Novas abordagens para o problema do alinhamento múltiplo de sequências

André Atanasio Maranhão Almeida

Zanoni Dias (Orientador)

IC - Unicamp

21 de Fevereiro de 2013

Roteiro

1. Motivação
2. Alinhamento
3. Avaliação de alinhadores
4. Resultados no contexto de alinhamento progressivo
5. Resultados no contexto de alinhamento baseado em consistência
6. Alinhamento baseado em blocos
7. Resultados no contexto de alinhamento iterativo
8. Considerações finais

Roteiro

1. Motivação
2. Alinhamento
3. Avaliação de alinhadores
4. Resultados no contexto de alinhamento progressivo
5. Resultados no contexto de alinhamento baseado em consistência
6. Alinhamento baseado em blocos
7. Resultados no contexto de alinhamento iterativo
8. Considerações finais

Motivação

- Procedimento comum em bioinformática
  - Buscas em bases de dados de sequências
  - Estudo de função de genes
  - Identificação de restrições estruturais ou funcionais
  - Visualização do efeito da evolução em família de proteínas
  - Estudo de relacionamentos evolucionários
  - Identificação de *motifs* preservados pela evolução
  - Predição de estrutura

Roteiro

1. Motivação
2. Alinhamento
3. Avaliação de alinhadores
4. Resultados no contexto de alinhamento progressivo
5. Resultados no contexto de alinhamento baseado em consistência
6. Alinhamento baseado em blocos
7. Resultados no contexto de alinhamento iterativo
8. Considerações finais

Alinhamento

Alinhamento de sequências homólogas

Consiste na tentativa de posicionar resíduos (nucleotídeos ou aminoácidos) em colunas que derivam de um resíduo de um ancestral comum.

Ou seja... Modelo hipotético de mutações ao longo da evolução.

Como? Pela introdução de *gaps*, que representam *indels*.

Melhor? Aquele mais parecido com um cenário evolucionário.

Alinhamento de par

- Quando se alinham duas sequências
- Existe algoritmo ótimo $O(mn)$ para este caso:
  - Needleman e Wunsch 1970 (global)
- Existem heurísticas também:
  - FASTA
  - BLAST

Alinhamento múltiplo de sequências (MSA)

- Quando três ou mais sequências são simultaneamente alinhadas
- NP-Difícil
- Abordagens:
  - Adaptação de Needleman e Wunsch 1970
  - Algoritmo de aproximação
  - Heurística

MSA: Exemplo entrada (FASTA)

```
>1aab_
GKGDPKKPRGKMSSYAFFVQTSREEHKKKHPDASVNFSEFSKKCSERWKT
MSAKEKGKFEDMAKADKARYEREMKTYIPPKGE
>1j46_A
MQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAE
KWPFFQEAQKLQAMHREKYPNYKYRPRRKAKMLPK
>1k99_A
MKKLKKHPDFPKKPLTPYFRFFMEKRAKYAKLHPEMSNLDLTKILSKKYK
ELPEKKKMKYIQDFQREKQEFERNLARFREDHPDLIQNAKK
>2lef_A
MHIKKPLNAFMLYMKEMRANVVAESTLKESAAINQILGRRWHALSREEQA
KYYELARKERQLHMQLYPGWSARDNYGKKKKRKREK
```

RV11/BB11001.tfa

MSA: Exemplo saída (MSF)

```
1j46_A      ------MQDR VKRPMNAFIV WSRDQRRKMA LENPRMRN-- SEISKQLGYQ
2lef_A      --------MH IKKPLNAFML YMKEMRANVV AESTLKES-- AAINQILGRR
1k99_A      MKKLKKHPDF PKKPLTPYFR FFMEKRAKYA KLHPEMSN-- LDLTKILSKK
1aab_       ---GKGDPKK PRGKMSSYAF FVQTSREEHK KKHPDASVNF SEFSKKCSER

1j46_A      WKMLTEAEKW PFFQEAQKLQ AMHREKYPNY KYRP---RRK AKMLPK
2lef_A      WHALSREEQA KYYELARKER QLHMQLYPGW SARDNYGKKK KRKREK
1k99_A      YKELPEKKKM KYIQDFQREK QEFERNLARF REDH---PDL IQNAKK
1aab_       WKTMSAKEKG KFEDMAKADK ARYEREMKTY IPPK---GE- ------
```

Clustal W

Abordagem: Alinhamento progressivo

- Constrói o MSA a partir de alinhamentos de pares
- Uma das maneiras mais simples e efetivas
- Pequeno requisito de tempo e memória
- Bom desempenho: sequências homólogas e bem conservadas
- Principal problema: natureza gulosa

Abordagem: Alinhamento iterativo

- Dependem de algoritmo(s) para gerar alinhamento(s) inicial(is)
- Sua tarefa é refinar o(s) alinhamento(s) através de uma série de ciclos
- Classes:
  - Não estocásticos
  - Estocásticos: HMM, SA e GA

Abordagem: Alinhamento baseado em consistência

Alinhamento ótimo

Dado um conjunto de sequências, o MSA ótimo é definido como aquele que está de acordo com a maioria de todos os possíveis alinhamentos ótimos de pares.

