
Táxons
0
10
20
30
40
50
60
70
80
10
20
30
40
50
60
70
80

# Questões pendentes

As árvores inferidas não possuem o melhor *least square* e a qualidade é inferior aos de demais implementações, por exemplo os softwares *fitch* e *kitsch* do pacote PHYLIP.
A otimização das distâncias não esta funcionando corretamente e necessita de ajustes.

# Sumário

# Conclusões sobre as otimizações do workflow

- as otimizações no *workflow*, estão sendo utilizadas nas pesquisas de genômica comparativa do DBBM/Instituto Oswaldo Cruz/FIOCRUZ e estão se mostrando de grande valia;
- o software agendador para múltiplas execuções do software *compass* pode ser modificado para usos em outras ocasiões onde deseja executar diversas instâncias do mesmo software num ambiente distribuído;
- não foi substituído a software PAUP* pelo software desenvolvido neste trabalho, porém é possível substitui-lo pelos os softwares do pacote PHYLIP.

# Conclusões sobre o algoritmo paralelizado

- através deste algoritmo, inédito na literatura, especificado e implementado neste trabalho, é possível inferir árvores filogenética em ambiente distribuído;
- o formato de entrada e saída da implementação são compatíveis com o pacote PHYLIP e pode ser utilizado em conjunto com os softwares deste.

# Extensões

## Sugestão para o workflow

- criação de uma interface gráfica para o *workflow*;

## Suguestões para o algoritmo

- utilização de algoritmos genéticos para busca dos melhores parâmetros para a execução do algoritmo;
- aperfeiçoar as técnicas de otimização dos comprimentos dos ramos;
- implementar um sistema de balanceamento de carga entre os processos trabalhadores;
- verificar a aplicabilidade do algoritmo em outras áreas, como por exemplo, o algoritmo *MinMax*.

# Resultados e ganhos pessoais

- aprofundado o conhecimento sobre genética, proteínas e filogenética molecular;
- publicação de um trabalho para a SEMINCO em 2005;
- cooperação com o grupo de Bioinformática da Fiocruz, principalmente com o Dr. Alberto M. R. Dávila;
- conhecimento e trabalho de computação distribuída, montando *clusters* na Fiocruz e grupo de bioinformática da UFSC.

# Bibliografia


Greg Burns, Raja Daoud, and James Vaigl.
Lam.
In *Proceedings of Supercomputing Symposium'94*, pages 379–386, Toronto, 1994. University of Toronto.

Joseph Felsenstein.
An alternating least squarers approach to inferring phylogenies from pairwise distances.
*Systematic Biology*, 46(1):101–111, Mar. 1997.

Douglas L. Theobald and Deborah S. Wuttke.
Divergent evolution within protein superfolds inferred from profile-based phylogenetics.
*Journal of Molecular Biology*, 354(3):722–737, Dec. 2005.