Questões algorítmicas em Biologia Molecular

CARLOS E. FERREIRA, MARIE-FRANCE SAGOT

Primeiro Relatório – Julho de 2005 a Agosto de 2006

O projeto de pesquisa é financiado com o apoio da FAPESP em cooperação conjunta do Departamento de Ciência da Computação do IME-USP com o Inria Rhône-Alpes (França). O tema central do projeto é o estudo de problemas algorítmicos e de otimização que surgem em Biologia Molecular. Este relatório traz os resultados obtidos no primeiro ano do projeto de colaboração.

1 Resumo do Projeto

O Departamento de Ciência da Computação do IME-USP é, desde julho de 2005, um *time associado* do Inria. A colaboração tem se dado com a unidade Rhône-Alpes daquela instituição, onde está o grupo de pesquisa de Marie-France Sagot. Este projeto é coordenado em conjunto por Marie-France Sagot e por Carlos E. Ferreira (IME-USP).

O interesse principal neste projeto é o estudo de problemas de Biologia Computacional. Os avanços tecnológicos em Biologia têm contribuído com uma enorme quantidade de informação que precisa ser tratada e estudada pelos pesquisadores. Um dos maiores desafios no momento é extrair destas informações conhecimento para entender melhor os processos envolvidos. O papel da Ciência da Computação nesta extração é fundamental. A análise de dados envolve a formalização dos problemas e a busca por algoritmos eficientes para resolvê-los. Estes algoritmos requerem teoria matemática e de Ciência da Computação bastante sofisticadas para a completa compreensão dos processos que envolvem organismos vivos.

Este projeto trata do estudo de técnicas avançadas de matemática e computação que são aplicadas na solução de problemas complexos vindos de Biologia Molecular. Os quatro principais tópicos de pesquisa do projeto são os seguintes:

- **Análise de seqüências e modelos para a evolução molecular**: neste tópico o interesse é pela busca de padrões (*motifs*) de tamanhos e tipos diferentes nas seqüências moleculares. Esta identificação requer também bons modelos de evolução destas seqüências.
- **Genômica comparada de larga escala e dinâmica de genomas**: A descoberta de que os genomas não são estáticos se deu na década de 40 do século passado. Os chamados *jumping genes* eram, então, chamados de elementos transportadores. O estudo do comportamento dinâmico dos genes, apesar de bastante avançado, ainda tem um longo caminho a ser trilhado até a completa compreensão dos fenômenos relacionados à expressão gênica.

- **Filogenia**: A inferência de filogenias é um campo importante de estudo. Rearranjos podem ser usados na inferência da história da evolução dos genes, e muitas vezes encontramos contradições entre o que é inferido a partir destas informações com outras inferências calculadas com informações mais gerais do organismo. O estudo destes problemas também é parte deste projeto.
- **Redes bioquímicas**: Atualmente é bastante aceito que os fenômenos bioquímicos que ocorrem nos organismos vivos são regulados por redes de interação entre genes, proteínas e pequenas moléculas. Estas redes têm comportamentos bastante complexos e seu estudo exige conhecimentos de teorias matemáticas e algorítmicas.

2 Atividades do período financiadas pelo projeto

Ocorreram no período de julho de 2005 a julho de 2006 duas atividades financiadas com recursos advindos deste projeto.

- **Visita de pesquisadores franceses ao IME-USP em setembro de 2005**: de 28 de agosto a 9 de setembro de 2005 o Departamento de Ciência da Computação do IME-USP recebeu uma missão de pesquisa formada pelos seguintes pesquisadores do grupo francês:

  - Marie-France Sagot (coordenadora do lado francês do projeto);
  - Christian Gautier (pesquisador chefe);
  - Claire Lemaitre;
  - Ludovic Cottret;
  - Maria Leonor G. Rodrigues Palmeira;
  - Pierre Peterlongo;
  - Vincent Lacroix;

  Nesta visita foi organizado um Workshop com apresentações dos pesquisadores franceses em que problemas de interesse comum foram apresentados e atividades conjuntas com pesquisadores do IME-USP foram iniciadas ou continuadas. As palestras apresentadas neste workshop foram as seguintes:

  - The life of a BAOBAB in Lyon, France. Marie-France Sagot, 30 de agosto de 2005, 14:00. Abaixo segue o resumo da palestra apresentada.