- Razões para utilizar OFs baseadas em consistências:
  - Não dependem de matriz de substituição
  - Dependente de posição
  - A maioria dos alinhamentos consistentes estão frequentemente próximos da verdade

Abordagem: Alinhamento baseado em consenso

- Método cuja entrada são MSAs
- Seu objetivo é computar um MSA que seja consistente com os alinhamentos de entrada
- Exemplo: M-COFFEE

Abordagem: Alinhamento baseado em modelos

- Parte do princípio de que a evolução é mais conservativa no que se refere aos elementos estruturais das proteínas
- Fortemente recomendado para menores similaridades
- Variações:
  - Extensão estrutural: usa estruturas (PDB, por exemplo) e sobreposição de estruturas
  - Extensão por homologia: usa perfil (PSI-BLAST, por exemplo) no lugar de estruturas

Abordagem: Alinhamento baseado em blocos

Bloco

Alinhamento de fragmentos de sequências, ou seja, alinhamento local.

- Utiliza blocos, vistos como âncoras, para guiar o alinhamento
- Objetiva diminuir a dependência dos parâmetros para penalidade de *gaps*

Exemplos de alinhadores múltiplos

Alinhador	Abordagem	Ano
Clustal W	Progressivo	1994
MAFFT	Progressivo	2005
MUSCLE	Progressivo	2004
PRRP	Iterativo não estocástico	1996
SAGA	Iterativo estocástico	1996
T-COFFEE	Baseado em Consistência	2000
ProbCons	Baseado em Consistência	2005
MUMMALS	Baseado em Consistência	2006
M-COFFEE	Baseado em Consenso	2006
3D-COFFEE	Baseado em Modelos	2004
DbClustal	Baseado em Modelos	2000
DiAlign	Baseado em Blocos	1996

Roteiro

1. Motivação
2. Alinhamento
3. Avaliação de alinhadores
4. Resultados no contexto de alinhamento progressivo
5. Resultados no contexto de alinhamento baseado em consistência
6. Alinhamento baseado em blocos
7. Resultados no contexto de alinhamento iterativo
8. Considerações finais

Avaliação de alinhadores

- Diversas ferramentas para *benchmark*:
  - BAliBASE: baseado em alinhamentos pré-compilados
  - SABmark: baseado em alinhamentos pré-compilados
  - IRMbase: baseado em alinhamentos pré-compilados
  - HOMSTRAD: baseado em alinhamentos estruturais automaticamente gerados
  - PREFAB: baseado em alinhamentos estruturais automaticamente gerados
  - APDB: baseado em alinhamentos estruturais, mas não possui uma base de alinhamentos

BAliBASE 3.0

- Primeiro construído para *benchmarking* em larga escala
- Refino manual dos alinhamentos
- Subdivisão dos conjuntos de referência
  - Sequências equidistantes (RV11 e RV12)
  - Sequências órfãs (RV20)
  - Grupos de sequências (RV30)
  - Sequências com extensões longas nas extremidades (RV40)
  - Sequências com inserções internas longas (RV50)

BAliBASE 3.0

- Pontuação:
  - SP: porcentagem de pares de bases corretamente alinhados
  - TC: porcentagem de colunas corretamente alinhadas
- Alinhamentos:
  - 386 entradas: 218 completas e 168 de regiões homólogas
  - Pode-se pontuar inteiro ou apenas *core blocks*

Desempenho de alinhadores conhecidos

Alinhador	Tempo (s)	SP	TC
ProbCons 1.12	21.978,50	86,38	55,66
T-COFFEE 8.14	14.631,89	86,11	55,46
MUMMALS 1.01	50.594,17	85,54	53,83
MUSCLE 3.7	1.636,36	82,19	47,59
DiAlign 2.2	7.286,96	77,49	41,52
Clustal W 2.0.10	2.040,10	75,36	37,38

Roteiro

1. Motivação
2. Alinhamento
3. Avaliação de alinhadores
4. Resultados no contexto de alinhamento progressivo
5. Resultados no contexto de alinhamento baseado em consistência
6. Alinhamento baseado em blocos
7. Resultados no contexto de alinhamento iterativo
8. Considerações finais

Resultados no contexto de alinhamento progressivo

- Construção de MSAs a partir de alinhamentos de pares
- Etapas:
  - Computação da matriz de distâncias
  - Geração da árvore guia
  - Construção do MSA
- Implementamos 342 alinhadores progressivos

Computação da matriz de distâncias

- Modelo PAM
- Modelo PMB
- Modelo das Categorias (PCM)
- Modelo Jones-Taylor-Thornton (JTT)
- Distância local recursiva (LD)
- Função logarítmica para penalização de *gaps* (LOGD)

