Projeto de pesquisa: Estudo de seqüências e aplicações à Biologia Computacional

Augusto Fernandes Vellozo

31 de agosto de 2006

Resumo

O interesse principal deste projeto é a elaboração de algoritmos eficientes que permitam tratar e extrair informações presentes em grandes volumes de dados como seqüências biológicas. Estruturas de dados apropriadas e algoritmos eficientes são particularmente importantes no estudo de seqüências biológicas devido à crescente quantidade de organismos sequenciados e ao grande comprimento destas seqüências.

Peça fundamental na comparação de seqüências biológicas é a obtenção de alinhamentos de seqüências, que no entanto costumam desprezar eventos mutacionais importantes como a duplicação e a inversão. Em nosso trabalho em Biologia Computacional temos obtido algoritmos ainda polinomiais exatos e que reintroduzem estes eventos em sua análise. Este projeto inclui avançar nos trabalhos já iniciados nestes assuntos, assim como trabalhar em novos problemas dentro de outros assuntos da Biologia Computacional, como pontos de quebra (*breakpoints*), motivos (*motifs*), genômica comparativa e técnicas para alinhamentos múltiplos.

O laboratório LBBE (*Laboratoire de Biométrie et Biologie Évolutive*) da Universidade Claude Bernardde - Lyon I ligado ao INRIA, onde será executado este plano de trabalho, assim como a supervisora Marie-France Sagot, têm sua excelência em pesquisa reconhecidas internacionalmente.

1 Conceitos gerais

1.1 Comparação de seqüências

À medida que o número de novos genomas completos aumenta, a comparação entre seqüências longas de DNA de espécies próximas torna-se mais importante para nossso entendimento da estrutura da seqüência do DNA. Devido a isto, a análise genômica comparativa [29], apesar de ser um novo campo na bioinformática, está se desenvolvendo rapidamente. Em muitas espécies próximas, a ordem dos genes é preservada para intervalos curtos [33], i.e. sintenia. Nesses casos, os genes são mais conservados do que as regiões intergênicas. Portanto, a ordem da seqüência de genes é muito útil para detectar reordenamentos cromossômicos como inversões. Estes tipos de comparações ganham maior significância à medida que mais segmentos de genomas ortólogos, fortemente relacionados pela evolução, são sequenciados.

Desde a finalização do rascunho do genoma humano novos projetos de sequenciamento têm sido desenvolvidos para comparação com o genoma humano. Muitos programas computacionais têm sido usados para esse propósito como VISTA [31, 15], GLASS [6], Mummer [12], PipMaker [40], e também BLAST 2 Sequences [41].

Assim como muitos outros estudos em bioinformática, essas comparações dependem fortemente da obtenção de um alinhamento ótimo.

Na história da evolução alguns eventos introduzem mudanças na seqüência do DNA. Alguns eventos biológicos típicos são as *substituições*, *remoções* e *inserções* de nucleotídeos. Portanto, qualquer comparação de seqüências precisa levar em consideração a possibilidade da ocorrência desses eventos, se é esperado identificar uma alta similaridade entre duas seqüências. Procedimentos de alinhamento típicos tentam identificar que partes não mudam e onde se localizam esses eventos biológicos. Após, apresentam um alinhamento ótimo de acordo com algum critério de otimização e sistema de pontuação associado aos eventos.

Alinhamentos podem ser associados a um conjunto de operações de edição que transformam uma seqüência em outra. Normalmente as únicas operações de edição consideradas são a *substituição* de um símbolo em outro, a *inserção* de um símbolo e a *remoção* de um símbolo. Se os custos são associados a cada operação, existe um procedimento de programação dinâmica clássico em

$O(n^2)$[1] que computa o conjunto mínimo de operações de edição com o custo total mínimo e apresenta o alinhamento associado, que tem boa qualidade e alta semelhança para custos realistas.

1.2 Inversões

Considere três novas possibilidades de operações de edição:

- a *reversão-por*-2, que reverte a ordem de *dois símbolos consecutivos*;
- a operação de *reversão*, que reverte a ordem de *qualquer segmento* de símbolos ao invés de um segmento de comprimento 2;
- a operação de *inversão*, que substitui qualquer segmento pela sua seqüência *reversa complementar*. A operação de inversão é a operação de interesse na obtenção de alinhamentos ótimos de seqüências biológicas.

Associados a quaisquer dessas três operações, nós podemos definir novos problemas de alinhamento. Por exemplo, dadas duas seqüências e custos fixos para cada tipo de operação de edição, o problema de *alinhamento com inversões* é um problema de otimização que indaga o custo total mínimo de um conjunto de operações de edição que transforma uma seqüência em outra. Além disso, pode também haver interesse na apresentação do alinhamento correspondente , ou seja das operações de edição. Da mesma forma, podemos definir os problemas de *alinhamento com reversões-por*-2 e de *alinhamento com reversões*.

