

T-COFFEE

Coleção de ferramentas e algoritmo para
alinhamentos múltiplos de DNA, RNA e proteínas

Lorraine M. Pepe e Everton de M. Camacho 

Agenda

- ☐ O que é o TCOFFEE?
- ☐ Como o TCOFFEE funciona?
- ☐ Algoritmo
- ☐ Características
- ☐ “O que o TCOFFEE pode e não pode fazer por você”
- ☐ Acurácia
- ☐ Complexidade
- ☐ Na prática
- ☐ Referências

❑ O que é o TCOFFEE?

Tree-based Consistency Objective Function For alignmEnt Evaluation ou função objetivo de consistência baseada em árvore para avaliação do alinhamento.

- Dr. Cedric Notredame, ano 2000 em Barcelona.
- É um método de alinhamento múltiplo de sequência (MSA) e também um conjunto de ferramentas úteis e poderosas para alinhamento de DNA, RNA e proteínas.
- Permite manipular outras informações como perfis de sequências, estruturas secundárias e terciárias de proteínas.
- Possui a versão online e a versão em linhas de comando.



❑ Como o TCOFFEE funciona?

- Realiza **alinhamento múltiplo progressivo** por meio de um sistema de pesos nas posições das sequências porém mais consistente que o alinhamento em pares.
- **Utiliza abordagem heurística.**
- **Alinhamento progressivo:** Constrói o alinhamento adicionando sequências por nível de similaridade. **As mais semelhantes primeiro**, até adicionar todas.
- Organiza as sequências utilizadas com base em **árvores filogenéticas**, onde as sequências são comparadas em pares. **Folhas são sequências mais similares.**

❑ Como o TCOFFEE funciona?

- Utiliza duas bibliotecas para organizar seus dados: **biblioteca primária (b.p.)** e **biblioteca estendida (b.e.)**.
- Bibliotecas são listas de restrições com a ponderação dos pares.
- **(b.p.)** Armazena dados sobre todos os alinhamentos par-a-par e os 10 melhores alinhamentos sem sobreposição.

$$\frac{n \times (n-1)}{2} \text{ elementos}$$

- **(b.e.)** Extensão: Compara cada entrada com todas as outras e calcula peso que representa **o grau de consistência dessa com as outras**.

❑ Algoritmo

Passo 1: Alinha todos os pares de sequências → **Gera a biblioteca primária**

Passo 2: Estender a biblioteca primária → **Gera a biblioteca estendida**

Passo 3: Calcula a matriz de distâncias a partir dos alinhamentos da biblioteca estendida.