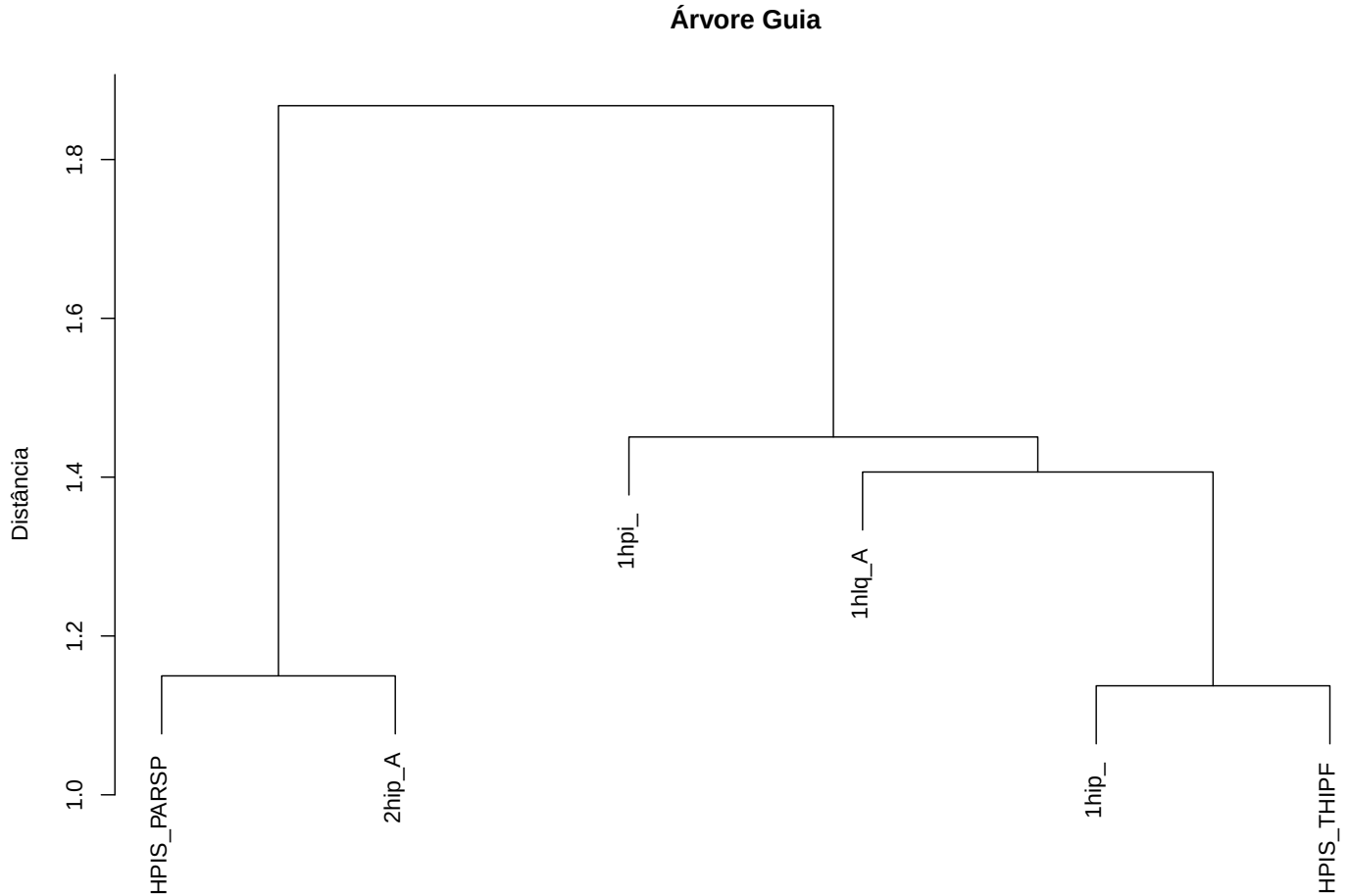
Exemplo matriz de distâncias (PAM)

	1j46_A	2lef_A	1k99_A	1aab_
1j46_A	0,000000	1,733140	2,496830	3,469881
2lef_A	1,733140	0,000000	2,666700	3,266899
1k99_A	2,496830	2,666700	0,000000	2,224336
1aab_	3,469881	3,266899	2,224336	0,000000

Geração da árvore guia

- Métodos:
  - UPGMA
  - NJ

Geração da árvore guia



UPGMA para RV12/BB12021.tfa

Construção do MSA

- Seleção do par
  - Por bloco único (BU)
  - Par mais próximo (NP)
- Agrupamento
  - **Alinhamento de consensos** global/semi (AC e ACb)
    - AC local recursivo (LC)
    - AC g/s com f. log. para penalização de *gaps* (ACLog e ACLogb)
  - **Alinhamento de perfil** g/s (AP e APb)
    - AP g/s com função afim (APA e APAb)
    - AP g/s com f. log. para penalização de *gaps* (APLog e APLogb)
    - AP com ajuste automático de parâmetros (APAp)
- Esquema de pesos (PM)