Em 1975, Wagner [45] estudou o alinhamento com reversões-por-2 e provou que ele admite uma solução polinomial se o custo de uma reversão-por-2 é nulo. Por outro lado, ele também provou que a obtenção de uma solução ótima é *NP-difícil*, se cada operação tem um custo positivo constante.

Em [10] podemos ver que o problema de decisão, associado ao problema de obter um alinhamento considerando inversões para um alfabeto de tamanho ilimitado, é NP-difícil.

Com o objetivo de tratar os alinhamentos com inversões, três estratégias principais têm sido consideradas:

- inversões sem sobreposições;

[1] Consideraremos neste texto que n é o comprimento da maior seqüência analisada

- ordenação de permutações sem sinal por reversões e;
- ordenação de permutações com sinal por reversões.

Em 1992, Schöniger e Waterman [39] introduziram uma *hipótese simplificadora*: *todas as regiões envolvendo inversões não se sobrepõem.* Isso levou ao problema do *alinhamento com inversões sem sobreposições* (primeira estratégia). Eles apresentaram uma solução em $O(n^6)$ para esse problema e também introduziram uma heurística que reduziu a complexidade do tempo de execução do problema. Essa heurística usa o algoritmo desenvolvido por Waterman e Eggert [46] que informa os K melhores alinhamentos locais não mutualmente intersectantes, com o objetivo de reduzir o tempo de execução para algo entre $O(n^2)$ e $O(n^4)$, dependendo dos dados.

A segunda estratégia se aplica bem a alinhamentos de *seqüências de genes* e tem sido bastante usada para genomas de mitocôndrias. Ela não se aplica a seqüências de nucleotídeos nem a seqüências de aminoácidos porque *repetições* de símbolos *não são* permitidas. Além disso, *nenhuma inserção* e *nenhuma remoção* são consideradas e a única operação permitida é a reversão em que a *reversão* é definida para transformar uma seqüência tal como $1, 2, 3, 4, 5$ em $1, 4, 3, 2, 5$. O problema, também chamado de *ordenação de permutações sem sinal por reversões*, produz a média do cálculo das distâncias de edição de duas permutações com a operação de reversão. Neste caso, os dados são duas permutações de $1, 2, 3, \ldots, n$, onde n é o número de genes. Kececioglu e Sankoff [27] propuseram um algoritmo de 2-aproximação em 1995 e Christie [11] propôs um algoritmo de aproximação de razão $3/2$ em 1998. De fato, Caprara [9] provou em 1999 que esse problema na verdade é NP-difícil.

A terceira estratégia é o problema chamado *ordenação de permutações com sinal por reversões.* Este é o mesmo problema de ordenação de permutações sem sinal por reversões até o ponto em que os sinais também são atribuídos a um gene e uma reversão também troca seu sinal. Por exemplo, uma reversão poderia transformar $1, 2, 3, 4, 5$ em $1, -4, -3, -2, 5$. Este sinal é normalmente associado à direção do gene (a qual filamento de DNA ele pertence). Hannenhalli e Pevzner [22, 23] propuseram o primeiro algoritmo polinomial para o problema em 1995 e iniciaram uma seqüência de artigos baseados nessa estratégia. O algoritmo de Hannenhalli e Pevzner era $O(n^4)$ e foi melhorado para $O(n^2)$ por Kaplan, Shamir e Tarjan [25, 26] em 1997. Em 2001, Bader, Moret, e Yan [4] propuseram um algoritmo que calcula a

distância de edição em $O(n)$ (a seqüência de reversões ainda requer $O(n^2)$). Estes estudos têm sido aplicados a estudos de reconstrução filogenética.

Em 2000, El-Mabrouk [17, 18] estudou a inclusão de duas operações: inserções e remoções de segmentos genéticos. Ela obteve resultados parciais e propôs uma solução polinomial exata para um caso e uma heurística polinomial com um testador polinomial para otimalidade no outro caso. Símbolos repetidos ainda não são permitidos. Em 2002, El-Mabrouk [19] também obteve alguns resultados parciais ao considerar reversões e duplicações.

Em 2003, Lago [13] propôs dois algoritmos exatos para o problema do alinhamento com inversões sem sobreposições, ou seja, a primeira estratégia para tratar o problema do alinhamento com inversões. Um algoritmo é uma solução que executa em tempo $O(n^4)$ e que usa espaço $O(n^2)$. O outro algoritmo é uma implementação dinâmica esparsa que reduz o uso de recursos se $o(n^2)$ atribuições são dadas. Isto é freqüentemente esperado se a cardinalidade do alfabeto for grande, como por exemplo quando as letras são fragmentos de DNA de comprimento fixo.

Em 2005, Alves, Lago, e Vellozo [2] propusemos um algoritmo exato para este mesmo problema do alinhamento com inversões não sobrepostas que executa em tempo $O(n^3 \log n)$ e que usa espaço $O(n^2)$.

Em 2006, Vellozo, Alves, e Lago [43] propusemos um algoritmo exato para o mesmo problema que executa em tempo $O(n^3)$ e que usa espaço $O(n^2)$.

