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

Pontifícia Universidade Católica
do Rio de Janeiro

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

Pontifícia Universidade Católica
do Rio de Janeiro

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

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 freqüência 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 freqüência. 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, Seqüênciamento por hibridação, Seqüênciamento 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 construtive 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 sequecing, Vocabulary building, Adaptive memory.

Sumário

Lista de Figuras

Lista de Tabelas