MSAProbs

Seminário de Bioinformática

Alunos

- Guilherme Gervaes RA: 151041946
- Marco Vinicius Guebarra RA: 151045054
- Paulo Victor de Queiroz Zanele RA: 151044244

Introdução

- Apesar de todos os avanços recentes no desenvolvimento de algoritmos de alinhamento múltiplo de sequências, o custo computacional na aplicação destes métodos, em diversas abordagens, ainda é relativamente grande. Deste modo, buscar uma solução eficiente e simples computacionalmente ainda é um grande desafio na área de Bioinformática.

- Desta maneira, em 2010, foi proposto o **MSAProbs**, um algoritmo de alinhamento múltiplo totalmente voltado à análise de **sequências proteicas**.

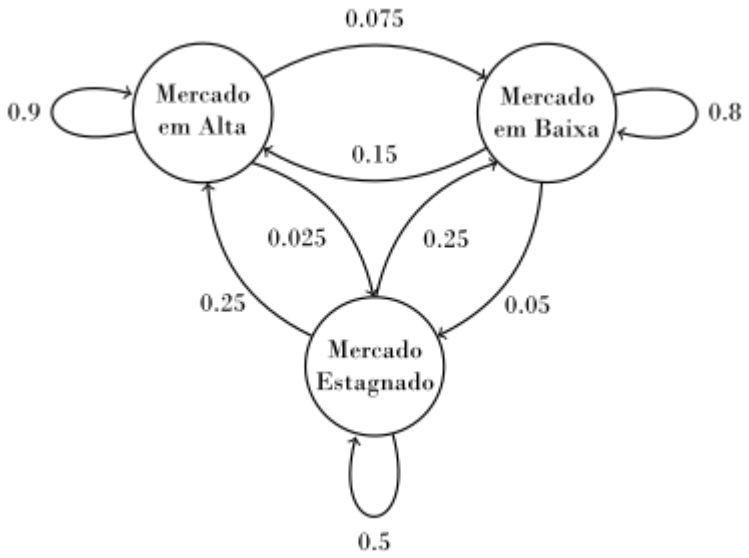
Introdução

- Basicamente, o **MSAProbs** combina um **par-HMM** (Modelo Oculto de Markov) com uma **função de partição** para calcular as probabilidades subsequentes.

Modelo Oculto de Markov

- Para entender o que é um **HMM**, primeiramente definiremos uma **Cadeia de Markov**.

- **Cadeia de Markov:** Seja um conjunto de Estados **N = S_1, S_2, ..., S_N**. Então, em um determinado tempo **t**, o sistema sofre uma alteração de estado, de acordo com a probabilidade de transição do estado atual para um novo estado.



Sejam os estados correspondentes:

S = {1 = Mercado em Alta,

2 = Mercado em Baixa,

3 = Mercado Estagnado},

Então temos a seguinte matriz de transição de estados:

$$P = \begin{bmatrix} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}.$$

Modelo Oculto de Markov

Um **HMM** é definido por:

1. Um conjunto S de N estados;
2. Um alfabeto discreto de M símbolos;
    a. No caso de sequências de proteínas, os símbolos são formados pelos 20 aminoácidos possíveis.
3. Uma matriz P de probabilidades de transição de estados;
4. Uma matriz B de probabilidades de símbolos emitidos em cada estado;
5. Um conjunto I de estados iniciais.

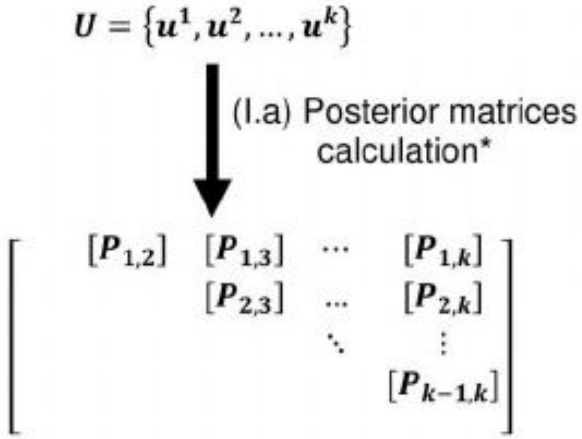
MSAProbs

O funcionamento do **MSAProbs** se dá da seguinte maneira:

