Alinhamento Múltiplo Progressivo de Sequências de Proteínas

Maria Angélica Lopes de Souza
Orientador: Zanoni Dias

23 de julho de 2010

Roteiro

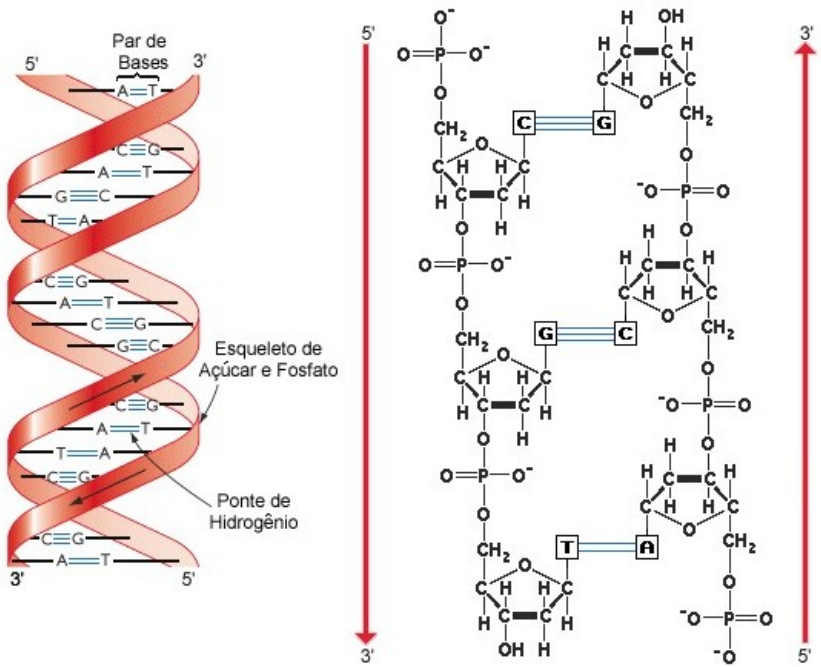
- Motivação
- Conceitos
- Alinhamento de Sequências
- Alinhamento Múltiplo Progressivo
- Alinhadores Progressivos
- Avaliação dos Alinhadores
- Conclusões e Trabalhos Futuros

Motivação

- Grande volume de dados produzido pelos projetos de sequenciamento
- Necessidade de analisar dados produzidos
- Bioinformática
- Alinhamento de sequências

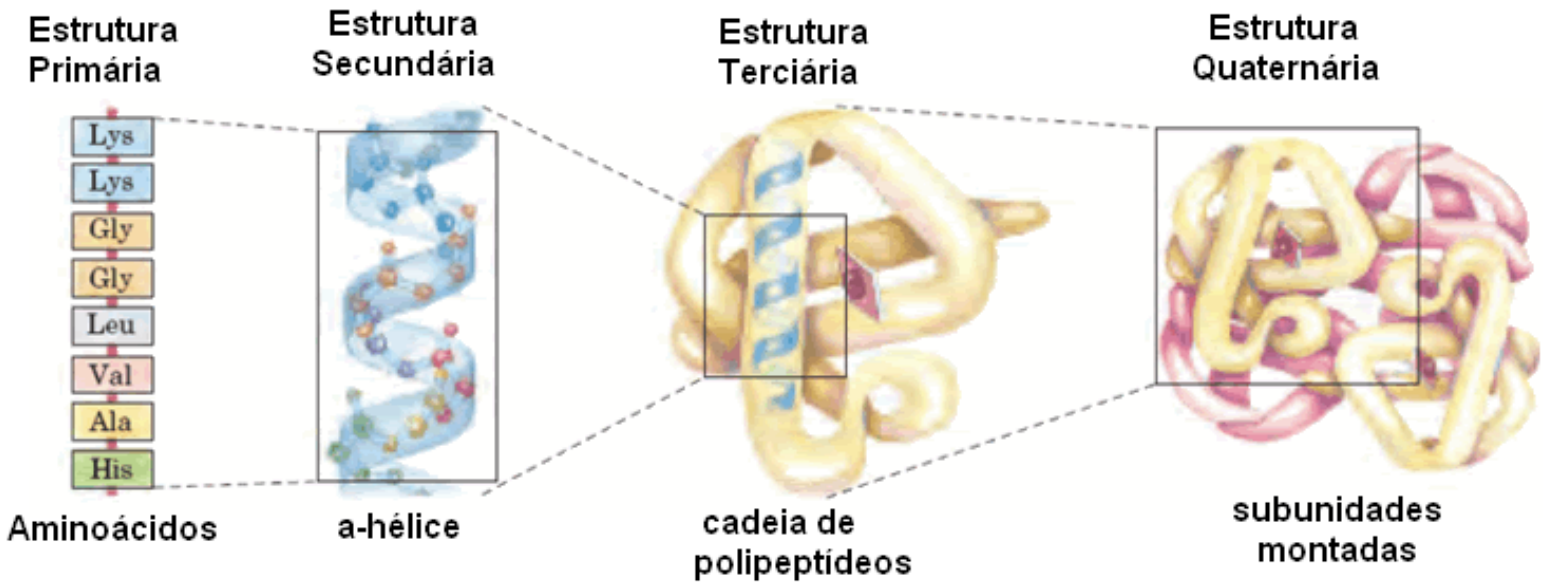
DNA, RNA e Proteínas

- Genômica é a área da biologia que estuda o genoma de diferentes organismos
- DNA → RNA → Proteínas
- Proteômica



DNA, RNA e Proteínas

- Genômica é a área da biologia que estuda o genoma de diferentes organismos
- DNA → RNA → Proteínas
- Proteômica



