

**Eraldo Luís Rezende  
Fernandes**

**Heurísticas para o Problema  
de Seqüenciamento de DNA  
por Hibridação**

**DISSERTAÇÃO DE MESTRADO**

**DEPARTAMENTO DE INFORMÁTICA  
Programa de Pós-graduação em  
Informática**

Rio de Janeiro  
Março de 2005



**Eraldo Luís Rezende Fernandes**

**Heurísticas para o Problema de  
Seqüenciamento de DNA por Hibridação**

**Dissertação de Mestrado**

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática da PUC-Rio

Orientador: Prof. Celso da Cruz Carneiro Ribeiro

Rio de Janeiro  
Março de 2005



**Eraldo Luís Rezende Fernandes**

**Heurísticas para o Problema de  
Seqüenciamento de DNA por Hibridação**

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

**Prof. Celso da Cruz Carneiro Ribeiro**

Orientador

Departamento de Informática — PUC-Rio

**Prof. Antônio Basílio de Miranda**

Departamento de Bioquímica e Biologia Molecular —

Fiocruz

**Prof. Luiz Fernando Bessa Seibel**

Departamento de Informática — PUC-Rio

**Prof. Luiz Satoru Ochi**

Instituto de Computação — UFF

**Prof. Noemi Rodriguez**

Departamento de Informática — PUC-Rio

**Prof. José Eugênio Leal**

Coordenador Setorial do Centro Técnico Científico —

PUC-Rio

Rio de Janeiro, 28 de Março de 2005

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

**Eraldo Luís Rezende Fernandes**

Graduou-se em Ciência da Computação na Universidade Federal de Mato Grosso do Sul em 2002 (Campo Grande, Brasil)

Ficha Catalográfica

Fernandes, Eraldo R.

Heurísticas para o Problema de Seqüenciamento de DNA por Hibridação/ Eraldo Luís Rezende Fernandes; orientador: Celso da Cruz Carneiro Ribeiro. — Rio de Janeiro : PUC–Rio, Departamento de Informática, 2005.

v., 80 f: il. ; 29,7 cm

1. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Dissertações. 2. Algoritmos. 3. Otimização combinatória. 4. Heurísticas. 5. Biologia computacional. 6. Seqüenciamento por hibridação. 7. Seqüenciamento de DNA. 8. Construção de vocabulário. 9. Memória adaptativa. I. Ribeiro, Celso C.. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

## Agradecimentos

Agradeço principalmente à toda minha família pelo apoio incondicional durante o desenvolvimento deste trabalho.

Ao Professor Celso Ribeiro pela contribuição neste trabalho e, principalmente, pela contribuição imensurável na minha formação como pesquisador.

Aos meus amigos pelo apoio. Em especial, à Michel Truyenque.

Aos meus novos amigos que conheci no Rio. Em especial, aos companheiros de república Pablo Soto e Bruno Silvestre.

À Valéria Quadros pelo seu companheirismo tão importante nos momentos mais difíceis.

À CAPES pela bolsa de fomento que viabilizou a realização do meu mestrado.

Ao vovô Leobino.

## Resumo

Fernandes, Eraldo R.; Ribeiro, Celso C.. **Heurísticas para o Problema de Seqüenciamento de DNA por Hibridação**. Rio de Janeiro, 2005. 80p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O seqüenciamento por hibridação é uma alternativa interessante para a tarefa de seqüenciamento de DNA. Este método ainda está sendo aperfeiçoado e pode superar as técnicas utilizadas em termos de tempo e custo. Uma etapa crucial do método consiste em resolver um problema combinatório que pode ser formulado como um caso especial do problema do caixeiro viajante com coleta de prêmios. Neste trabalho, propõe-se uma nova heurística construtiva multi-partida para resolver este problema. Uma estratégia de aprendizado baseada em uma memória adaptativa e um procedimento de construção de vocabulário são utilizados para melhorar o desempenho da heurística multi-partida. A memória adaptativa é utilizada para intensificar as construções de novas soluções com os elementos que aparecem com uma frequência maior nas melhores soluções encontradas anteriormente pela heurística multi-partida. O procedimento de construção de vocabulário consiste em construir novas soluções através da combinação de partes comuns a boas soluções. Testes computacionais mostraram que estas duas estratégias aumentam significativamente o desempenho da heurística multi-partida e são particularmente indicadas para problemas de escalonamento nos quais as melhores soluções são na maioria dos casos formadas por blocos de elementos que aparecem juntos com muita frequência. A heurística proposta supera os resultados dos melhores algoritmos encontrados na literatura, tanto em termos da qualidade das soluções encontradas, como do tempo de computação.

## Palavras-chave

Algoritmos, Otimização combinatória, Heurísticas, Biologia computacional, Seqüenciamento por hibridação, Seqüenciamento de DNA, Construção de vocabulário, Memória adaptativa.

## Abstract

Fernandes, Eraldo R.; Ribeiro, Celso C.. **Heuristics for the Problem of DNA Sequencing by Hybridization**. Rio de Janeiro, 2005. 80p. MSc. Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Sequencing by hybridization is an attractive alternative for DNA sequencing. This novel method can be less time and cost consuming than the techniques applied nowadays. A very important step of this method is to solve a combinatorial problem formulated as a special case of the prize-collecting traveling salesman problem. In this work, we propose a new multistart constructive heuristic to solve this problem. A learning strategy based on adaptive memory and a vocabulary building procedure are used to improve the performance of the multistart heuristic. The adaptive memory is used to intensify the construction of new solutions with the elements that appear frequently in the best solutions previously found by the multistart heuristic. The objective of the vocabulary building procedure is to construct new solutions combining parts of good solutions. Computational experiments have shown that these two methods significantly improves the performance of the multistart heuristic and are particularly suitable for scheduling problems whose best solutions are in most cases built by blocks of elements that appear together very often. The proposed heuristic obtains systematically better solutions and is less time consuming than the best algorithms found in the literature.