1. A partir de um par-HMM, usando o algoritmo Forward-Backward e uma função de partição, calculam-se todas as matrizes de probabilidade posterior possíveis;
2. Cálculo de uma matriz de distância par-a-par, utilizando as matrizes de probabilidade posterior;
3. Construção de uma árvore guia, utilizando as distâncias entre pares e calculando o peso de cada sequência;
4. Realiza-se uma transformação de consistência probabilística ponderada, a fim de melhorar a acurácia das probabilidades posteriores de cada par de sequência;
5. Utilizando o resultado dos itens 3 e 4, realiza-se um alinhamento progressivo.

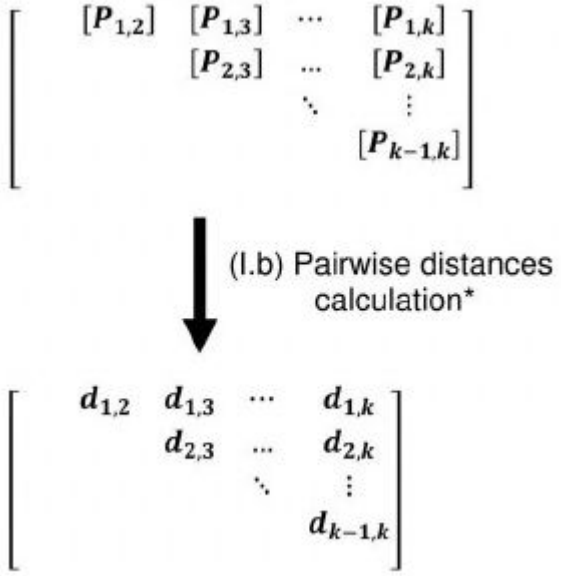
MSAProbs

- Cálculo par-a-par das matrizes de probabilidade posterior utilizando tanto o modelo oculto de Markov quanto uma função partição



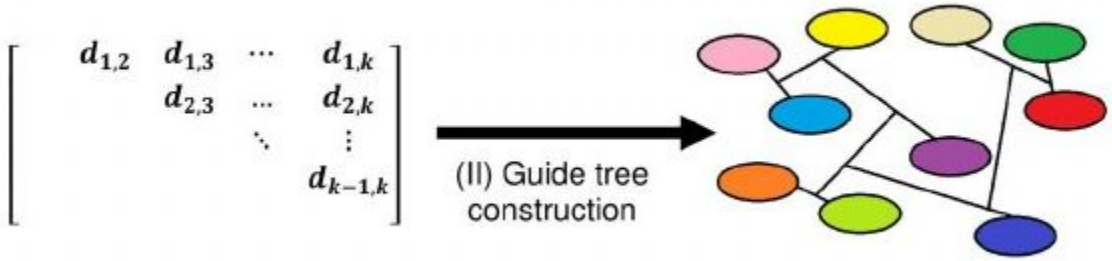
MSAProbs

- Cálculo par-a-par da matriz de distância utilizando as matrizes de probabilidade posterior



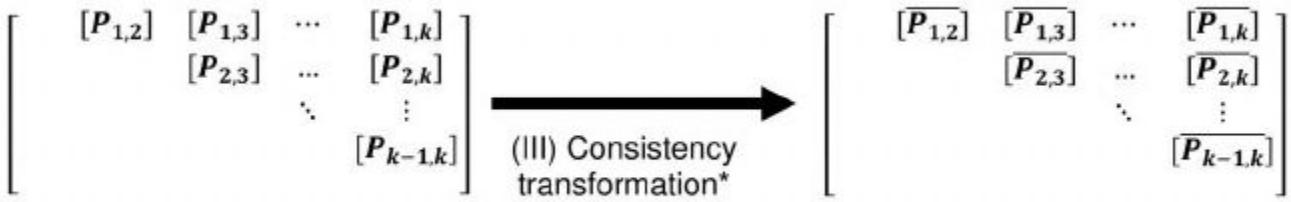
MSAProbs

- Construção da árvore guia usando a matriz de distância par-a-par e calculando os pesos da sequência