Alinhamento de Sequências

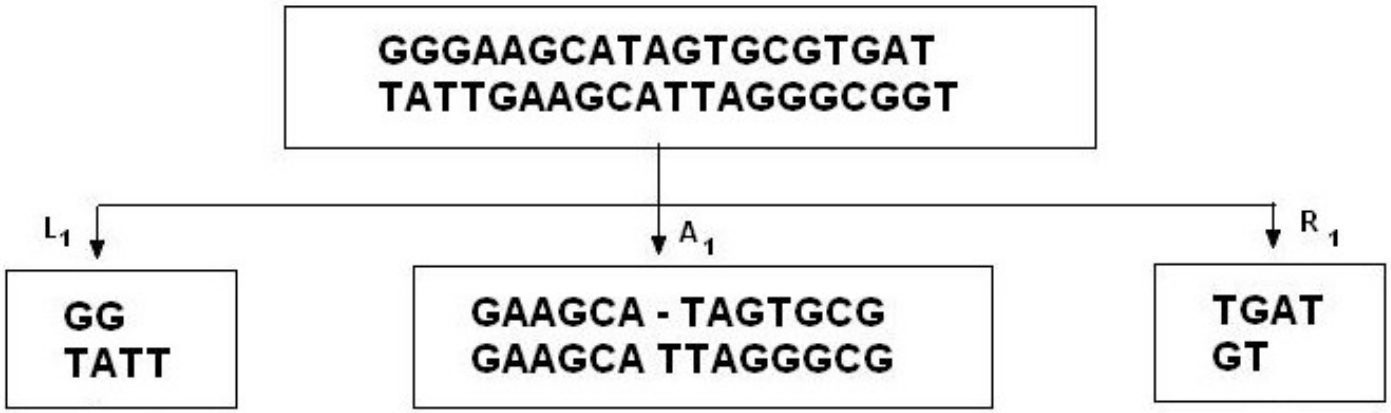
- Entendimento da estrutura, função e evolução dos genes que compõem um organismo
- Busca posicionamento de bases iguais, ou semelhantes, em uma mesma coluna
  Alinhamento das sequências `CCAACTGGACACT` e `CGAACATGGAG`:
  `CCAAC-TGGACACT`
  `CGAACATGG--A-G`
- Pontuação: *match*, *mismatch*, *gap*
- Alinhamento de Pares de Sequências
  - Global (Needleman e Wunsch, 1970)
  - Semi-Global
  - Local (Smith e Waterman, 1981)

Alinhamento Local Recursivo

- Alinhar corretamente as partes bem conversadas das sequências
- Alinhamento local do par de sequências. Recursão à esquerda e à direita para as subsequências não alinhadas
- `GGGAAGCATAGTGCGTGAT` e `TATTGAAGCATTAGGGCGGT`

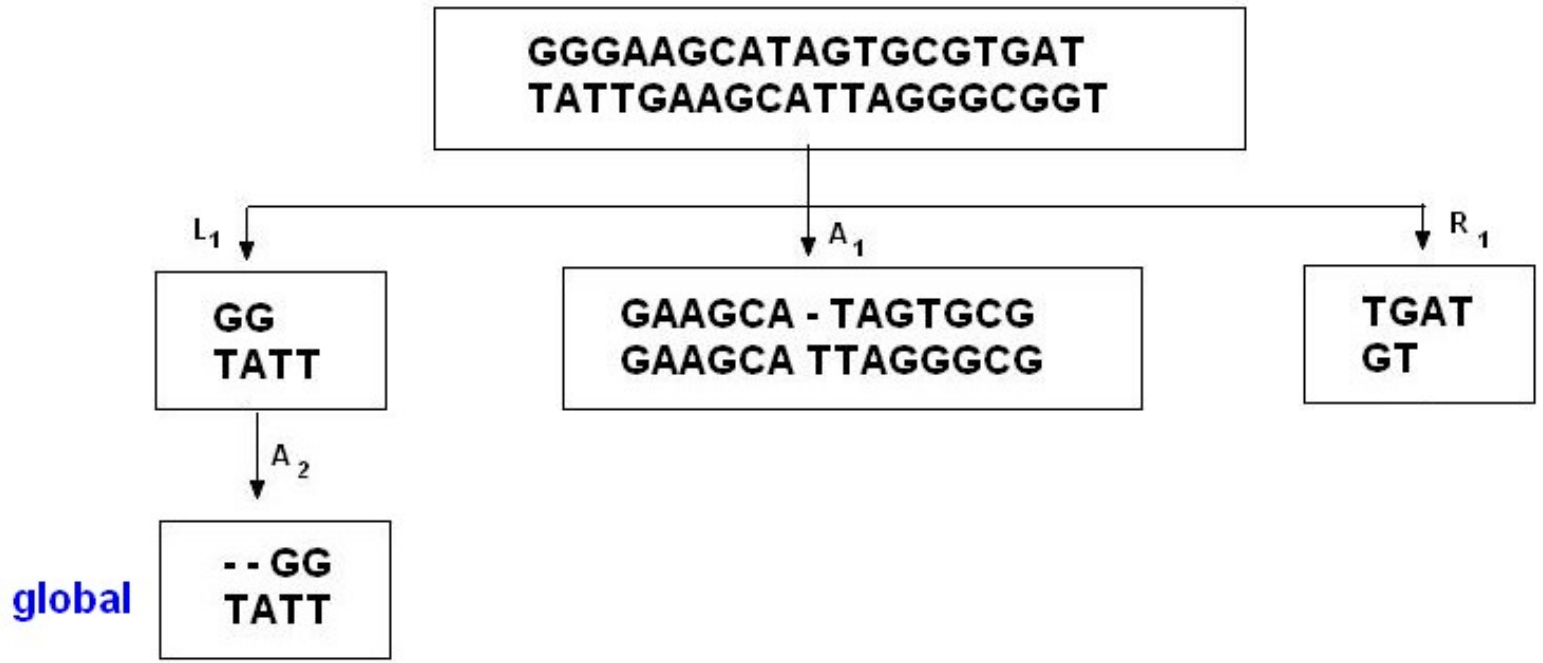
Alinhamento Local Recursivo

- Alinhar corretamente as partes bem conversadas das sequências
- Alinhamento local do par de sequências. Recursão à esquerda e à direita para as subsequências não alinhadas
- `GGGAAGCATAGTGCGTGAT` e `TATTGAAGCATTAGGGCGGT`