    > This talk will present an overview of the research activities of the BAOBAB team. It will also serve to place in context three other talks from members of the team that will follow this one, one the same day and two the day after. The BAOBAB team is part of two

distinct bigger French research structures, the HELIX project of the INRIA, the national institute of research in computer science, and the "Laboratoire de Biometrie et Biologie Evolutive"(LBBE) which is a laboratory affiliated to both the CNRS and the "Universite Claude Bernard"(Lyon I). The LBBE is composed mainly of biologists, bio-mathematicians and bio-informaticians. The BAOBAB team mirrors this multidisciplinarity in a perhaps starker way. Its members come thus from mathematics, statistics, (theoretical) computer science and biology (molecular biology, evolution, biochemistry). The team has a number of objectives that reflect its varied backgrounds yet common passion for biology. Through it's head, the team has been collaborating with the IME for many years. Since January of this year, the DCC-IME is an official research partner of BAOBAB-HELIX, with support from both the INRIA and the FAPESP for developing its collaboration. It is in the context of this partnership that the talks and visits of seven members of BAOBAB to the IME are taking place. For more information on the activities of BAOBAB, you may consult:
`http://www.inrialpes.fr/helix/people/sagot/team/`
For information on the partnership between BAOBAB and the DCC-IME, you may consult:

`http://www.inrialpes.fr/helix/people/sagot/team/projects/`

`associated_team_usp_helix/associated_team_usp_helix.html.`

– Lossless Filter for Finding Long Multiple Approximate Repetitions Using a New Data Structure, the Bi-Factor Array. Pierre Peterlongo, 30 de agosto de 2005, 14:00. Abaixo segue o resumo da palestra apresentada.

> Similarity search in texts, notably biological sequences, has received substantial attention in the last few years. Numerous filtration and indexing techniques have been created in order to speed up the resolution of the problem. However, previous filters were made for speeding up pattern matching, or for finding repetitions between two sequences or occurring twice in the same sequence. In this talk, we present an algorithm called NIMBUS for filtering sequences prior to finding repetitions occurring more than twice in a sequence or in more than two sequences. NIMBUS uses gapped seeds that are indexed with a new data structure, called a bi-factor array, that is also presented in this paper. Experimental results show that the filter can be very efficient: preprocessing with NIMBUS a data set where one wants to find functional elements using a multiple local alignment tool such as GLAM, the overall execution time can be reduced from 10 hours to 6 minutes while obtaining exactly the same results.

– Similarity between neighbors: selection or mutational bias inside genomes?

Christian Gautier, 31 de agosto de 2005, 14:00. Abaixo segue o resumo da palestra apresentada.

> In all three biological worlds (eukaryotes, prokaryotes, archea) it appears that neighbor sequences (genes or non coding sequences) are more similar than random ones. These similarities are very diverse and can involve base frequencies, transcription direction, expressivity, etc. An important evolutionary question is ”what are the mechanisms that create and maintain such regional similarities?”As very often, the debate between natural selection and mechanistic effects of genome functioning appears and some relevant mathematical analysis from the lab are presented.

– Dinucleotide over- and under-representation in complete bacterial genomes (Bacteria and Archaea). Leonor Palmeira, 31 de agosto de 2005, 15:00. Abaixo segue o resumo da palestra apresentada.

> Statistical analysis of the representation of oligonucleotides has been widely used to try and shed light over sequence structure – to try and determine the degree in which randomness and selection play a part within biological sequences. In order to understand these features, we conducted a systematic study of the over- and under-representation of dinucleotides in completed Bacteria and Archaea genomes. In this talk, we will first present the general statistical methodology used. We will then discuss some main results.