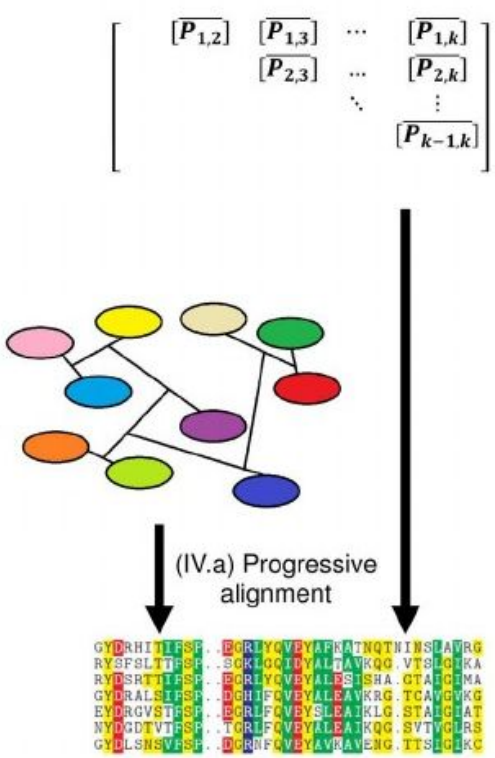
MSAProbs

- Realização de uma transformação de consistência probabilística ponderada de todas as matrizes par-a-par posteriores de probabilidade
- Passo esse realizado paralelamente ao cálculo da matriz de distâncias



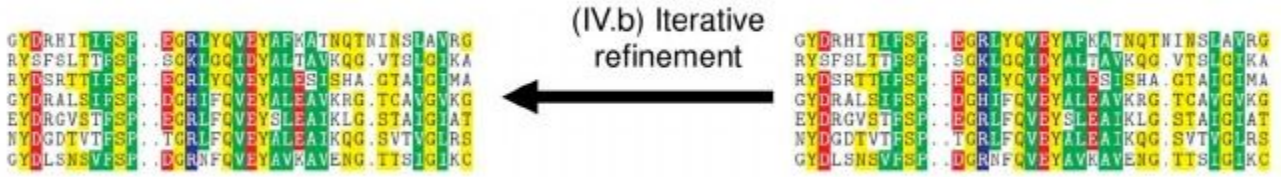
MSAProbs

- Cálculo de um alinhamento progressivo ao longo da árvore guia usando as matrizes posteriores de probabilidade transformadas



MSAProbs

- Após a etapa do alinhamento é feito um refinamento iterativo adicional como uma etapa de pós processamento
- Com o objetivo de aumentar a precisão do alinhamento



Resultados

- Comparações com outros cinco algoritmos de alinhamento múltiplo de sequências
- Exemplo de teste realizando o benchmark BAliBASE 3.0

Aligner	RV11	RV12	RV20	RV30	RV40	RV50
MSAProbs	**74.63**	**94.86**	**94.35**	**88.20**	92.32	**90.90**
MUSCLE	65.75	92.32	91.50	84.23	86.31	85.28
MAFFT	69.18	93.68	93.62	87.81	**92.53**	90.14
Probalign	71.27	94.65	93.54	86.45	92.21	89.12
ProbCons	74.00	94.59	93.70	87.54	90.03	90.15
ClustalW	58.16	88.36	88.79	77.14	78.94	76.91

Referências

Modelo Oculto de Markov:

https://repositorio.ufpe.br/bitstream/123456789/2561/1/arquivo4987_1.pdf

https://pt.wikipedia.org/wiki/Cadeias_de_Markov

MSAProbs:

https://academic.oup.com/bioinformatics/article/26/16/1958/218540