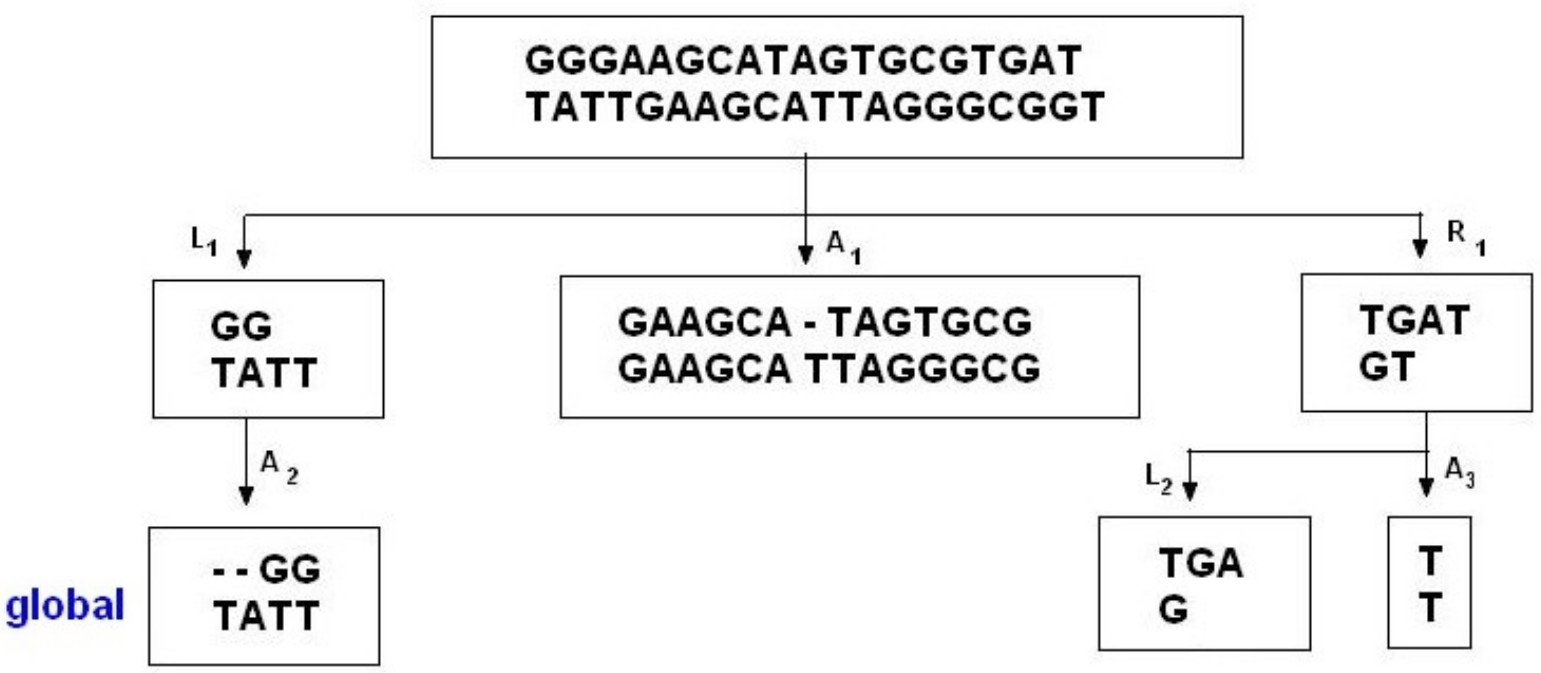
Alinhamento Local Recursivo

- Alinhar corretamente as partes bem conversadas das sequências
- Alinhamento local do par de sequências. Recursão à esquerda e à direita para as subsequências não alinhadas
- `GGGAAGCATAGTGCGTGAT` e `TATTGAAGCATTAGGGCGGT`



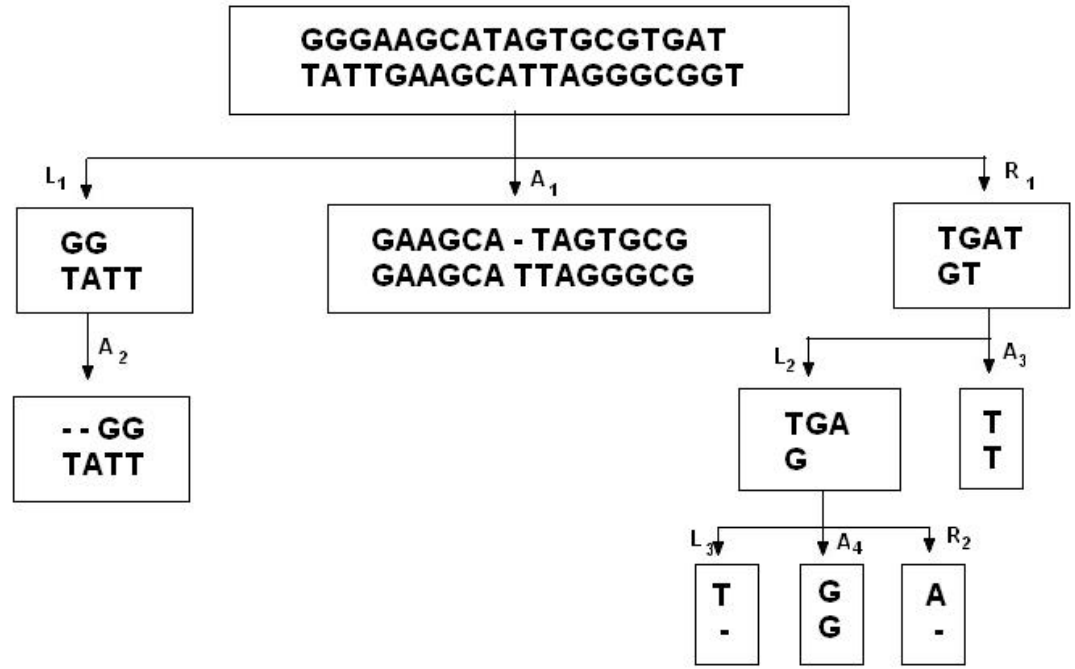
Alinhamento Local Recursivo

- Alinhar corretamente as partes bem conversadas das sequências
- Alinhamento local do par de sequências. Recursão à esquerda e à direita para as subsequências não alinhadas
- `GGGAAGCATAGTGCGTGAT` e `TATTGAAGCATTAGGGCGGT`