Os pesquisadores franceses receberam diárias do projeto para sua estadia durante o período. Além do contato com o grupo de pesquisa inicialmente envolvido no projeto, coordenado por Carlos E. Ferreira e com a participação de Yoshiko Wakabayashi, Cristina G. Fernandes, José Augusto Ramos Soares, José Coelho de Pina Junior, Carlos Eduardo Rodrigues Alves e Alair P. do Lago (além de vários estudantes de mestrado e doutorado envolvidos), durante sua estadia em São Paulo o grupo francês pode estabelecer outros contatos com pesquisadores do IME-USP. Destaque-se o contato com o Roberto Marcondes Cesar Jr, que discutiu algumas técnicas de análise de dados de microarray e de inferência de redes de genes que vem aplicando no estudo de problemas da área, e com o Eduardo Jordão Neves, do Departamento de Estatística do IME-USP.

- **Visita de pesquisadores do IME-USP ao Inria**: De 3 a 19 de dezembro de 2005 as pesquisadoras Cristina G. Fernandes e Yoshiko Wakabayashi realizaram uma missão de pesquisa ao Inria Rhône-Alpes. Além desta instituição as pesquisadoras visitaram a Université Claude Bernard em Lyon e participaram do COMPBIONETS, um evento da área organizado na França. Nesta missão as pesquisadoras do IME foram acompanhadas pelo aluno de doutorado Augusto Fernandes Vellozo, orientado de Alair P. do Lago. Diversos problemas interessantes de pesquisa foram estudados no período, dando continuidade à colaboração existente e gerando artigos em conferências e revistas internacionais da área, como destacamos abaixo. As passagens aéreas das pesquisadoras foram pagas com recursos advindos do projeto.

3 Outras missões do período

Além da viagem das pesquisadoras relatada acima, que foi financiada pelo projeto, durante o período deste relatório outras viagens foram realizadas entre alunos e ex-alunos de pesquisadores do projeto ao Inria em Lyon.

- Said Sadique Adi. Orientado por Carlos E. Ferreira, o ex-aluno foi bolsista da FAPESP e realizou durante o seu doutorado um estágio de pesquisa no Inria. Após concluir o doutorado esteve em Lyon em janeiro de 2006, a fim de dar continuidade à pesquisa iniciada naquele estágio;
- Fábio Henrique Viduani Martinez. Orientado por José Augusto Ramos Soares, concluiu seu doutoramento em 2004. Esteve em julho de 2006 em visita ao Inria, onde iniciou colaboração com os pesquisadores deste projeto;
- Augusto Fernandes Vellozo. Orientado por Alair Pereira do Lago. Deve completar seu doutorado no ano de 2007, e manifestou interesse de cumprir um estágio de pós-doutoramento junto ao Inria após finalizar seu doutorado.

4 Publicações dos pesquisadores no período relacionadas com o projeto

- S.S. Adi, C.E. Ferreira, *Gene prediction by syntenic alignment*, Proceedings of the Brazilian Simposium on Bioinformatics, 2005, São Leopoldo, RS, Lecture Notes in Bioinformatics, Springer Verlag, 2005. v. 3594, 246-250 (2005).

  Abstract: The search for a gene has undeniable practical importance given, for example, the close relation between these elements and genetic diseases. The gene prediction problem can be addressed in several ways using homology infomation given by genomic DNA and previous annotated sequences. In this paper we present a new comparative-based approach for the gene prediction problem. It is based on a syntenic alignment of two genomic sequences. We have implemented the proposed algorithm and tests on a benchmark including 50 pairs of human and mouse genomic sequences. The results in both nucleotide and exon levels look promising considering the fact that our approach lies manily in homology information between two unannotated sequences.

- S.S. Adi and C.E. Ferreira, *Gene prediction by multiple syntenic alignment*, Journal of Integrative Bioinformatics, v. 13, 2005.

