Questões algorítmicas em Biologia Molecular

Carlos E. Ferreira, Marie-France Sagot

Relatório Final – Julho de 2005 a Junho de 2007

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 principais resultados obtidos neste 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 (Inria) 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 melhor compreensão dos processos que envolvem organismos vivos.

Este projeto trata do estudo de técnicas avançadas de matemática e computação e sua aplicação na resolução 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 financiadas pelo projeto

Nesta seção relatamos as atividades realizadas no período 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 à 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.

- **Visita de pesquisadores do IME-USP ao Inria**: de 1 a 13 de setembro de 2006 o pesquisador Alair P. do Lago realizou uma missão de pesquisa à Université Claude Bernard em Lyon. Nesta missão o pesquisador deu continuidade aos trabalhos realizados em cooperação com Marie-France Sagot, Pierre Peterlongo e Claire Lemaitre. A passagem aérea do pesquisador foi paga com recursos do projeto. A estadia foi paga com recursos do Inria.

- **Workshop on Combinatorial Optimization and Graph Theory**: realizado em Itanhaém, São Paulo, de 8 a 11 de outubro de 2006. Neste workshop participaram vários pesquisadores do projeto e demos continuidade às pesquisas sobre o problema da seqüência comum mais longa sem repetição de símbolos, que estudamos em parceria com o grupo da França.

- **Visita de pesquisadores franceses ao IME-USP em dezembro de 2006**: recebemos a visita de vários pesquisadores franceses na semana de 4 a 9 de dezembro de 2006. Antes disso, eles estiveram em Campo Grande-MS, onde mantiveram contato e deram continuidade a pesquisas desenvolvidas em outras missões com Fábio Henrique V. Martinez, Said S. Adi e Marco Aurelio Stefanes. A lista de pesquisadores visitantes do grupo da pesquisadora Marie-France Sagot segue abaixo:

  - Marie-France Sagot (coordenadora francesa do projeto);
  - Christian Gautier (pesquisador chefe);
  - Claire Lemaitre;
  - Emmanuel Gabriel Prestat;
  - Ludovic Eric Maurice Cottret;
  - Marília Dias Vieira Braga;
  - Paulo Gustavo Soares da Fonseca;
  - Pierre Peterlongo;
  - Vicente Acuña;

  Durante a visita ocorreram várias reuniões de trabalho entre os dois grupos, com especial destaque às atividades desenvolvidas pelo pesquisador Alair Pereira do Lago e os pesquisadores Claire Lemaitre e Pierre Peterlongo. Ainda, houve uma intensa colaboração entre os pesquisadores Carlos E. Ferreira, Cristina G. Fernandes, Yoshiko Wakabayashi, Marie-France Sagot e Marília Braga.

  Após a estadia do grupo a pesquisadora Marília Braga ficou em visita de trabalho no IME-USP por mais uma semana em janeiro de 2007, quando pudemos avançar na busca de modelos e algoritmos para o problema da busca de seqüência comum mais longa sem repetições já mencionado. Ainda com relação a essa pesquisa, durante o ano de 2007 houve algumas visitas curtas dos pesquisadores Fábio H.V. Martinez e Marco A. Stefanes ao IME-USP.

3 Outras missões do período

Além das viagens dos pesquisadores relatadas acima, que foram financiadas pelo projeto, durante o período deste relatório outras viagens foram realizadas entre alunos e ex-alunos de pesquisadores do projeto a 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. 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 a Lyon, onde iniciou colaboração com os pesquisadores deste projeto;
- Augusto Fernandes Vellozo. Orientado por Alair Pereira do Lago. Completou seu doutorado no ano de 2007, e inicia em 3 de julho de 2007 um estágio de pós-doutoramento junto ao grupo da pesquisadora Marie-France Sagot na Université Claude Bernard.
- Roberto Marcondes Cesar. O pesquisador estreitou o contato com o grupo Baobab da Université Claude Bernard. Esteve em visita àquela universidade em maio de 2007. Recentemente submeteu (com a participação de vários pesquisadores deste projeto) uma proposta na chamada STIC-Amsud envolvendo o grupo do IME-USP, do Baobab em Lyon, da UFMS e da Universidad de Chile.

4 Publicações dos pesquisadores no período relacionadas com o projeto

- S.S. Adi, M.D.V. Braga, C.G. Fernandes, C.E. Ferreira, F.H.V. Martinez, M.F. Sagot, M.A. Stefanes, C. Tjandraatmadja, and Y. Wakabayashi, *Repetition-free longest common subsequence*, submetido (2007).