## Keywords

Algorithms, Combinatorial optimization, Heuristics, Computational biology, Sequencing by hybridization, DNA sequencing, Vocabulary building, Adaptive memory.



## Sumário

1	Introdução	11
2	Seqüenciamento por Hibridação	14
2.1	DNA, proteínas e os seres vivos	14
2.2	Seqüenciamento de DNA	15
2.3	Arranjo de DNA	17
2.4	O problema de montagem	18
2.5	Reconstrução única	20
2.6	Erros de hibridação	22
2.7	Formulação do problema de montagem com erros	22
2.8	Outras aplicações para o arranjo de DNA	25
3	Algoritmos	26
3.1	Revisão bibliográfica	26
3.2	Cálculo das sobreposições entre provas	28
3.3	Heurística multi-partida	30
3.4	Memória adaptativa	32
3.5	Construção de vocabulário	36
4	Resultados experimentais	52
4.1	Condições dos experimentos	52
4.2	Instâncias	52
4.3	Qualidade das soluções	54
4.4	Ajuste de parâmetros	57
4.5	Comparação de resultados	63
5	Conclusões e trabalhos futuros	74
	Referências Bibliográficas	76
A	Códigos de acesso das seqüências	80

## Lista de Figuras

2.1	Duas visões de uma molécula de DNA	16
2.2	Esquema de um arranjo de DNA após hibridação	18
2.3	Grafo da redução a um problema do caminho Hamiltoniano	19
2.4	Grafo da redução a um problema do caminho Euleriano	19
2.5	Problema ambíguo de reconstrução	21
2.6	Grafo da instância do exemplo da Figura 2.2	24
2.7	Grafo da instância do exemplo da Figura 2.2, com erros	24
3.1	Árvore de sufixo do texto ATAG	29
3.2	Algoritmo construtivo aleatorizado	32
3.3	Heurística multi-partida proposta neste trabalho	32
3.4	Heurística multi-partida com memória adaptativa	34
3.5	Operação $y = \text{Int}(x, x')$	38
3.6	Operação $y = \text{Int}(X)$ , onde $X = \{x, x', x''\}$	39
3.7	Algoritmo para encontrar palavras	40
3.8	Operação $z = \text{EInt}(y, y')$	41
3.9	Operação $z = \text{EInt}(Y)$ , onde $Y = \{y, y', y''\}$	42
3.10	Algoritmo que combina as palavras fornecidas para gerar frases	43
3.11	Operação $z = \text{EInt}(y, y')$ , violando grau máximo de entrada	44
3.12	Operação $z = \text{EInt}(y, y')$ , violando grau máximo de saída	45
3.13	Exemplos de frases geradas pelo algoritmo CombinaPalavras	46
3.14	Algoritmo para gerar uma solução viável a partir de uma frase	47
3.15	Algoritmo que tenta inserir a primeira prova em uma frase	47
3.16	Exemplo da inserção da primeira prova em uma frase (i)	48
3.17	Exemplo da inserção da primeira prova em uma frase (ii)	48
3.18	Exemplo da inserção da primeira prova em uma frase (iii)	48
3.19	Iterações do algoritmo Viabiliza	49
3.20	Procedimento de construção de vocabulário	50
3.21	Heurística multi-partida com construção de vocabulário	50
3.22	Heurística MP com memória e construção de vocabulário	51
4.1	Alinhamento entre as seqüências ATAGG e ATCGA	55
4.2	Histogramas do algoritmo construtivo aleatorizado (300 bp)	59
4.3	Histogramas do algoritmo construtivo aleatorizado (600 bp)	60
4.4	Porcentagem de provas das heurísticas propostas para o grupo R	65
4.5	Similaridade das heurísticas propostas para o grupo R	66
4.6	Tempo de computação das heurísticas propostas para o grupo R	67
4.7	Similaridade para diferentes taxas de erro	68
4.8	Porcentagem de provas para diferentes taxas de erro	68
4.9	Similaridade para diferentes tamanhos de prova	69
4.10	Porcentagem de provas para diferentes tamanhos de prova	69
4.11	Similaridade das heurísticas MP+Mem+CV e SOPAS (grupo R)	73
4.12	Tempo das heurísticas MP+Mem+CV e SOPAS (grupo R)	73

## Lista de Tabelas

2.1	Valores da função $w$ para o exemplo da Figura 2.7	25
4.1	Valores escolhidos para o parâmetro $\alpha$	58
4.2	Resultados dos testes para ajustar os parâmetros $q$ e $d$	62
4.3	Resultados dos testes para ajustar os parâmetros $d'$ e $s_{min}$	63
4.4	Similaridade média para o grupo A	70
4.5	Quantidade de ótimos encontrados para o grupo A	70
4.6	Tempo médio de computação para o grupo A	71
4.7	Resultados dos algoritmos para o grupo B	71
4.8	Quantidade de ótimos encontrados para o grupo R	72