  Abstract: Given the increasing number of available genomic sequences, one now faces the task of identifying their functional parts, like the

> protein coding regions. The gene prediction problem can be addressed in several ways. One of the most promising methods makes use of similarity information between the genomic DNA and previously annotated sequences (proteins, cDNAs and ESTs). Recently, given the huge amount of newly sequenced genomes, new similarity-based methods are being successfully applied in the task of gene prediction. The so-called **comparative-based** methods lie in the similarities shared by regions of two evolutionary related genomic sequences. Despite the number of different gene prediction approaches in the literature, this problem remains challenging. In this paper we present a new comparative-based approach to the problem. It is based on a syntenic alignment of three or more genomic sequences. With syntenic alignment we mean an alignment that is constructed taking into account the fact that the involved sequences included conserved regions intervened by unconserved ones. We have implemented the proposed algorithm in a computer program and confirm the validity of the approach on a benchmark including triples of human, mouse and rat genomic sequences.

- C.E.R. Alves, A.P. do Lago, and A. F. Vellozo, *Alignment with non-overlapping inversions in $O(n^3 \log n)$-time*, Proceedings of GRACO2005, Electronic Notes Discrete Mathematics (19), pages 365–371 (electronic), Amsterdam, 2005. Elsevier.

  > Abstract: Alignment of sequences is widely used for biological sequence comparisons, and only biological events like mutations, insertions and deletions are usually modeled. Other biological events like inversions are not automatically detected by the usual alignment algorithms. Alignment with inversions does not have a known polynomial algorithm and a simplification to the problem that considers only non-overlapping inversions were proposed by Schöniger and Waterman in 1992 as well as a corresponding $O(n^6)$ solution[1]. Recent works improved these results to $O(n^4)$-time and $O(n^2)$-space complexity. In this present extended abstract, we announce an algorithm that solves this simplified problem in $O(n^3 \log n)$-time and $O(n^2)$-space.

- C.G. Fernandes, V. Lacroix, and M-F. Sagot, *Reaction motifs in metabolic networks*, IEEE ACM Transactions on Computational Biology and Bioinformatics, accepted.

  > Abstract: The classic view of metabolism as a collection of metabolic pathways is being questioned with the currently available possibility of studying whole networks. Novel ways of decomposing the network into modules and motifs that could be considered as the building blocks of a network are being suggested. In this work, we introduce a new definition of motif in the context of metabolic networks. Unlike in previous works on (other) biochemical networks, this definition is not based only on

[1] In this paper, n denotes the maximal length of the two aligned sequences.

topological features. We propose instead to use an alternative definition based on the functional nature of the components that form the motif, which we call a *reaction motif.* After introducing a formal framework motivated by biological considerations, we present complexity results on the problem of searching for all occurrences of a reaction motif in a network, and introduce an algorithm that is fast in practice in most situations. We then show an initial application to the study of pathway evolution. Finally, we give some general features of the observed number of occurrences in order to highlight some structural features of metabolic networks.

- V. Lacroix, C.G. Fernandes, and M-F. Sagot, *Reaction motifs in metabolic networks*, Proceedings of the Workshop on Algorithms in Bioinformatics (WABI) 2005, R. Casadio and G. Myers, eds., Lecture Notes in Bioinformatics, 3692, Springer Verlag (2005).

  Este artigo é uma versão anterior do artigo publicado na revista acima, que foi enviado à conferência.

- A.P. do Lago, I. Muchnik, and C. Kulikowski, *A sparse dynamic programming algorithm for alignment with non-overlapping inversions*, Theor. Inform. Appl., 39(1):175–189, 2005.

Abstract: Alignment of sequences is widely used for biological sequence comparisons, and only biological events like mutations, insertions and deletions are considered. Other biological events like inversions are not automatically detected by the usual alignment algorithms, thus some alternative approaches have been tried in order to include inversions or other kinds of rearrangements. Despite many important results in the last decade, the complexity of the problem of alignment with inversions is still unknown. In 1992, Schöniger and Waterman proposed the simplification hypothesis that the inversions do not overlap. They also presented an $O(n^6)$ exact solution for the alignment with non-overlapping inversions problem and introduced a heuristic for it that brings the average case complexity down. (In this work, n is the maximal length of both sequences that are aligned.) The present paper gives two exact algorithms for the simplified problem. We give a quite simple dynamic program with $O(n^4)$-time and $O(n^2)$-space complexity for alignments with non-overlapping inversions and exhibit a sparse and exact implementation version of this procedure that uses much less resources for some applications with real data.

