Heurísticas para Problemas de Rearranjo de Genomas

Ulisses Martins Dias

Zanoni Dias (Orientador)

Campinas, 03 de Abril de 2009

Roteiro

- Motivação
- Conceitos
- Descrição do Projeto
- Estágio no Exterior
- Cronograma de Atividades
- Resultados Parciais

Motivação

- Processo Evolutivo
- Como relacionar espécies?
  - Características observáveis
  - Dados genéticos
- Rearranjo de Genomas
  - Estuda mutações que afetam largos trechos de DNA
  - Alguns rearranjos geram sequüências com praticamente as mesmas características do original
  - A distância de edição classificaria esses genomas como muito divergentes
  - Comuns em plantas, mamíferos, vírus e bactérias

Conceitos

- Representação do Genoma
  - $\pi = (\pi_1\ \pi_2\ \ldots\ \pi_n)$
- Reversão: operação em que a ordem de um segmento da permutação é invertida.
  - Seja $\rho(i,j)$, $i < j$, uma reversão no segmento $[i,j]$.
  - $\rho(i,j)\pi = (\pi_1\ \ldots\ \pi_{i-1}\ \pi_j\ \pi_{j-1}\ \ldots\ \pi_{i+1}\ \pi_i\ \pi_{j+1}\ \ldots\ \pi_n)$
- Transposição: operação em que um segmento é cortado da permutação e colado em uma posição diferente.
  - Seja $\rho(i,j,k)$, $1 \leq i < j < k \leq n+1$, uma transposição nos segmentos $[i,j-1]$ e $[j,k-1]$.
  - $\rho(i,j,k)\pi = (\pi_1\ \ldots\ \pi_{i-1}\ \pi_j\ \ldots\ \pi_{k-1}\ \pi_i\ \ldots\ \pi_{j-1}\ \pi_k\ \ldots\ \pi_n)$.

Reversão

- Orientação do Gene não conhecida
  - Problema NP-Completo
    - Caprara (1999)
  - Fator de aproximação 1.375
    - Berman, Hannenhalli e Karpinski (2002)
- Orientação do Gene conhecida
  - Algoritmo polinomial
    - $O(n^4)$ - Hannenhali e Pevzner (1995)
    - $O(n^2)$ - Bergeron (2001)
    - $O(n\sqrt{n \log n})$ - Tannier e Sagot (2004)
  - Cálculo de $d(\pi)$
    - $O(n)$ - Bader, Moret e Yan (2001)

Transposição

- Ordenação por Transposições
  - Problema em aberto
- Algoritmos de aproximação
  - Bafna e Pevzner (1995)
    - Fator 1.5 - $O(n^2)$
  - Christie (1998)
    - Fator 1.5 - $O(n^4)$
  - Walter, Dias e Meidanis (2000)
    - Fator 2.25 - $O(n^2)$
  - Elias e Hartman (2005)
    - Fator 1.375 - $O(n^2)$

Transposição

- Problemas relacionados
  - Troca de pares de elementos adjacentes (Jerrum, 1985).
  - Um dos blocos possui tamanho 1 (Heath e Vergara, 1998)
  - Transposição de prefixos (Dias e Meidanis, 2002)
  - Os blocos não precisam ser adjacentes (Christie, 1996)
- Classes polinomiais
  - Fortuna, 2005
  - Labarre, 2006

Descrição do projeto

- Heurísticas
  - Programação Linear Inteira
  - Programação em Lógica com Restrições
  - Melhorias aplicadas ao algoritmo de Bafna e Pevzner (1998)
- Classes de permutações
  - Permutações que atingem o limitante inferior de *breakpoint*
    - Seria a maior classe polinomial já estudada
  - Permutações calculadas de forma incorreta pelo algoritmo de Bafna e Pevzner (1998)
- Objetivos secundários
  - Estudar o problema do diâmetro de transposição
  - Estudar novos algoritmos de aproximação
  - Fornecer novas ferramentas

Estágio no exterior

- Estágio no *Virginia Tech* - Estados Unidos
  - Professor Dr. João Carlos Setubal
- Participação no projeto de pesquisa para reconstrução de genomas ancestrais para bactérias da ordem *Rhizobiales*
- Contribuição com estudo de métodos baseados em Rearranjo de Genomas.
  1. Matriz de distância
  2. Estimativas estatísticas
  3. Rearranjo múltiplos de genomas
  4. Estudo de métodos específicos para o grupo de bactérias em questão.