  > ABSTRACT: We study the problem of, given two sequences x and y over a finite alphabet, finding a repetition-free longest common subsequence of x and y. We show several algorithmic results, a complexity result, and we describe a preliminary experimental study based on the proposed algorithms.

- S.S. Adi and 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.

- S.S. Adi and C.E. Ferreira, *A dynamic programming based heuristic for the gene prediction problem*, submetido (2007).

  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, and one of the most promising methods makes use of homology information from evolutionary related sequences. In this paper we present a mathematical formulation of the problem and propose a dynamic programming algorithm to solve it. Like others comparative-based algorithms, the one proposed here is based on the search for similarities between regions of two unannotated genomic sequences. Differently from previous methods, our approach relies on a syntenic alignment of two genomic sequences. In this alignment it is taken into account the fact that the sequences include conserved regions intervened by unconserved ones. The algorithm was implemented and used as a heuristic tool to identify genes in eukaryotic genomes. In this heuristic we incorporate some biological restrictions in order to improve the quality of the results. We confirme the validity of the model and the related algorithm on a benchmark including 50 pairs of human and mouse 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, where n denotes the maximal length of the two aligned sequences. 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, **3** (4), 360-368, 2006.

  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 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.

- A. Fujita, K.B. Massirer, A.M. Durham, C.E. Ferreira, and M.C. Sogayar, *GATO gene annotation tool for research laboratories*, Brazilian Journal of Medical and Biological Research, Ribeirão Preto, SP, **38**, (11), 2005.

  ABSTRACT: The gene annotation tool (GATO) is a Bioinformatics pipeline designed to facilitate routine functional annotation and access to annotated genes. It was designed based on the frequent need, on the part of genomic research groups, to access data pertaining to a common set of genes. Annotation is generated through querying of Web-accessible resources and the information is stored in a local database. It is implemented in PHP and Perl, and may be run on any suitable Web server.

- A. Fujita, J.R. Sato, H.M. Garay-Malpartida, P.A. Morettin, M.C. Sogayar, and C.E. Ferreira, *Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method*, Bioinformatics (Oxford), (to appear) 2007.

  ABSTRACT: Motivation: A variety of biological cellular processes are achieved through a variety of extracellular regulators, signal transduction, protein-protein interactions and differential gene expression. Understanding of the mechanisms underlying these processes requires detailed molecular description of the protein and gene networks involved. To better understand these molecular networks, we propose a statistical method to estimate time-varying gene regulatory networks from time series microarray data. One well known problem when inferring connectivity in gene regulatory networks is the fact that the relationships found constitute correlations that do not allow inferring causation, for which, a priori biological knowledge is required. Moreover, it is also necessary to know the time period at which this causation occurs. Here, we present the Dynamic Vector Autoregressive model as a solution to these problems. Results: We have applied the Dynamic Vector Autoregressive model to estimate time-varying gene regulatory networks based on gene expression profiles obtained from microarray experiments. The network is determined entirely based on gene expression profiles data, without any prior biological knowledge. Through construction of three gene regulatory networks (of p53, NF-B and c-myc) for HeLa cells, we were able to predict the connectivity, Grangercausality and dynamics of the information flow in these networks. Additional figures may be found at http://mariwork.iq.usp.br/dvar/.

- A. Fujita, J.R. Sato, H.M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M.C. Sogayar, and C.E. Ferreira, *Modeling gene expression regulatory networks with the sparse vector autoregressive model*, submetido (2007).

  ABSTRACT: Background: To understand the molecular mechanisms underlying biological important processes, a detailed description of the

> gene products networks involved is required. In order to comprehend such molecular networks, some statistical methods are proposed in the literature to estimate gene regulatory networks from time series microarray data. However, several problems still need to be overcome. Firstly, causality has to be inferred in addition to the correlation between genes. Secondly, we usually try to identify large networks from a large number of genes (parameters) from a smaller number of microarray experiments (samples). Due to the latter, which is rather frequent in Bioinformatics, it is difficult to perform statistical tests using methods that model large gene-gene networks. In addition, most of the models are based on dimension reduction using clustering techniques, therefore, the resulting network is not a gene-gene network but a module-module network. Here, we present the Sparse Vector Autoregressive model as a solution to these problems. Results: We have applied the Sparse Vector Autoregressive model to estimate gene regulatory networks based on gene expression profiles obtained from time-series microarray experiments. Through extensive simulations, by applying the SVAR method to artificial regulatory networks, we show that SVAR can infer true positive edges even under conditions in which the number of samples is smaller than the number of genes. Moreover, it is possible to control the false positives, a significant advantage when compared to other methods described in the literature, which are based on ranks or score functions. By applying SVAR to actual HeLa cell cycle gene expression data, we were able to identify well known transcription factor targets. Conclusions: SVAR method is able to model gene regulatory networks where the number of samples is often lower than the number of genes. It is possible to naturally infer partial Granger causalities without any a priori information and we present a statistical test to control the false discovery rate, which was not previously possible in several gene regulatory network models.