Alinhamento Local Recursivo

- Alinhar corretamente as partes bem conversadas das sequências
- Alinhamento local do par de sequências. Recursão à esquerda e à direita para as subsequências não alinhadas
- `GGGAAGCATAGTGCGTGAT` e `TATTGAAGCATTAGGGCGGT`



Pontuação

- Linear: $w(k) = gap \times k$
- Afim: $w(k) = gop + gep \times k$
- Logarítmica: $w(k) = gop + c \times \log(k)$
- Modelos de Substituição
  - PAM e BLOSUM
  - DCMut: Adaptação PAM
  - JTT: Jones-Taylor-Thornton
  - PMB: Matriz de Probabilidade de Blocos
  - Modelo de Categorias: adaptação do modelo de dois parâmetros de Kimura para proteínas

Alinhamento Múltiplo de Sequências

- Encontrar características conservadas de famílias de proteínas
- Escolha das sequências, escolha da função objetivo e otimização da função objetivo
- Uma função objetivo matematicamente perfeita geralmente não é biologicamente perfeita
- Problema NP-Completo (Lusheng Wang e Tao Jiang, 1994)
- Heurísticas
  - Progressiva
  - Iterativa
  - Consistência
  - Consenso
  - Modelos
  - Blocos

Alinhamento Múltiplo Progressivo

- Três etapas principais:
  - Computação da matriz de distâncias de pares de sequências
  - Construção da árvore guia
  - Geração do alinhamento múltiplo
- Clustal W, MUSCLE, Pileup e MultiAlign
- Sensível aos parâmetros utilizados
- Estratégia Gulosa

Alinhadores Progressivos

- Implementamos 342 alinhadores
- Linguagem JAVA, uso da biblioteca Biojava
- Uso do PHYLIP e do R
- Combinação dos métodos para realizar as etapas do alinhamento progressivo

Alinhadores Progressivos

- Implementamos 342 alinhadores
- Linguagem JAVA, uso da biblioteca Biojava
- Uso do PHYLIP e do R
- Combinação dos métodos para realizar as etapas do alinhamento progressivo

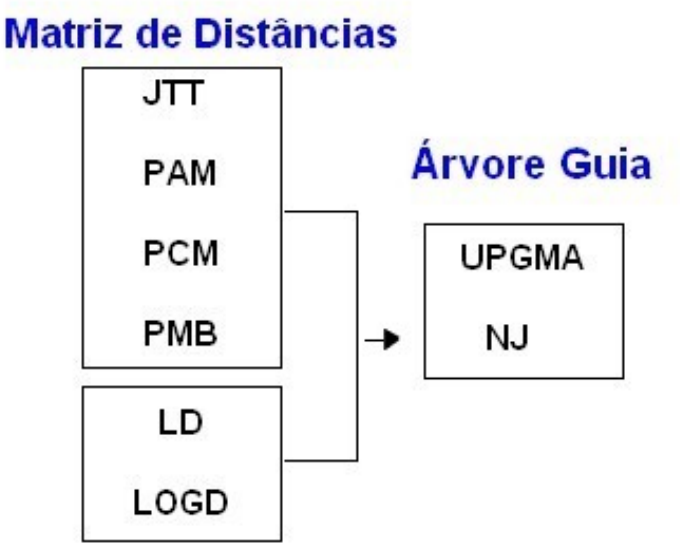
Matriz de Distâncias

JTT
PAM
PCM
PMB

LD
LOGD

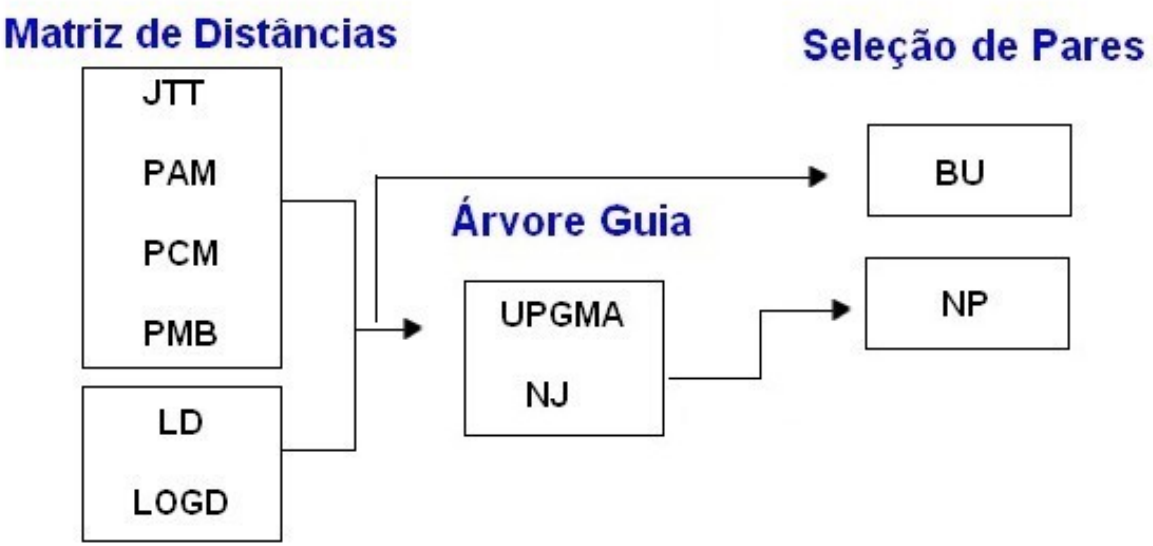
Alinhadores Progressivos

- Implementamos 342 alinhadores
- Linguagem JAVA, uso da biblioteca Biojava
- Uso do PHYLIP e do R
- Combinação dos métodos para realizar as etapas do alinhamento progressivo