Alinhamento de consensos

```
======================== 1° alinhamento e consenso ==================================

SAPANAVAADNATAIALKYNQDATKSERVAAARPGLPPEEQHCADCQFMQADAAGATDEWKGCQLFPGKLINVNGWCASWTLKAG
EDLPHVDAATNPIAQSLHYIEDANASERNPVTKTELPGSEQFCHNCSFIQADSGA----WRPCTLYPGYTVSEDGWCLSWAHKTA
AAPLVAETDAN--AKSLGYVADTTKADK---TKYPKHTKDQSCSTCALYQ----GKTAPQGACPLFAGKEVVAKGWCSAWA-KKA
MERLSED---DPAAQALEYRHDAS------SVQHPAYEEGQTCLNC-LLYTDASAQ--DWGPCSVFPGKLVSANGWCTAWVAR--

SAPLNADAATNPTAQALHYIQDATKSERNPATKHPLPPEEQHCANCSFLQADAGGQTDDWGPCPLFPGKLVSANGWCTAWAHKTA

======================== 2° alinhamento e consenso ==================================

QDLPPLDPSAEQAQALNYVKDTAEAADHPAHQEGEQCDNCMFF-QADSQGCQL-----FPQNSVEPAGWCQSWTAQN
------EPRAEDGHAHDYVNEAADASGHPRYQEGQLCENCAFWGEAVQDGWGRCTHPDFDEVLVKAEGWCSVYAPAS

QDLPPLDPRAEDGHAHNYVNDTADAADHPRHQEGQQCDNCMFWGQADQDGWGRCTHPDFPQNLVEPEGWCQSWTPQN

========================== Alinhamento dos consensos ================================

SAPLNADAATNPTAQALHYIQDATKSERNPATKHPLPPEEQHCANCSFLQADAGGQTDDWGPC--PLFPGKLVSANGWCTAWAHKTA
QDLPPLDPRAE-DGHAHNYVNDTA-----DAADHPRHQEGQQCDNCMFW---GQADQDGWGRCTHPDFPQNLVEPEGWCQSWTPQN-

========================== Alinhamento resultante ===================================

SAPANAVAADNATAIALKYNQDATKSERVAAARPGLPPEEQHCADCQFMQADAAGATDEWKGC--QLFPGKLINVNGWCASWTLKAG
EDLPHVDAATNPIAQSLHYIEDANASERNPVTKTELPGSEQFCHNCSFIQADSGA----WRPC--TLYPGYTVSEDGWCLSWAHKTA
AAPLVAETDAN--AKSLGYVADTTKADK---TKYPKHTKDQSCSTCALYQ----GKTAPQGAC--PLFAGKEVVAKGWCSAWA-KKA
MERLSED---DPAAQALEYRHDAS------SVQHPAYEEGQTCLNC-LLYTDASAQ--DWGPC--SVFPGKLVSANGWCTAWVAR--
QDLPPLDPSAE-QAQALNYVKDTA-----EAADHPAHQEGEQCDNCMFF----QADSQGCQL-----FPQNSVEPAGWCQSWTAQN-
------EPRAE-DGHAHDYVNEAA-----DASGHPRYQEGQLCENCAFW---GEAVQDGWGRCTHPDFDEVLVKAEGWCSVYAPAS-
```

Alinhamento de perfis

```
============================== 1° alinhamento ====================================

SAPANAVAADNATAIALKYNQDATKSERVAAARPGLPPEEQHCADCQFMQADAAGATDEWKGCQLFPGKLINVNGWCASWTLKAG
EDLPHVDAATNPIAQSLHYIEDANASERNPVTKTELPGSEQFCHNCSFIQADSGA----WRPCTLYPGYTVSEDGWCLSWAHKTA
AAPLVAETDAN--AKSLGYVADTTKADK---TKYPKHTKDQSCSTCALYQ----GKTAPQGACPLFAGKEVVAKGWCSAWA-KKA
MERLSED---DPAAQALEYRHDAS------SVQHPAYEEGQTCLNC-LLYTDASAQ--DWGPCSVFPGKLVSANGWCTAWVAR--

============================== 2° alinhamento ====================================

QDLPPLDPSAEQAQALNYVKDTAEAADHPAHQEGEQCDNCMFF-QADSQGCQL-----FPQNSVEPAGWCQSWTAQN
------EPRAEDGHAHDYVNEAADASGHPRYQEGQLCENCAFWGEAVQDGWGRCTHPDFDEVLVKAEGWCSVYAPAS

========================== Alinhamento resultante ====================================

SAPANAVAADNATAIALKYNQDATKSERVAAARPGLPPEEQHCADCQFM-QADAAGATDEWKGCQ-L-FPGKLINVNGWCASWTLKAG
EDLPHVDAATNPIAQSLHYIEDANASERNPVTKTELPGSEQFCHNCSFI-QADSGA----WRPCT-L-YPGYTVSEDGWCLSWAHKTA
AAPLVAETDAN--AKSLGYVADTTKADK---TKYPKHTKDQSCSTCALY-Q----GKTAPQGACP-L-FAGKEVVAKGWCSAWA-KKA
MERLSED---DPAAQALEYRHDAS------SVQHPAYEEGQTCLNC-LL-YTDASAQ--DWGPCS-V-FPGKLVSANGWCTAWVAR--
QDLPPLDPSAEQ-AQALNYVKDTA--E---AADHPAHQEGEQCDNCMFF-QADSQG----CQL-----FPQNSVEPAGWCQSWTAQN-
------EPRAED-GHAHDYVNEAA--D---ASGHPRYQEGQLCENCAFWGEAVQDG----WGRCTHPDFDEVLVKAEGWCSVYAPAS-
```