- P. Peterlongo, N. Pisanti, F. Boyer, A.P do Lago, and M-F. Sagot, *Lossless filter for multiple repetitions*, submitted, 2006.

Abstract: Similarity search in texts, notably in biological sequences, has received substantial attention in the last few years. Numerous filtration and indexing techniques have been created in order to speed up

> the solution of the problem. However, previous filters were made for speeding up pattern matching, or for finding repetitions be- tween two strings or occurring twice in the same string. In this paper, we present an algorithm called Nimbus for filtering strings prior to finding repetitions occur- ring more than twice in a string or in more than two strings. Nimbus uses gapped seeds that are indexed with a new data structure, called a bi-factor array, that is also presented in this paper. Experimental results show that the filter can be very effcient: preprocessing with Nimbus a data set where one wants to find functional elements using a multiple local alignment tool such as Glam, the overall execution time can be reduced from 10 hours to 6 minutes.

- J.C. de Pina Junior, F.V. Martinez, and J.A.R. Soares, *Algorithms for Terminal Steiner Trees*, Computing and Combinatorics: 11th Annual International Conference (COCOON), 2005. Lecture Notes in Computer Science 3595, 369–379 (2005).

> Abstract: The terminal Steiner tree problem (TST) consists of finding a minimum cost Steiner tree where each terminal is a leaf. We describe a factor $2\rho - \rho/(3\rho - 2)$ approximation algorithm for the TST, where ρ is the approximation factor of a given algorithm for the Steiner tree problem. Considering the current best value of ρ, this improves a previous 3.10 factor to 2.52. For the TST restricted to instances where all edge costs are either 1 or 2, we improve the approximation factor from 1.60 to 1.42.

- E.M. Rodrigues, M-F. Sagot and Y. Wakabayashi, *The Maximum Agreement Forest Problem: approximation algorithms and computational experiments*, Theoretical Computer Science, accepted.

> Abstract: There are various techniques for reconstructing phylogenetic trees from data, and in this context the problem of determining how distant two such trees are from each other arises naturally. Various metrics for measuring the distance between two phylogenies have been defined. Another way of comparing two trees T and U is to compute the so called *maximum agreement forest* of these trees. Informally, the number of components of an agreement forest tells how many edges from each of T and U need to be cut so that the resulting forests agree, after performing all forced edge contractions. This problem is NP-hard even when the input trees have maximum degree 2. Hein, Jiang, Wang and Zhang presented an approximation algorithm for it, claimed to have performance ratio 3. We show that the performance ratio of Hein's algorithm is at least 4, and we also present two new 3-approximation algorithms for this problem. We show how to modify one of the algorithms into a $(d + 1)$-approximation algorithm for trees with bounded degree d, $d \geq 2$. Finally, we report on some computational experiments comparing the performance of the algorithms presented in this paper.

- J.A.R. Soares, and E. Araújo. *Scoring matrices that induce metrics on sequences*, Latin American Theoretical INformatics LATIN'2006, 2006, Valdivia - Chile. Lecture Notes in Computer Science 3887, 68–79 (2006).

  > ABSTRACT: Scoring matrices are widely used in sequence comparisons. A scoring matrix γ is indexed by symbols of an alphabet. The entry in γ in row `a` and column `b` measures the cost of the *edit operation* of replacing symbol `a` by symbol `b`.
  >
  > For a given scoring matrix and sequences s and t, we consider two kinds of induced scoring functions. The first function, known as *weighted edit distance*, is defined as the sum of costs of the edit operations required to transform s into t. The second, known as *normalized edit distance*, is defined as the minimum quotient between the sum of costs of edit operations to transform s into t and the number of the corresponding edit operations.
  >
  > In this work we characterize the class of scoring matrices for which the induced weighted edit distance is actually a metric. We do the same for the normalized edit distance.

- R. Tavares, S.S. Adi, P. Blayo, M-F. Sagot, *Utopia: an exact generic core algorithm for gene prediction using homology*, submitted.