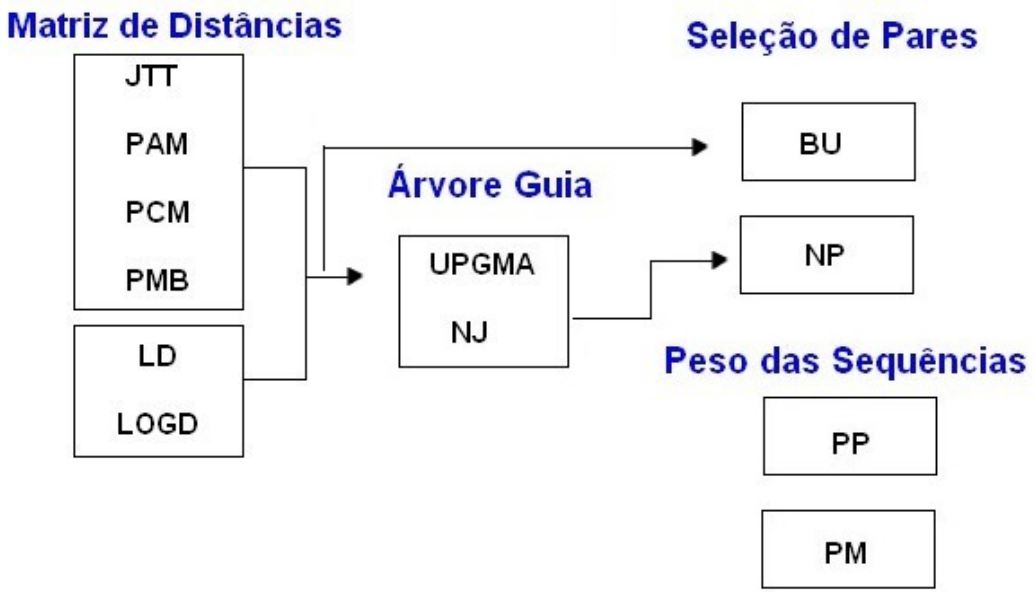
Alinhadores Progressivos

- Implementamos 342 alinhadores
- Linguagem JAVA, uso da biblioteca Biojava
- Uso do PHYLIP e do R
- Combinação dos métodos para realizar as etapas do alinhamento progressivo



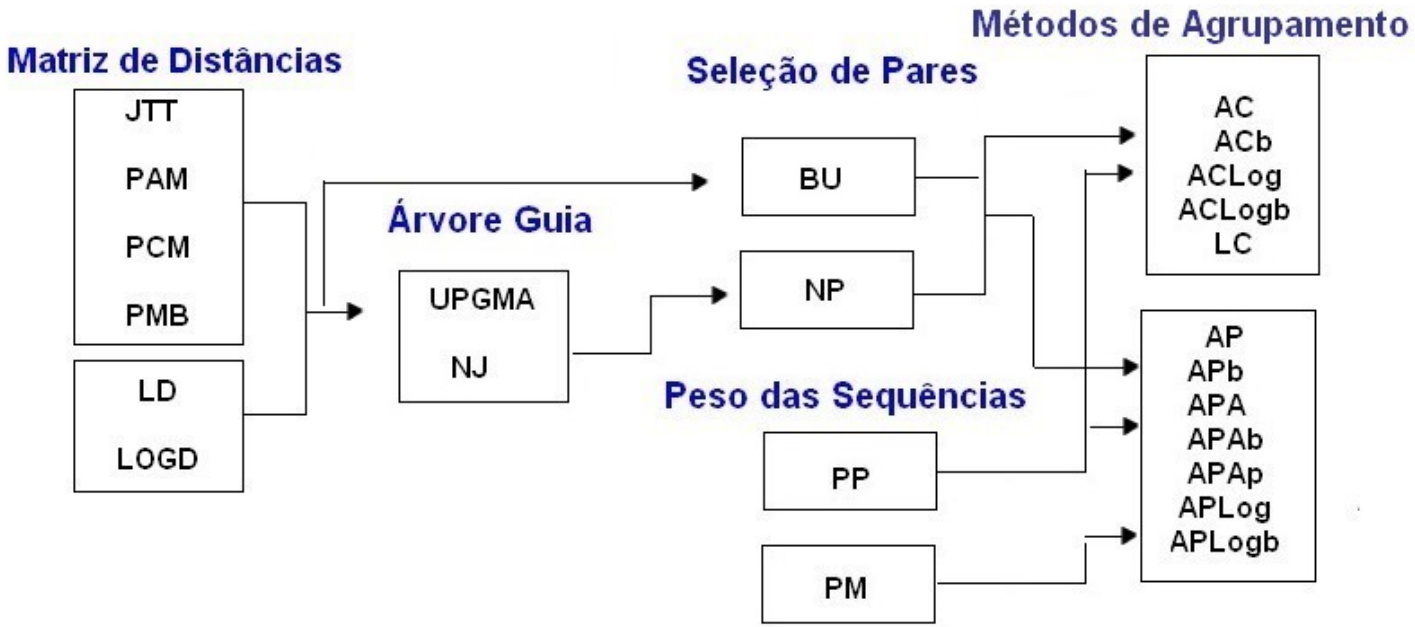
Alinhadores Progressivos

- Implementamos 342 alinhadores
- Linguagem JAVA, uso da biblioteca Biojava
- Uso do PHYLIP e do R
- Combinação dos métodos para realizar as etapas do alinhamento progressivo



Alinhadores Progressivos

- Implementamos 342 alinhadores
- Linguagem JAVA, uso da biblioteca Biojava
- Uso do PHYLIP e do R
- Combinação dos métodos para realizar as etapas do alinhamento progressivo



Agrupamento por Consenso

- O agrupamento por consenso é baseado no alinhamento dos consensos dos alinhamentos de entrada

Alinhamento 1:

```
SAPANAVAADNATAIALKYNQDATKSERVAAARPGLPPEEQHCADCQFMQADAAGATDEWKGCQLFPGKLIN
EDLPHVDAATNPIAQSLHYIEDANASERNPVTKTELPGSEQFCHNCSFIQADSGA----WRPCTLYPGYTVS
AAPLVAETDAN--AKSLGYVADTTKADK---TKYPKHTKDQSCSTCALYQ----GKTAPQGACPLFAGKEVV
MERLSED---DPAAQALEYRHDAS------SVQHPAYEEGQTCLNC-LLYTDASAQ--DWGPCSVFPGKLVS
```