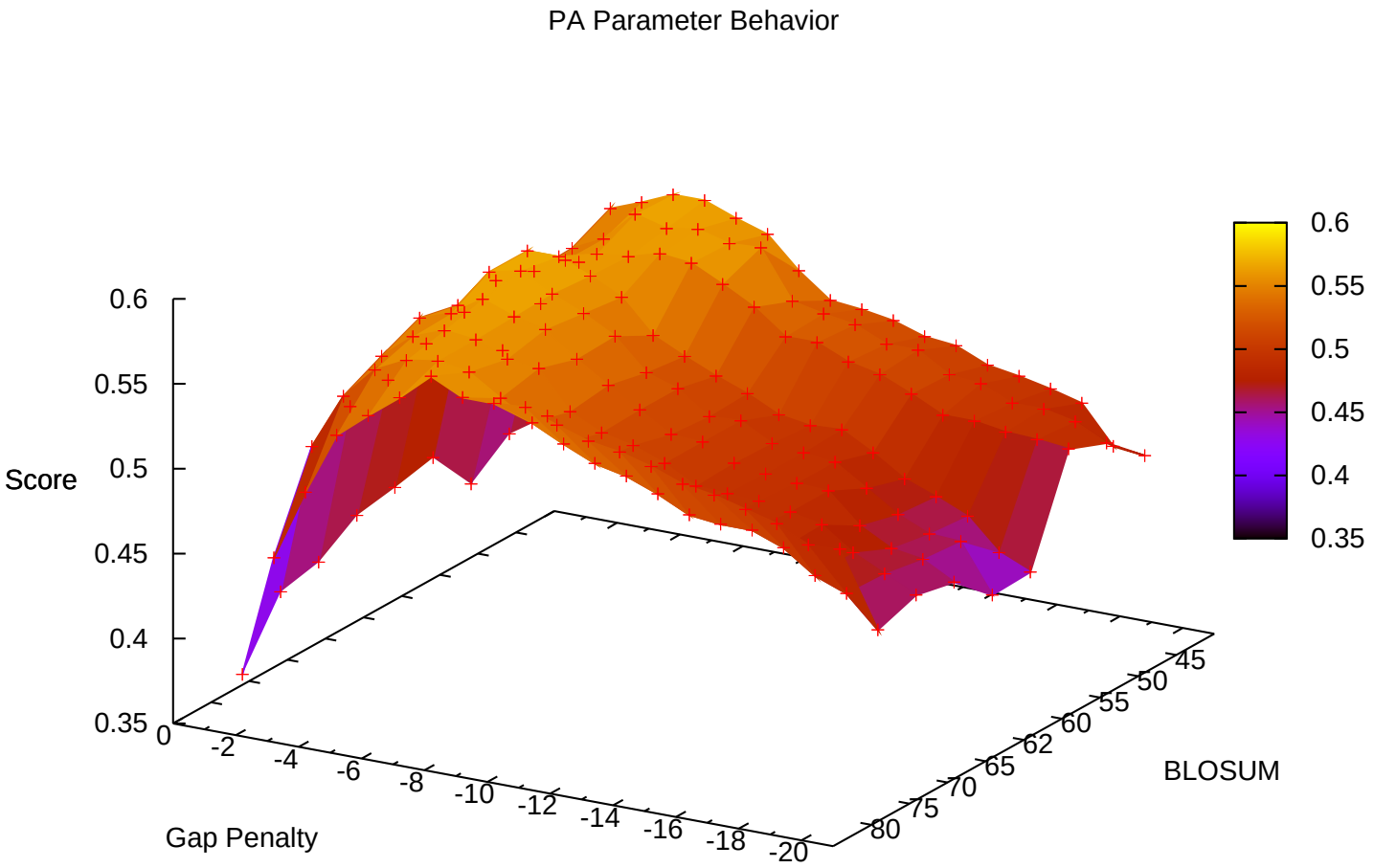
AP com ajuste automático de parâmetros

- Utiliza função afim para penalizar *gaps*
- Define matriz de substituição, *gop* e *gep* de acordo com as sequências de entrada
- Baseado no método empregado pelo Clustal W