- A. Fujita, J.R. Sato, L.O. Rodrigues, C.E. Ferreira, and M.C. Sogayar, *Evaluating different methods of microarray data normalization*, BMC Bioinformatics, **7**, 469, 2006.

> ABSTRACT: With the development of DNA hybridization microarray technologies, nowadays it is possible to simultaneously assess the expression levels of thousands to tens of thousands of genes. Quantitative comparison of microarrays uncovers distinct patterns of gene expression, which define different cellular phenotypes or cellular reponses to drugs. Due to technical biases, normalization of the intensity levels is a pre-requisite to performing further statistical analyses. Therefore, choosing a suitable approach for normalization can be critical, deserving judicious consideration. Here, we consider three commonly used normalization approaches, namely: Loess, Splines, and Wavelets, and two non-parametric regression methods, which have yet to be used for

normalization, namely, the Kernel smoothing and Support Vector Regression. The results obtained were compared using artificial microarray data and benchmark studies. The results indicate that the Support Vector Regression is the most robust to outliers and that Kernel is the worst normalization technique, while no practical difference were observed between Loess, Splines and Wavelets.

- A. Fujita, J.R. Sato, M.C. Sogayar, and C.E. Ferreira, *Non-parametric Regression and Canonical Correlation Analysis in Tumor Classification*, submetido (2007).

  ABSTRACT: The analysis of SAGE (Serial Analysis of Gene Expression) expression data is one of the main challenges for biologists and mathematicians involved with Bioinformatics. In this paper, we study the use of a mathematical model to analyze tumor samples and classify them, distinguishing normal samples from the tumoral ones. Specifically, we focused on describing applications of statistical approaches, particularly, the non-parametric regression and the canonical correlation analysis. We used the Kernel regression in order to classify human tumor samples based on SAGE gene expression data and predict the patient2̆019s outcome and canonical correlation analysis for dimension reduction. The approach has been tested in real data from glia, prostate and breast tumoral tissues. This methodology was implemented in R and may be downloaded at http://mariwork.iq.usp.br/kernel/.

- 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 Computer Science, 3692, Springer Verlag (2005).

  Este artigo é uma versão preliminar 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.

- F.V. Martinez, G.P. Telles, N.F. Almeida, *Algoritmos e heurísticas para comparação exata e aproximada de seqüências.* In: Marinho P. Barcelos; Antônio A. F. Loureiro. (Org.). XXIV Jornadas de Atualização em Informática / XXV Congresso da Sociedade Brasileira de Computação. Editora da SBC, 2005, v. 24, p. 1545-1586.

  Abstract:

- F.V. Martinez, J.Soares, *Steiner Trees with a Terminal Order.* In: 15th International Conference on Computing, 2006, Cidade do México. Proceedings of the 15th International Conference on Computing. Los Alamitos, California, USA : IEEE Publishing, 2006. p. 254-259.

  Abstract:

- F.V. Martinez, J.C. de Pina, J. Soares, *Algorithms for Terminal Steiner Trees*, Theoretical Computer Science, 2007, aceito.

  Abstract:

- F.V. Martinez, J. Soares, *Steiner Trees With a Terminal Order*, Information Processing Letters, 2007, aceito.

  Abstract:

- 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 **374** (1-3), 91–110, 2007.

  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 Computer Science 4175, 186-196 (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[1]. 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

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

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: "Análise de dados de expressão gênica: normalização de *microarrays* e modelagem de redes regulatórias".
- Augusto Fernandes Vellozo (Doutorado, orientador: Alair Pereira do Lago), concluiu o trabalho em 2007. Título do trabalho: "Alinhamento de seqüências com inversões não sobrepostas".
- Christian Tjandraatmadja (Iniciação Científica, orientador: Carlos Eduardo Ferreira). Título do trabalho: "Variantes do problema da subseqüência comum mais longa".
- 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".
- Gerardo Valdisio Rodrigues Vianna (Doutorado, orientador: Carlos Eduardo Ferreira), concluído em 2007. Título do trabalho: "Técnicas para construção de árvores filogenéticas".
- 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".

6 Uso da reserva técnica

Informo que não houve uso da verba da reserva técnica deste projeto.