Consenso 1:

```
SAPLNADAATNPTAQALHYIQDATKSERNPATKHPLPPEEQHCANCSFLQADAGGQTDDWGPCPLFPGKLVS
```

$$score(S) = score(S,S) + score(S,E) + score(S,A) + score(S,M) = 4 + 0 + 1 - 1 = 4$$

$$score(E) = score(E,S) + score(E,E) + score(E,A) + score(E,M) = 0 + 5 - 1 - 2 = 2$$

$$score(A) = score(A,S) + score(A,E) + score(A,A) + score(A,M) = 1 - 1 + 4 - 1 = 3$$

$$score(M) = score(M,S) + score(M,E) + score(M,A) + score(M,M) = -1 - 2 - 1 + 5 = 1$$

Agrupamento por Consenso

Alinhamento 2:

QDLPPLDPSAEQAQALNYVKDTAEAADHPAHQEGEQCDNCMFF-QADSQGCQL-----FPQNSVE

------EPRAEDGHAHDYVNEAADASGHPRYQEGQLCENCAFWGEAVQDGWGRCTHPDFDEVLVK

Consenso 2:

QDLPPLDPRAEDGHAHNYVNDTADAADHPRHQEGQQCDNCMFWGQADQDGWGRCTHPDFPQNLVE

Alinhamento dos consensos foi o seguinte:

SAPLNADAATNPTAQALHYIQDATKSERNPATKHPLPPEEQHCANCSFLQADAGGQTDDWGPC--PLFPGKLVS

QDLPPLDPRAE-DGHAHNYVNDTA-----DAADHPRHQEGQQCDNCMFW---GQADQDGWGRCTHPDFPQNLVE

União dos alinhamentos:

SAPANAVAADNATAIALKYNQDATKSERVAAARPGLPPEEQHCADCQFMQADAAGATDEWKGC--QLFPGKLIN
EDLPHVDAATNPIAQSLHYIEDANASERNPVTKTELPGSEQFCHNCSFIQADSGA----WRPC--TLYPGYTVS
AAPLVAETDAN--AKSLGYVADTTKADK---TKYPKHTKDQSCSTCALYQ----GKTAPQGAC--PLFAGKEVV
MERLSED---DPAAQALEYRHDAS------SVQHPAYEEGQTCLNC-LLYTDASAQ--DWGPC--SVFPGKLVS
QDLPPLDPSAE-QAQALNYVKDTA-----EAADHPAHQEGEQCDNCMFF----QADSQGCQL-----FPQNSVE
------EPRAE-DGHAHDYVNEAA-----DASGHPRYQEGQLCENCAFW---GEAVQDGWGRCTHPDFDEVLVK

Agrupamento por Perfil

- Alinha dois alinhamentos de entrada com base na pontuação dos pares de colunas.

```
SAPANAVAADNATAIALKYNQDATKSERVAAARPGLPPEEQHCADCQFMQADAAG
EDLPHVDAATNPIAQSLHYIEDANASERNPVTKTELPGSEQFCHNCSFIQADSGA
AAPLVAETDAN--AKSLGYVADTTKADK---TKYPKHTKDQSCSTCALYQ----G
MERLSED---DPAAQALEYRHDAS------SVQHPAYEEGQTCLNC-LLYTDASA
e,
QDLPPLDPSAEQAQALNYVKDTAEAADHPAHQEGEQCDNCMFF-QADSQG

------EPRAEDGHAHDYVNEAADASGHPRYQEGQLCENCAFWGEAVQDG
```

Agrupamento por Perfil

- Alinhamento da décima segunda coluna do primeiro alinhamento com a décima segunda do segundo alinhamento, respectivamente $A, P, -, P$ e Q, D, cuja pontuação é:
  $score(A, Q) + score(A, D) + 2 \times score(P, Q) + 2 \times score(P, D) +$
  $score(-, Q) + score(-, D) = -1 - 2 + (2 \times -1) + (2 \times -1) - 8 - 8 = -23$
- Alinhamento da décima segunda coluna do primeiro com *gaps*:
  $6 \times gap + 2 \times score(gap, gap) = -48 + 0 = -48$.
- Alinhamento da décima segunda coluna do segundo com uma coluna de gaps: $8 \times gap = -64$.

```
SAPANAVAADNATAIALKYNQDATKSERVAAARPGLPPEEQHCADCQFM-QADAAG
EDLPHVDAATNPIAQSLHYIEDANASERNPVTKTELPGSEQFCHNCSFI-QADSGA
AAPLVAETDAN--AKSLGYVADTTKADK---TKYPKHTKDQSCSTCALY-Q----G
MERLSED---DPAAQALEYRHDAS------SVQHPAYEEGQTCLNC-LL-YTDASA
QDLPPLDPSAEQ-AQALNYVKDTA--E---AADHPAHQEGEQCDNCMFF-QADSQG
------EPRAED-GHAHDYVNEAA--D---ASGHPRYQEGQLCENCAFWGEAVQDG
```

Ajustes de Parâmetros

- Alta sensibilidade do alinhador de perfil em relação aos parâmetros de entrada
  - Matriz substituição: escolha pela similaridade das sequências
  - gop: Considera média dos valores excluindo a diagonal da matriz e tamanho das sequências
  - gep: Considera tamanho das sequências

BAliBASE

- Banco de dados de alinhamentos múltiplos de sequências manualmente refinado
- Programa *bali_score*, calcula pontuações de soma-de-pares (SP) e total de colunas (TC)
- Versão 3.0 do BAliBASE, possui sequências divididas em 5 conjuntos

Escolha de Parâmetros

Alinhamento Local Recursivo

- Porcentagem dos *core blocks* corretamente alinhados
- Combinação de todos os pares de sequências de cada um dos dez maiores arquivos de cada conjunto de referência do BAliBASE, totalizando 2903 pares.