Alinhadores progressivos

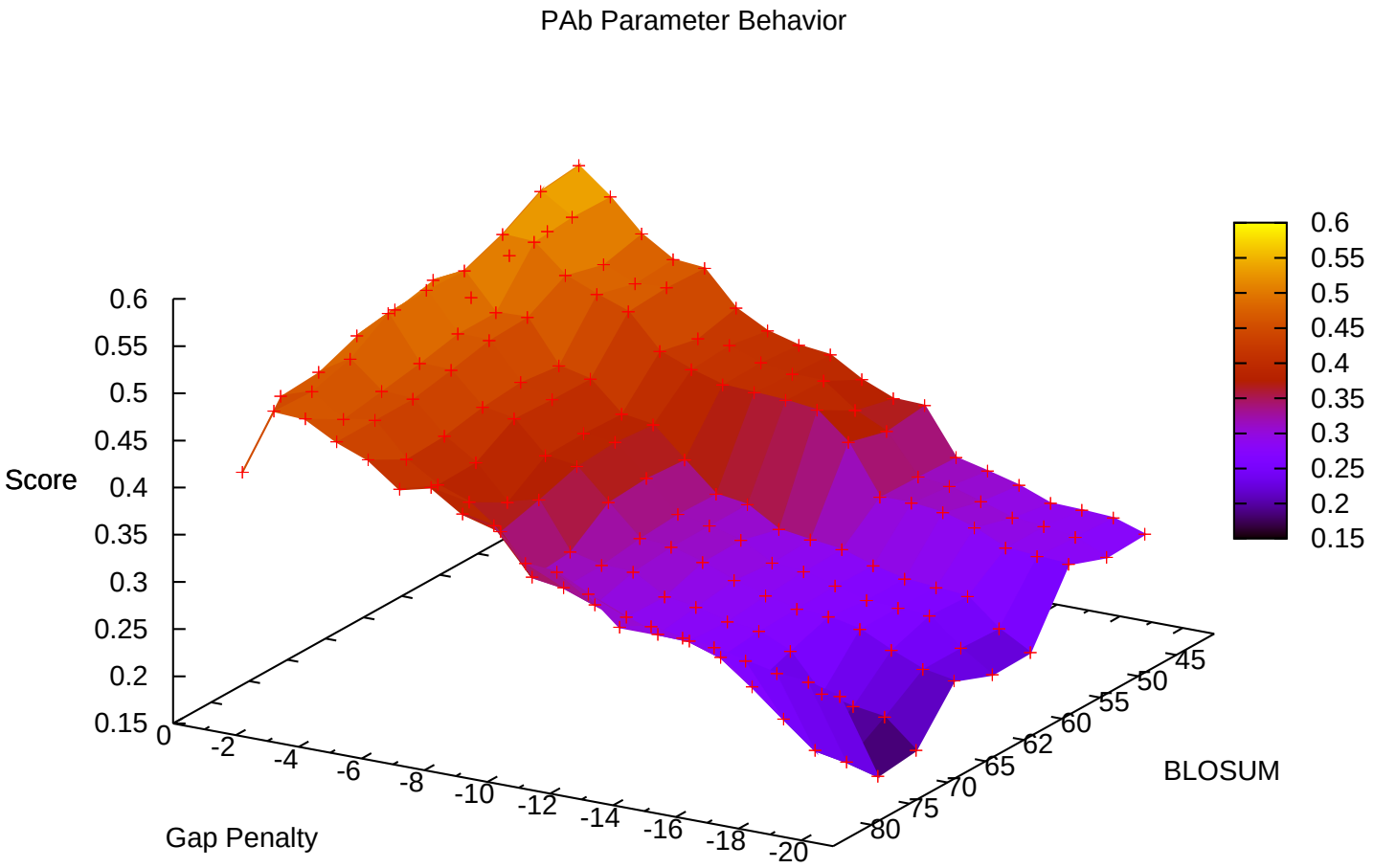
Distância	Árvore	Seleção	Agrupamento	Pesos
PAM	UP	BU	AC	PM
PMB	NJ	NP	ACb	PP
PCM			LC	
JTT			ACLog	
LD			ACLogb	
LOGD			AP	
			APb	
			APA	
			APAb	
			APLog	
			APLogb	
			APAp	
			Total:	342

Definição de parâmetros para alinhamento de perfil (AP)



BLOSUM62 e $gap = -5$

Definição de parâmetros para alinhamento de perfil (APb)



BLOSUM45 e $gap = -2$

Definição de parâmetros para alinhamento de perfil

- Com função afim para penalizar *gaps*:
  - As mesmas matrizes foram avaliadas
  - $-1 \leq gop \leq -20$
  - $-1 \leq gep \leq -10$
  - Global (APA): BLOSUM55, $gop = -17$ e $gep = -1$
  - Semi-global (APAb): BLOSUM45, $gop = -10$ e $gep = -1$

Como os testes foram realizados

- Entradas com sequências completas
- Pontuação SP apenas nos *core blocks*
- Entradas:

RVS1	RVS2
BB11001	BB11025
BB12020	BB12021
BB20020	BB20001
BB30017	BB30006
BB40032	BB40010
BB50004	BB50002

- Grupo 1: conjunto completo dos 342 alinhadores
- Grupo 2: conjunto dos 195 que não usam função logarítmica
- Grupo 1 foi executado apenas para RVS1

Resultados: Grupo 1 no RVS1

Categoria	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
Esquema de pesos	PM	45,43	82,63	68,22	69,47
	PP	39,07	83,40	65,70	66,90

Resultados: Grupo 2 no RVS1 e RVS2

Categoria	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
Esquema de pesos	PM	31,49	67,07	53,93	53,71
	PP	31,26	67,39	51,93	51,67

Resultados: melhores do Grupo 1

Alinhador	SP	TC	CD	AG	SL	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
053	82,67	64,83	JTT	UP	NP	APA	PP
077b	82,63	63,00	PMB	UP	NP	APAb	PP
077	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
059p	82,08	64,50	JTT	–	BU	APA	PP
083p	82,05	65,50	PMB	–	BU	APA	PP
065	82,03	64,17	JTT	NJ	NP	APA	PP
138	82,03	62,83	PCM	NJ	NP	APA	PM
101b	82,03	62,67	PAM	UP	NP	APAb	PP
089	81,75	63,50	PMB	NJ	NP	APA	PP
066	81,70	63,50	JTT	NJ	NP	APA	PM
060p	81,65	61,67	JTT	–	BU	APA	PM
138p	81,63	65,17	PCM	NJ	NP	APA	PM
137p	81,60	63,00	PCM	NJ	NP	APA	PP
059	81,47	64,33	JTT	–	BU	APA	PP
090	81,42	64,17	PMB	NJ	NP	APA	PM