  > ABSTRACT: The Utopia algorithm presented here is exact and uses a species independent gene model. The homologous genes in both sequences do not need to have the same number of exons. To explore the behaviour of Utopia and to analyse the current limitations of making gene prediction using only similarity information, we used a set of homologous genomic sequences, for which the cognate mRNA sequence is available, belonging to different eukaryotic species. The sequences contain genes presenting a wide variety of structures and degrees of similarity among the members of the family. We show that Utopia is able to produce good results when the homologous genes have dramatically different structures, or when they are not very well conserved, provided that they are better conserved than the flanking genomic sequences. We also show that Utopia obtains good results when compared to the only other exact, generic method publicly available for predicting genes from genomic sequences, Pro-Gen, having the further advantage of being able to deal efficiently with frameshits and partial and multiple genes.

- A.F. Vellozo, C.E.R. Alves, and A.P. do Lago, *Alignment with non-overlapping inversions in $O(n^3)$-time*, Proceedings of the 6th Workshop on Algorithms in Bioinformatics (WABI), Lecture Notes in Bioinformatics (2006).

  > ABSTRACT: Alignments of sequences are widely used for biological sequence comparisons. Only biological events like mutations, insertions

and deletions are usually modeled and other biological events like inversions are not automatically detected by the usual alignment algorithms. Alignment with inversions does not have a known polynomial algorithm and a simplification to the problem that considers only non-overlapping inversions were proposed by Schöniger and Waterman in 1992 as well as a corresponding $O(n^6)$ solution[2]. An improvement to an algorithm with $O(n^3 \log n)$-time complexity was announced in an extended abstract and, in this present paper, we give an algorithm that solves this simplified problem in $O(n^3)$-time and $O(n^2)$-space in the more general framework of an edit graph. Inversions have recently been discovered to be very important in Comparative Genomics and Scherer et al. in 2005 experimentally verified inversions that were found to be polymorphic in the human genome. Moreover, 10% of the 1,576 putative inversions reported overlap RefSeq genes in the human genome. We believe our new algorithms may open the possibility to more detailed studies of inversions on DNA sequences using exact optimization algorithms and we hope this may be particularly interesting if applied to regions around known rearrangements boundaries. Scherer report 29 such cases and prioritize them as candidates for biological and evolutionary studies.

5 Orientações no período relacionadas com o projeto

Os seguintes alunos são orientados por pesquisadores do IME-USP participantes do projeto e têm temas de pesquisa relacionados com o mesmo:

- Andrea Nakazato (Mestrado, orientadora: Yoshiko Wakabayashi). Título do trabalho: “Ordenação por reversão: algoritmos e aplicação à biologia computacional”.
- André Fujita (Doutorado, orientador: Carlos E. Ferreira). Título do trabalho:“Algoritmos para identificação de marcadores gênicos de câncer de glia”.
- Augusto Fernandes Vellozo (Doutorado, orientador: Alair Pereira do Lago). Título do trabalho: “Alinhamento de seqüências com inversões não sobrepostas”.
- Daniel Yugo Nakazato (Mestrado, orientador: Alair Pereira do Lago). Título do trabalho: “Vetores de sufixos: métodos de construção, de busca e aprimoramentos”.
- Domingos Soares Neto (Mestrado, orientador: José Augusto Ramos Soares). Título do trabalho: “Filtros de busca de padrões em seqüências biológicas”.
- Francisco Elói Soares de Araújo (Doutorado, orientador: José Augusto Ramos Soares). Título do trabalho: “Alinhamento de seqüências”.

[2] In this case, n denotes the maximal length of the two aligned sequences.

- Lennon Machado (Mestrado, orientador: Alair Pereira do Lago). Título do trabalho: "Indexação de grandes seqüências e buscas inexatas".
- Manoel Alonso Gadi (Mestrado, orientador: Alair Pereira do Lago). Título do trabalho: "Reconhecimento adaptativo de padrões raros através de Sistemas Imunológicos Artificiais".
- Said Sadique Adi (Doutorado concluído, orientador: Carlos E. Ferreira), concluiu o doutorado em 2005. Título do trabalho: "Identificação de genes por comparação de seqüências".