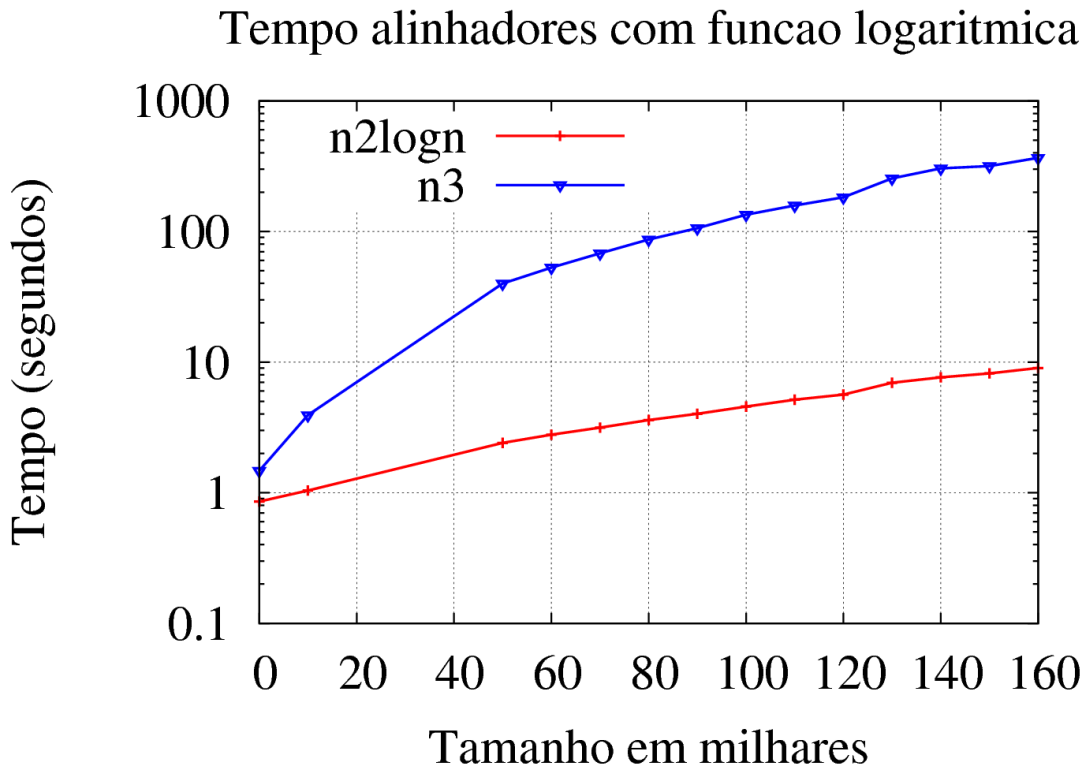
Tamanho	Média *mm*	Média *mM*	Média *MM*	Média *Total*
0	65,22	56,88	57,65	58,60
50	68,88	61,88	62,20	63,09
100	**71,19**	**65,09**	**65,72**	**66,29**
150	70,52	63,54	64,78	65,05
200	69,54	62,87	64,42	64,23
Global	66,79	57,83	59,02	59,16

Escolha de Parâmetros para Alinhamento de Perfil

- 5 grupos de sequências, cada um com seis conjuntos de sequências
- Média da pontuação de soma de pares em regiões *core blocks* contra os alinhamentos de referência do BAliBASE
- Pontuação Linear:
  Global: deve utilizar BLOSUM 62 e *gap* -5
  Semi-Global: deve usar BLOSUM 45 e *gap* -2
- Pontuação Afim:
  Global: BLOSUM 55, *gop* -17 e *gep* -1
  Semi-Global: BLOSUM 45, *gop* -10 e *gep* -1
- Pontuação Logarítmica:
  Global: qualquer BLOSUM, exceto a 45, gop = -9 e c = -4
  Semi-Global: qualquer BLOSUM, com gop = -8 e c = -5

Algoritmo para Pontuação Logarítmica

Alinhamento de 237 pares de sequências, combinações de pares de sequência dos conjuntos BB11001.tfa, BB12020.tfa, BB20020.tfa e BB30017.tfa do BAliBASE



Testes

- Entrada:
  RVS1, arquivos com menor número de sequências de cada um dos seis conjuntos do BAliBASE
  RVS2, arquivos com segundo menor número de sequências de cada um dos seis conjuntos do BAliBASE.
- Alinhadores:
  Grupo 1: todos os alinhadores. Tiveram como entrada o RVS1
  Grupo 2: 195 alinhadores, excluindo os que utilizam pontuação logarítmica. Tiveram como entrada RVS1 e RVS2
- Máquina Intel Core 2 Duo de 2.33GHz e 3GB de memória

Tempo de Execução

Escolha pelas menores sequências de cada conjunto do BAliBASE devido ao alto tempo de processamento necessário para executar os 342 alinhadores implementados

Método de Agrupamento	tempo RVS1	tempo RVS2
AC	187,74	401,58
ACb	220,08	39,11
ACLog	192,33	-
ACLogb	191,45	-
LC	92,18	356,35
AP	232,77	591,70
APb	143,80	152,20
APA	246,65	327,21
APAb	131,74	58,68
APLog	4013,65	-
APLogb	3681,00	-
APAp	129,70	994,80

Pontuações do Grupo 1

	Método	Mínimo	Máximo	Média	Mediana
Distância	JTT	52,40	82,87	70,44	71,53
	PAM	49,87	82,13	69,79	70,87
	PCM	52,50	83,40	70,47	70,73
	PMB	52,33	82,63	70,42	71,67
	LD	47,27	76,88	61,26	61,58
	LOGD	39,07	73,22	54,93	52,73
Árvore	NJ	48,37	82,72	68,14	70,53
	UP	39,07	83,40	65,33	67,00
Seleção de Pares	BU	47,02	82,08	65,99	65,92
	NP	39,07	83,40	66,71	67,95
Agrupamento	AC	43,88	75,70	67,52	70,89
	ACb	43,88	58,30	53,27	53,72
	ACLog	58,88	74,50	69,04	72,19
	ACLogb	57,17	72,37	66,71	66,59
	LC	57,18	68,18	62,19	61,92
	AP	39,75	71,65	64,04	68,75
	APb	39,75	57,83	53,17	54,50
	APA	53,65	83,40	77,48	80,81
	APAb	53,65	82,87	73,62	76,44
	APAp	47,43	82,08	72,13	79,19
	APLog	39,07	68,17	62,79	65,78
	APLogb	39,07	75,77	68,81	72,27
Média das Distâncias	PM	45,43	82,63	68,22	69,47
	PP	39,07	83,40	65,70	66,90