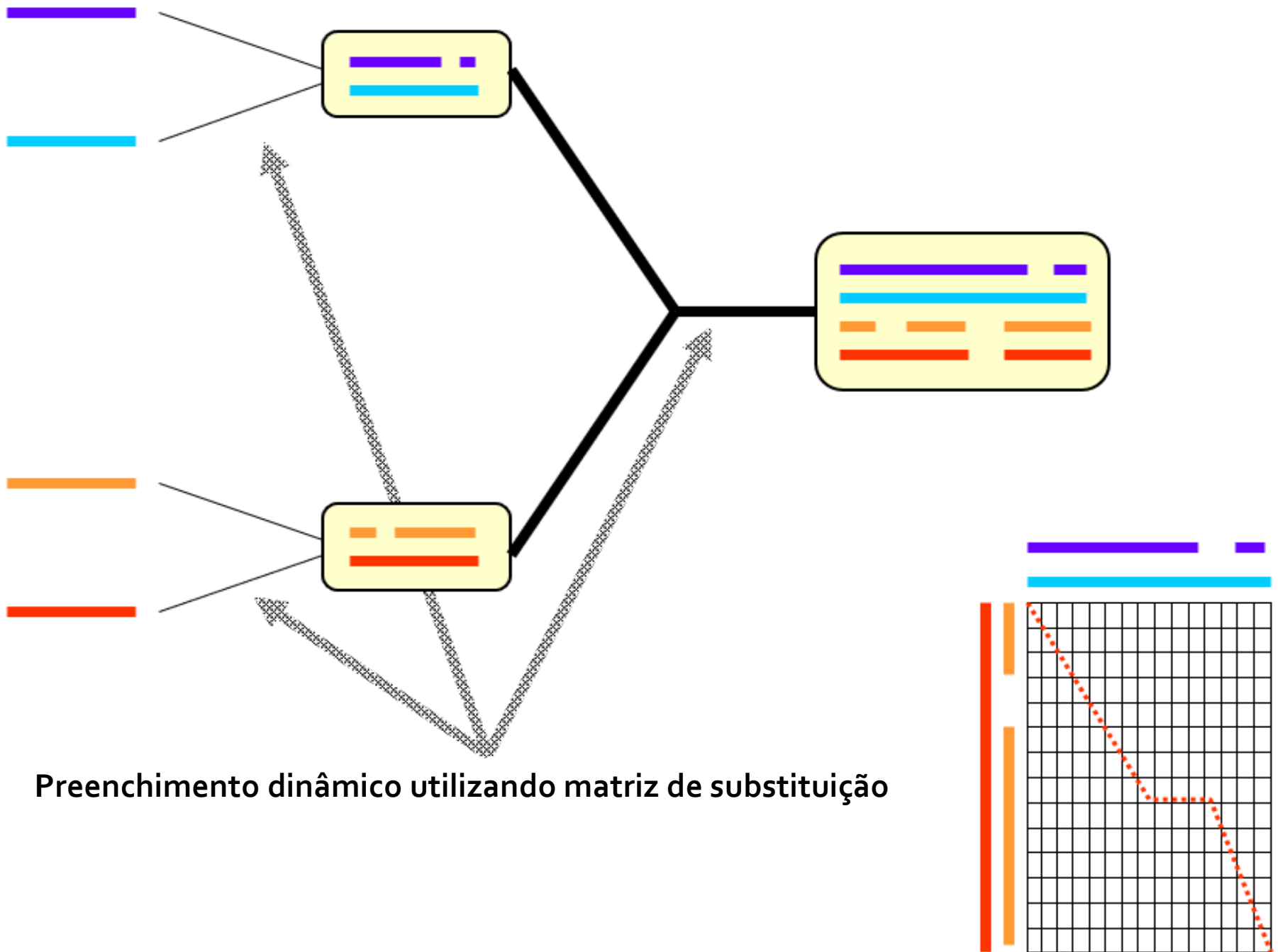
Passo 4: Constrói a árvore filogenética a partir da matriz de distâncias. Utiliza o algoritmo *Neighbor Joining* para agrupar as sequências vizinhas.

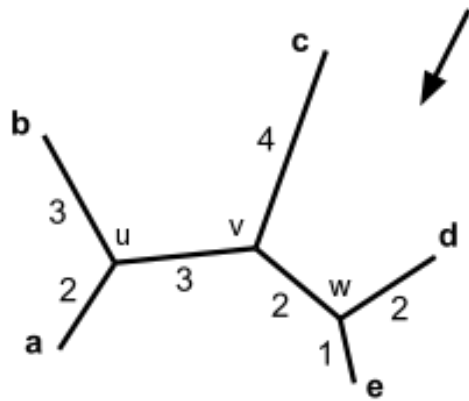
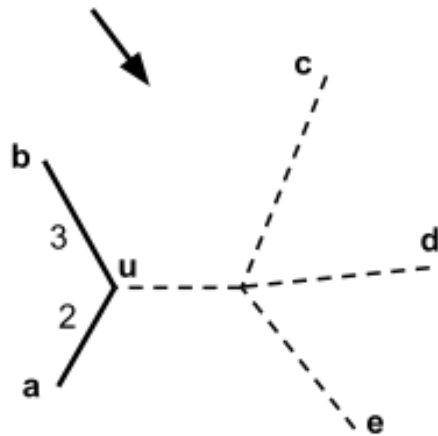
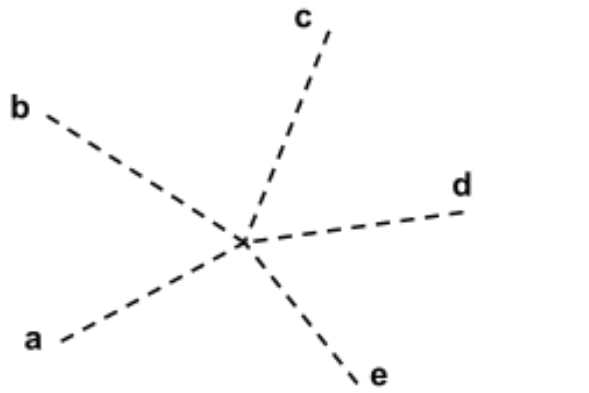
Passo 5: Alinha progressivamente as sequências seguindo a árvore criada.