Resultados: melhores do Grupo 2

Alinhador	SP	TC	CD	AG	SL	MA	PS
053b	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
077b	66,16	42,00	PMB	UP	NP	APAb	PP
083b	65,70	40,08	PMB	-	BU	APAb	PP
060b	65,47	41,25	JTT	-	BU	APAb	PM
066	65,29	42,92	JTT	NJ	NP	APA	PM
138	65,24	42,25	PCM	NJ	NP	APA	PM
090	65,22	42,58	PMB	NJ	NP	APA	PM
084b	65,21	39,50	PMB	-	BU	APAb	PM
125b	65,15	41,08	PCM	UP	NP	APAb	PP
059b	65,14	40,33	JTT	-	BU	APAb	PP
114	65,11	42,67	PAM	NJ	NP	APA	PM
126b	65,10	40,83	PCM	UP	NP	APAb	PM
060p	64,83	43,33	JTT	-	BU	APA	PM
126	64,68	42,25	PCM	UP	NP	APA	PM
125	64,60	40,75	PCM	UP	NP	APA	PP
108b	64,53	38,83	PAM	-	BU	APAb	PM
053	64,33	40,83	JTT	UP	NP	APA	PP

Roteiro

1. Motivação
2. Alinhamento
3. Avaliação de alinhadores
4. Resultados no contexto de alinhamento progressivo
5. Resultados no contexto de alinhamento baseado em consistência
6. Alinhamento baseado em blocos
7. Resultados no contexto de alinhamento iterativo
8. Considerações finais

Alinhamento múltiplo baseado em consistência

1. O primeiro método foi descrito por Kececioglu
2. O método foi avaliado com o uso do SAGA
3. Foi desenvolvido um método de otimização mais eficiente, T-COFFEE
4. ProbCons foi desenvolvido, utilizando consistência probabilística
5. MUMMALS foi desenvolvido, utilizando múltiplos estados de *match* descrevendo estruturas locais
6. Implementamos alterações no MUMMALS

MUMMALS

1. Computação de uma matriz de distâncias baseada no método de contagem de k-mer
2. Geração da árvore guia, usando UPGMA
3. Computação da medida de consistência probabilística
4. Construção do alinhamento progressivamente, usando a função de pontuação baseada em consistência

Alfabetos comprimidos

Alfabeto	Classes
Dayhoff(6)	AGPST,C,DENQ,FWY,HKR,ILMV
SE-B(6)	AST,CP,DEHKNQR,FWY,G,ILMV
SE-B(8)	AST,C,DHN,EKQR,FWY,G,ILMV,P
Li-A(10)	AC,DE,FWY,G,HN,IV,KQR,LM,P,ST
Li-B(10)	AST,C,DEQ,FWY,G,HN,IV,KR,LM,P
Murphy(10)	A,C,DENQ,FWY,G,H,ILMV,KR,P,ST
SE-B(10)	AST,C,DN,EQ,FY,G,HW,ILMV,KR,P
SE-V(10)	AST,C,DEN,FY,G,H,ILMV,KQR,P,W
Solis-D(10)	AM,C,DNS,EKQR,F,GP,HT,IV,LY,W
Solis-G(10)	AEFIKLMQRVW,C,D,G,H,N,P,S,T,Y
SE-B(14)	A,C,D,EQ,FY,G,H,IV,KR,LM,N,P,ST,W

```
Original:   MDPFLVLLHSVSSSLSSSELTELKYLCLCAGRVGKRKLERVQATE
Convertida: FCADFFFFEAFAAAFAAACFACFEDFBFBAAEFAEEEFCEFCAAC
```

Alterações realizadas

1. Variamos o valor do k no método de contagem k-mer. O padrão é $k = 6$. Avaliamos $3 \leq k \leq 14$
2. Variamos o alfabeto comprimido. Foram avaliados dez outros alfabetos e para cada um deles $6 \leq k \leq 10$
3. Variamos computação da matriz de distâncias e método para geração da árvore. Testamos PAM+NJ e PAM+UPGMA

No total foram implementados 89 alinhadores.

Resultados

- Foram utilizadas todas as 218 entradas de sequências completas.

Alinhador	Tempo (s)	SP	TC
MUMMALS Original	50.594	85,54	53,83
MUMMALS $k = 8$	81.017	86,27	55,70
MUMMALS SE-B(10) $k = 7$	87.692	86,70	56,52
MUMMALS PAM+NJ 0,7-1,0	135.890	86,43	56,07

- Todas as alterações alcançaram melhorias.
- O melhor alinhador reduziu erros em $7,98\%$ numa avaliação pela pontuação SP e em $5,81\%$ pela TC.
- O tempo variou de $3.311s$ a $135.967s$

Roteiro

1. Motivação
2. Alinhamento
3. Avaliação de alinhadores
4. Resultados no contexto de alinhamento progressivo
5. Resultados no contexto de alinhamento baseado em consistência
6. Alinhamento baseado em blocos
7. Resultados no contexto de alinhamento iterativo
8. Considerações finais

Alinhamento baseado em blocos

- Utilizado na geração da população inicial do GA
- Abordagens:
  - Primeira: usa janela delizante, grafos orientados e ordenação topológica
  - Segunda: recursão que constrói o MSA com base em blocos gerados a partir de *substrings* que ocorrem em um maior número de sequências

Roteiro

1. Motivação
2. Alinhamento
3. Avaliação de alinhadores
4. Resultados no contexto de alinhamento progressivo
5. Resultados no contexto de alinhamento baseado em consistência
6. Alinhamento baseado em blocos
7. Resultados no contexto de alinhamento iterativo
8. Considerações finais

Alinhamento iterativo não estocástico

- Implementamos módulos de refino:
  - R11: a cada passo da iteração: divide as sequências em dois grupos e o alinhamento em dois de acordo com estes grupos, colunas de *gaps* são removidas e reagrupa por APAb. Ao final do ciclo é mantido o melhor alinhamento. Os ciclos são interrompidos quando se falha em melhorar por 5 vezes consecutivas.
  - R12: neste sempre atualiza o alinhamento corrente, mas guarda o de melhor pontuação.