Pontuações do Grupo 2

	Método	Mínimo	Máximo	Média	Mediana
Distância	JTT	42,22	67,39	55,72	55,00
	PAM	42,26	66,84	55,82	59,01
	PCM	41,82	67,11	55,72	56,38
	PMB	42,08	66,16	55,50	55,10
	LD	31,26	62,21	44,50	44,41
Árvore	NJ	32,77	65,29	52,66	52,92
	UP	31,26	67,39	53,26	52,85
Seleção de Pares	BU	36,14	67,11	52,00	51,66
	NP	31,26	67,39	52,96	52,92
Agrupamento	AC	48,18	56,73	53,76	54,29
	ACb	41,59	48,81	45,33	45,56
	LC	44,61	52,65	49,37	49,01
	AP	40,62	53,97	50,40	51,76
	APb	31,26	44,78	41,33	42,53
	APA	54,23	65,29	62,95	63,67
	APAb	42,88	67,39	60,70	63,34
	APAp	32,87	64,83	52,04	60,88
Média das Distâncias	PM	31,49	67,07	53,93	53,71
	PP	31,26	67,39	51,93	51,67

Melhores Alinhadores do Grupo 1

Alinhador	SP	TC	MD	AG	SP	MA	PS
125	83,40	64,67	PCM	UP	NP	APA	PP
053b	82,87	64,00	JTT	UP	NP	APAb	PP
137	82,72	64,50	PCM	NJ	NP	APA	PP
53	82,67	64,83	JTT	UP	NP	APA	PP
077b	82,63	63,00	PMB	UP	NP	APAb	PP
77	82,43	64,17	PMB	UP	NP	APA	PP
114	82,13	63,50	PAM	NJ	NP	APA	PM
113	82,08	64,17	PAM	NJ	NP	APA	PP
59p	82,08	64,50	JTT	–	BU	APA	PP
83p	82,05	65,50	PMB	–	BU	APA	PP
65	82,03	64,17	JTT	NJ	NP	APA	PP

- Destaque para agrupamento de perfil com pontuação afim ou agrupamento de perfil semi-global com pontuação afim
- Par mais próximo é superior ao Bloco Único
- Melhor desempenho quando média das distâncias desativada

Melhores Alinhadores do Grupo 2

Alinhador	SP	TC	MD	AG	SP	MA	PS
53b	67,39	43,42	JTT	UP	NP	APAb	PP
131b	67,11	42,33	PCM	-	BU	APAb	PP
132b	67,07	41,42	PCM	-	BU	APAb	PM
101b	66,84	42,58	PAM	UP	NP	APAb	PP
77b	66,16	42,00	PMB	UP	NP	APAb	PP
83b	65,70	40,08	PMB	-	BU	APAb	PP
60b	65,47	41,25	JTT	-	BU	APAb	PM
66	65,29	42,92	JTT	NJ	NP	APA	PM
138	65,24	42,25	PCM	NJ	NP	APA	PM
90	65,22	42,58	PMB	NJ	NP	APA	PM

- Alinhadores de perfil com pontuação afim, e os de perfil semi-global com pontuação afim foram os melhores
- Métodos para determinar de distância apresentam-se bem distribuídos

Piores Alinhadores do Grupo 1

Alinhador	SP	TC	MD	AG	SP	MA	PS
329	47,02	6,83	LOGD	–	BU	AC	PP
329b	47,02	6,83	LOGD	–	BU	ACb	PP
324	45,43	19,00	LOGD	UP	NP	AP	PM
324b	45,43	19,00	LOGD	UP	NP	APb	PM
321	43,88	7,17	LOGD	UP	NP	AC	PP
321b	43,88	7,17	LOGD	UP	NP	ACb	PP
323	39,75	13,83	LOGD	UP	NP	AP	PP
323b	39,75	13,83	LOGD	UP	NP	APb	PP
327	39,07	26,83	LOGD	UP	NP	APLog	PP
327b	39,07	26,83	LOGD	UP	NP	APLogb	PP

- Baixo desempenho da distância logarítmica e do UPGMA

Piores Alinhadores do Grupo 2

Alinhador	SP	TC	MD	AG	SP	MA	PS
153	40,68	18,08	LD	–	BU	AP	PP
147	40,62	18,08	LD	UP	NP	AP	PP
153b	39,14	15,64	LD	–	BU	APb	PP
162p	38,13	18,67	LD	NJ	NP	APA	PM
161p	37,98	19,5	LD	NJ	NP	APA	PP
154b	37,32	14,83	LD	–	BU	APb	PM
160b	35,65	14,83	LD	NJ	NP	APb	PM
159b	32,77	14,58	LD	NJ	NP	APb	PP
148b	31,49	13,08	LD	UP	NP	APb	PM
147b	31,26	12,83	LD	UP	NP	APb	PP

- Baixo desempenho da distância local aliada ao alinhamento de perfil semi-global

Avaliação dos Alinhadores

- Alinhamento local recursivo melhorou com a inclusão do limite mínimo de tamanho para o alinhamento local, 7,35% para a pontuação SP, mas não superou o alinhamento de perfil
- Ajustes de Parâmetros proporcionaram um pequeno aumento de pontuação para as entradas individualmente.

Conclusões

- Estudo dos métodos para construção do Alinhamento Progressivo
- Estudo empírico para determinar os parâmetros
- Implementação e avaliação de 342 alinhadores progressivos
- Agrupamento por perfil com pontuação afim obtiveram melhores desempenhos.
- Destaque para os alinhadores de consenso com pontuação logarítmica.
- Os métodos do PHYLIP para determinar a matriz de distâncias, os métodos para construção da árvore e os métodos para seleção de pares, apresentaram-se bem equilibrados

Trabalhos Futuros

- Utilização de técnicas como o *bootstrapping* para melhorar a árvore guia
- Estudo em relação à penalidade de alinhamentos de pares de *gaps*
- Uso de estratégias iterativas para refinar os alinhamentos produzidos. Alinhamentos produzidos para cada arquivo de entrada podem ser usados como população para os algoritmos genéticos.

Alinhamento Múltiplo Progressivo de Sequências de Proteínas

Maria Angélica Lopes de Souza
Orientador: Zanoni Dias

23 de julho de 2010