SeqA	GARFIELD	THE	LAST	FAT	CAT	Prim. Weight = 88
SeqB	GARFIELD	THE	FAST	CAT	---	
SeqA	GARFIELD	THE	LAST	FA-T	CAT	Prim. Weight = 77
SeqC	GARFIELD	THE	VERY	FAST	CAT	
SeqA	GARFIELD	THE	LAST	FAT	CAT	Prim. Weight = 100
SeqD	-----	THE	----	FAT	CAT	
SeqB	GARFIELD	THE	----	FAST	CAT	Prim Weight = 100
SeqC	GARFIELD	THE	VERY	FAST	CAT	
SeqB	GARFIELD	THE	FAST	CAT		Prim. Weight = 100
SeqD	-----	THE	FA-T	CAT		
SeqC	GARFIELD	THE	VERY	FAST	CAT	Prim. Weight = 100
SeqD	-----	THE	----	FA-T	CAT	



SeqA	GARFIELD	THE	LAST	FAT	CAT	Weight = 88
SeqB						
SeqA	GARFIELD	THE	LAST	FAT	CAT	Weight = 77
SeqC	GARFIELD	THE	VERY	FAST	CAT	
SeqB	GARFIELD	THE		FAST	CAT	
SeqA	GARFIELD	THE	LAST	FAT	CAT	Weight = 100
SeqD		THE		FAT	CAT	
SeqB	GARFIELD	THE		FAST	CAT	





Neighbor Joining Algorithm

❑ Características

- Alinhar múltiplas sequências
- Simples, flexível e preciso
- Fonte heterogênea de dados
- Ameniza os erros de cálculo causado pelas sequências tomadas inicialmente
- Combina de forma eficiente informações de alinhamento global e local
- Bom para alinhamentos de sequências com baixa similaridade

❑ “O que o TCOFFEE pode e não pode fazer por você”

➤ **Não pode:** Buscar sequências. Devem ser previamente selecionadas e organizadas em formato fasta.

➤ **Pode:**

Aceitar qualquer tipo de sequência, embora existam módulos específicos.

Integrar métodos de terceiros para melhorar os resultados

Auxiliar a reformatar, colorir e estruturar dados de entrada e saída

Combinar os resultados obtidos

Combinar virtualmente vários MSAs e tenta produzir um novo

Combinar sequências e estruturas

❑ Acurácia

- O TCOFFEE está na categoria de alinhadores baseados em consistência, atualmente os melhores existentes.
- É muito preciso por si só.
- Pode ser mais preciso ainda mesclando métodos que também são considerados precisos para potencializar o método.
- Chega a detectar similaridade entre sequências com menos de 30% de identidade.
- Mostrou-se 10% mais precisos que os alinhadores tradicionais (2018).