Alinhamento iterativo não estocástico

Alinhador	PO	PR11	PR12	Efeito R11	Efeito R12
053	68,62	69,90	68,85	1,87%	0,34%
053b	60,27	64,07	63,36	6,30%	5,13%
077b	61,25	63,85	65,05	4,24%	6,20%
125	68,68	68,55	69,25	-0,19%	0,83%
137	68,36	68,08	68,10	-0,41%	-0,38%
321b	45,48	60,00	57,85	31,93%	27,20%
323	40,34	39,74	41,50	-1,49%	2,88%
323b	39,81	40,23	39,59	1,06%	-0,55%
327	56,72	57,87	58,12	2,03%	2,47%
327b	45,44	50,58	49,23	11,31%	8,34%
			Média	5,67%	5,25%

Alinhamento iterativo estocástico: GA

Algoritmo 1: GA utilizado na implementação do alinhador iterativo.

Input: Seqs, MaxPopulationSize, Generations
Output: The best alignment
population $\leftarrow$ createInitialPopulation(Seqs, MaxPopulationSize)
for $n \leftarrow 1$ **to** *Generations* **do**
    breedingPopulation $\leftarrow$ selectForBreeding(population)
    population $\leftarrow$ population $\cup$ offspring(breedingPopulation)
    population $\leftarrow$ select(population, MaxPopulationSize)
return bestAlignment(population)

ALGAe: ambiente parametrizável para alinhamento múltiplo utilizando algoritmos genéticos.

GA para MSA

- Geração da população inicial
  - Seleção de pares de sequências e tipo de alinhamento g/s. (`MaxPopulationSize` indivíduos)
  - Define uma âncora e então a cada passo seleciona uma sequência, que é adicionada por alinhamento global. ($2 \times |sequences|$ indivíduos)
- Seleção dos indivíduos para reprodução
  - Roleta: todos têm oportunidade, mas aqueles mais adaptados têm maior probabilidade de serem selecionados
  - 50% é selecionada para reprodução e cada indivíduos tem 20% de chance de ser selecionado para mutação e a mesma probabilidade para cruzamento
- Corte: Seleciona os `MaxPopulationSize` indivíduos mais adaptados
- Operadores: 3 de mutação e 2 de *crossover*

Testes iniciais

- Definição da função de aptidão
  - Soma dos pares
  - Soma dos pares com função afim para penalidade de *gaps*
- Evolução da aptidão ao longo das gerações

Novos testes

- Definição da penalização para *gaps*
  - $gap \times gap$
  - *gop* e *gep*
  - SP 80,81 e TC 56,80
- Função de aptidão baseada em alinhamento de estruturas
- Métodos alternativos para geração da população inicial
- Operador baseado em consensos
  - SP 87,42 e TC 70,84

Estes resultados referem-se a RV12.

Roteiro

1. Motivação
2. Alinhamento
3. Avaliação de alinhadores
4. Resultados no contexto de alinhamento progressivo
5. Resultados no contexto de alinhamento baseado em consistência
6. Alinhamento baseado em blocos
7. Resultados no contexto de alinhamento iterativo
8. Considerações finais

Considerações finais

Alinhador	Tempo (s)	SP	TC
SE-B 10 K=7	87.692,75	86,70	56.52
ProbCons 1.12	21.978,50	86,38	55,66
T-COFFEE 8.14	14.631,89	86,11	55,46
MUMMALS 1.01	50.594,17	85,54	53,83
MUSCLE 3.7	1.636,36	82,19	47,59
DiAlign 2.2	7.286,96	77,49	41,52
Clustal W 2.0.10	2.040,10	75,36	37,38
53b	58.810,98	66,45	26,71

Trabalhos futuros

- Novos testes usando o ALGAe:
  - Avaliar novos operadores baseados em alinhamentos estruturais
  - Avaliar novos métodos para geração da população inicial
- Alinhamento baseado em consistência
- Alinhamento baseado em modelo
- Desenvolvimento de ferramentas para visualização de MSAs

Publicações

A. Almeida and Z. Dias.
Improvements to a multiple protein sequence alignment tool.
In *International Conference on Bioinformatics Models, Methods and Algorithms*, pages 226–233, Vilamoura, Portugal, 2012.

A. Almeida, M. Souza, and Z. Dias.
Progressive multiple protein sequence alignment.
In *6th International Symposium on Bioinformatics Research and Applications - Short Abstracts*, pages 102–105, Storrs, CT, USA, 2010.

S. Ordine, A. Almeida, and Z. Dias.
An empirical study for gap penalty score using a multiple sequence alignment genetic algorithm.
In *Brazilian Symposium on Bioinformatics 2012 Digital Proceedings*, pages 108–113, 2012.

S. Ordine, A. Grilo, A. Almeida, and Z. Dias.
ALGAe:a test-bench environment for a genetic algorithm-based multiple sequence aligner.
In *Brazilian Symposium on Bioinformatics 2011 Digital Proceedings*, pages 57–60, 2011.