Cronograma de atividades

Etapas Iniciais

- Etapa 1:
  a. Cumprimento de créditos
  b. Revisão bibliográfica
  c. Apresentação do EQE
- Etapa 2:
  a. Programação Linear Inteira
  b. Programação em Lógica com Restrições
  c. Outras heurísticas
- Etapa 3:
  a. Definição de classes de permutações
  b. Estudo do problema do diâmetro

Cronograma de atividades

Etapas	2007		2008			2009
	mar-jul	ago-dez	jan-abr	mai-ago	set-dez	jan-abr
1	a	a,b	b	b	b	b
2			a	a,b	a,b	b,c
3						a
4						
5						
6					a	a

Tabela: Cronograma de atividades para os dois primeiros anos

Cronograma de atividades

Etapas Finais

- Etapa 4: estágio no exterior
  a. Matriz de distância
  b. Estimativas estatísticas
  c. Rearranjo múltiplo de genomas
  d. Geração de métodos específicos para as bactérias da ordem *Rhizobiales*
- Etapa 5: estudo de algoritmos de aproximação
  a. Transposição
  b. Reversão sem conhecer a orientação dos genes
- Etapa 6:
  a. Escrita da tese
  b. Defesa da tese

Cronograma de atividades

Etapas	2009		2010			2011
	mai-ago	set-dez	jan-abr	mai-ago	set-dez	jan-fev
1	b,c					
2	b,c					
3	a	a	a,b	b	b	
4		a,b,c	c	c,d	d	
5	a,b	a,b	a,b	a,b	a,b	
6	a	a	a	a	a	b

Tabela: Cronograma de atividades para os dois últimos anos

Resultados parciais

- Heurísticas aplicadas ao algoritmo de Bafna e Pevzner, 1998
  - Definir uma ordem para aplicação das regras
  - Definir uma ordem para utilização dos ciclos
  - Utilizar um good 0-move antes de um valid 0-move
  - Analisar os valid 2-moves de modo a fazer uma escolha mais apropriada da transposição a ser aplicada

Resultados parciais

Size	WDM	Ch/WCO	H/H	BP/W	M	DD
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	6	0	2	0	0	0
7	72	0	108	0	0	0
8	1167	40	1517	34	124	0
9	14327	1182	25425	1007	2977	0
10	-	-	-	-	69297	5907

- **WDM** Walter, Dias e Meidanis, 2000
- **Ch/WCO** Christie, 1998, com heurísticas implementado por Walter, Curado e Oliveira, 2003
- **H/H** Hartman, 2003, implementado por Honda, 2004
- **BP/W** Bafna e Pevzner, 1998, implementado por Walter et. al, 2005, com heurísticas
- **M** Mira et. al, 2008, usando formalismo algébrico
- **DD** Dias e Dias baseado no algoritmo de Bafna e Pevzner.

Resultados parciais

- Modelos de programação por restrição para o problema da distância de transposição
  - Analisar a técnica de programação por restrição aplicada a distância de transposição
  - Utilizar os limitantes conhecidos para o problema e verificar o mais adequado
  - Implementar o problema da distância de transposição como um problema de satisfação de restrições e como um problema de otimização.

Resultados parciais

	CSP				COP		
Size	**def_csp**	**br_csp**	**cg_csp**	**cg_red_csp**	**def_cop**	**cg_cop**	**gg_cop**
4	0.006	0.010	0.005	0.005	0.134	0.019	0.039
5	0.069	0.087	0.009	0.006	2.530	0.061	0.149
6	1.887	2.367	0.022	0.013	–	1.842	4.421
7	51.69	30.707	0.045	0.024	–	3.797	39.024
8	–	–	0.233	0.104	–	–	
9	–	–	0.946	0.313	–	–	
10	–	–	6.816	2.016	–	–	
11	–	–	20.603	4.212	–	–	

Média de tempo (em segundos) para calcular a distância de transposição entre 1000 permutações escolhidas ao acaso e a identidade. Um *gap* – é inserido caso o modelo demore mais do que 15 horas para terminar

Resultados parciais

- Dias e Souza, 2007, usando 100 instâncias escolhidas ao acaso
  - Permutações de tamanho 9
    - 143.5 segundos
  - Permutações de tamanho 10
    - Proibitivo