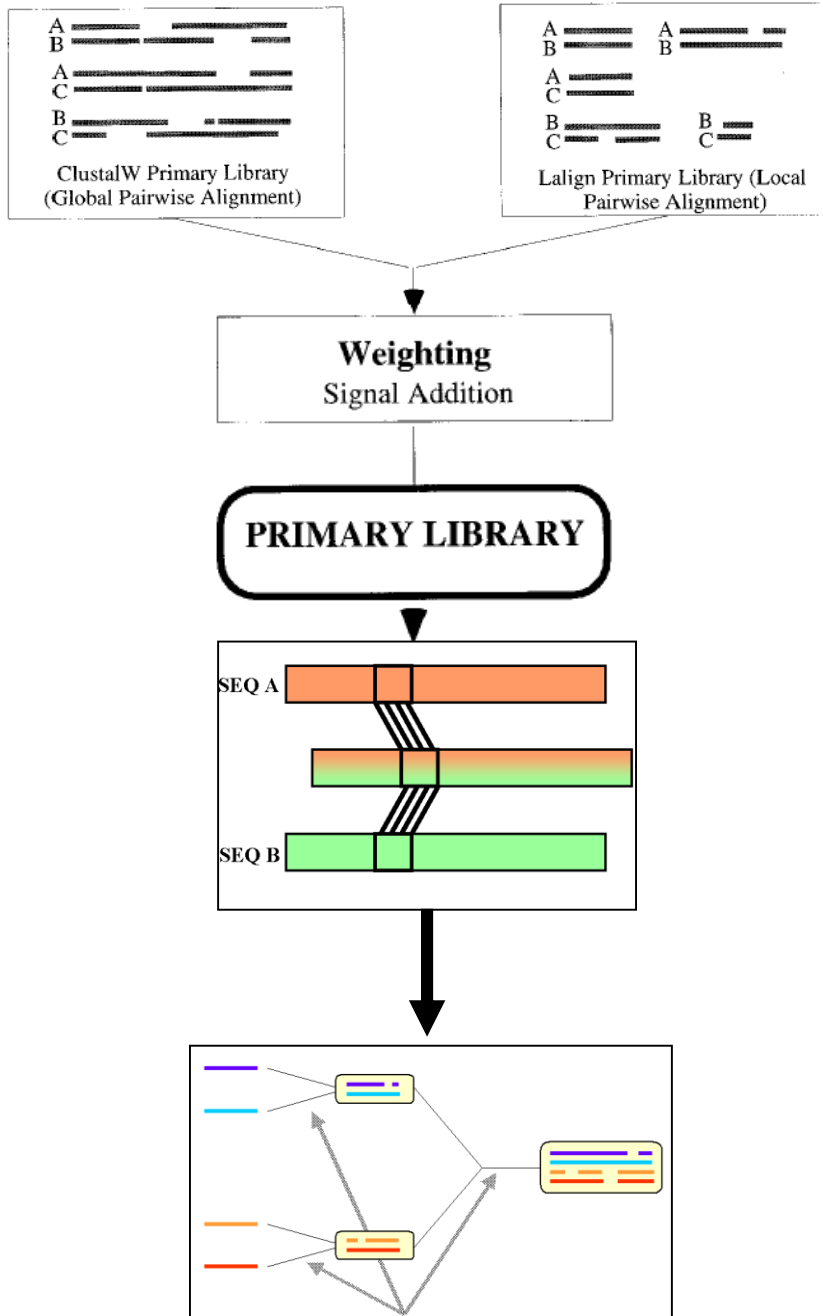
Complexidade

Biblioteca primária: $O(N^2L^2)$

N vezes mais lento
que o ClustalW

Extensão: $O(N^3L^2)$

Árvore: $O(N^2L^2) + O(N^3)$
Alinhamento: $O(NL^2)$



T-Coffee

Alinha DNA, RNA ou Proteínas usando o padrão T-Coffee

Seqüências de entrada

Cole ou envie seu conjunto de seqüências no formato FASTA

Seqüências para alinhar
[Clique aqui para usar o arquivo de amostra](#)

```
>seq1
AATCTATGAGCTAGCATGATCGA
>seq2
ACTCCGTCTATCG
>seq3
AACGGCTGATCTATGCTA
>seq4
GGATCTAGACTCCGTAGCTGATAGCTGATCCTAGCTGTATAGATCTGATCGTATGTATACGATCGTAAC
```

- OU - [Clique aqui para carregar um arquivo](#)

```
>seq1
AATCTATGAGCTAGCATGATCGA
>seq2
ACTCCGTCTATCG
>seq3
AACGGCTGATCTATGCTA
>seq4
GGATCTAGACTCCGTAGCTGATAGCTGATCCTAGCTGTATAGATCTGATCGTATGTATACG
ATCGTAAC
```

MSA

The multiple sequence alignment result as produced by T-coffee.

T-COFFEE, Version_11.00.d625267 (2016-01-11 15:25:41 - Revision d625267 - Build 507)

Cedric Notredame


SCORE=516

*

BAD AVG GOOD

*

seq1 : 65
seq2 : 63
seq3 : 65
seq4 : 40
cons : 51

seq1	AA-T-----CTA-----TGAGC-TAG-----CAT-----GATC---GA
seq2	AC-T-----CCG-----TC-TA-----T-----CG
seq3	AA-C-----GGC-----TGATC-T-----AT-----G--C---TA
seq4	GGATCTAGACTCCGTAGCTGATAGCTGATCCTAGCTGTATAGATCTGATCGTATGTATACGATCGTAAC
cons	


Result files

9 output files - [download them all](#)

Input(s)	Input sequences (151 B)
System	Command line (217 B) Log file (52KB)
Tree	dnd file (71 B)
Multiple Alignment	score_html file (6KB) clustalw_aln file (479 B) fasta_aln file (304 B) score_ascii file (603 B) phylip file (361 B)

Referências

- NOTREDAME, Cédric; HIGGINS, Desmond G.; HERINGA, Jaap. T-Coffee: A novel method for fast and accurate multiple sequence alignment. **Journal of molecular biology**, v. 302, n. 1, p. 205-217, 2000.
- LIMA, Daniel Sundfeld. Estratégia paralela exata para o alinhamento múltiplo de seqüências biológicas utilizando Unidades de Processamento Gráfico (GPU). 2012.
- Documentação TCOFFEE – Reliase 11, 2018.
- PERES, Patrícia Silva. Alinhamento múltiplo de seqüências através de técnicas de agrupamento. 2006. 61 f. Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas, Manaus, 2006.
- <http://www.tcoffee.org/homepage.html>
- <http://tcoffee.org.cat/apps/tcoffee/tutorial.html>



Obrigado pela atenção.
Dúvidas?