1.3 Duplicações

A duplicação é um outro tipo de evento biológico que ocorre em seqüências biológicas e que em geral não é considerada nos algoritmos de alinhamento de seqüências. O evento da duplicação ocorre quando um fragmento da seqüência é copiado e inserido em algum lugar da própria seqüência. Existe um tipo de duplicação muito comum em seqüências de DNA (envolvendo supostamente 10% ou mais do genoma humano) que são as duplicações encadeadas (*tandem duplication*). Uma *duplicação encadeada* é o evento que ocorre quando um fragmento da seqüência é duplicado e inserido na própria seqüência numa posição adjacente ao início ou término do fragmento original. Nestes casos, normalmente o número de cópias encontradas de um mesmo trecho é muito grande. Uma duplicação encadeada resulta numa *repetição encadeada* (*tandem repeat*) na seqüência. Cada cópia destas repetições encadeadas muitas vezes não aparecem exatamente iguais à original, pois algumas vezes eventos de mutação ocorrem durante ou após o evento da duplicação.

Algumas doenças humanas são associadas a estes eventos de duplicação, tais como: retardação mental *fragile-X* [44], doença de Huntington [1], distrofia miotônica [20] e ataxia de Friedreich [8]. *Tandem repeats* podem estar ligados a regras de regulação gênica [21, 30, 34], ligação DNA-proteína [37, 48] e evolução [24].

O número de cópias num *tandem repeat* pode ser variável entre indivíduos diferentes (polimórfico). Locais polimórficos são úteis em várias tarefas de laboratório [16, 47]. *Tandem repeats* tem sido utilizados para sustentar algumas hipóteses da evolução humana [3, 42] e da evolução de micro-satélites (*tandem repeats* cujo tamanho é de apenas algumas unidades de nucleotídeos) em primatas [32].

Em 1997, Benson [7] propôs um modelo para o alinhamento de seqüências que considera o evento das duplicações encadeadas de mesmo tamanho. Ele propôs dois algoritmos exatos para obter um tal alinhamento ótimo. O primeiro algoritmo proposto executa em tempo $O(n^5)$ e espaço $O(n^2)$. O segundo algoritmo proposto executa em tempo $O(n^4)$ e espaço $O(n^3)$.

Estamos elaborando um artigo com um algoritmo que obtém um alinhamento ótimo considerando o mesmo modelo proposto por Benson, porém com tempo de execução $O(n^3)$ e espaço $O(n^2)$.

1.4 *Breakpoints*

Ponto de quebra, ou *breakpoint*, é o local dentro de um genoma onde ocorreu um rearranjo na comparação com outros organismos próximos.

A detecção exata dos *breakpoints* a nível de seqüência de DNA é um problema que, pelo que temos conhecimento, ainda não tem nenhum resultado relevante. Várias aproximações têm sido feitas, tal que os métodos que detectam regiões conservadas do genoma entre diferentes organismos, tentam delimitar uma área relativamente ampla em torno dos possíveis *breakpoints*. Estes métodos, em geral, utilizam em algum momento os procedimentos de alinhamentos para obter estas regiões bem conservadas.

A detecção exata dos *breakpoints* poderia ajudar a entender os mescanismos que estão por trás dos rearranjos e quais na verdade aconteceram. Isto poderia ajudar a melhorar os modelos de comparação de genomas e sua evolução. A hipótese de *Hotspots* (regiões que são mais suscetíveis a rearranjos [5, 28, 35, 36, 38]) é um assunto que poderia ser melhor estudado com a detecção exata dos *breakpoints*.

1.5 *Motifs*

Em genética, um motivo, ou *motif*, é um padrão de seqüência de nucleotídeos ou aminoácidos que é amplamente encontrado e tem, ou espera-se que tenha, um significado biológico.

Um exemplo é o *N-glycosylation motif*, cujo padrão é descrito a seguir: Asn, seguido por qualquer aminoácido exceto Pro, seguido por Ser ou Thr, seguido por qualquer aminoácido exceto Pro, onde as abreviações de 3 letras identificam um tipo de aminoácido. Este padrão pode ser escrito como $N\{P\}[ST]\{P\}$, onde $N = Asn$, $P = Pro$, $S = Ser$, $T = Thr$; $\{X\}$ significa qualquer aminoácido exceto X; e $[XY]$ significa ou X ou Y.

A notação $[XY]$ não mostra qualquer probabilidade da ocorrência de X ou Y. Algumas vezes, estes padrões são definidos de acordo com algum modelo estatístico.

Existem também os *motifs* estruturais, aplicados principalmente em proteínas e que estabelecem padrões tridimensionais.

2 Objetivos

Os objetivos deste projeto são relacionados a seguir:

- A elaboração de algoritmos eficientes que permitam tratar e extrair informações de seqüências biológicas;
- produzir artigos em publicações com reconhecimento internacional na área;
- possibilitar uma grande troca de conhecimentos científicos e acadêmicos entre os pesquisadores envolvidos;
- estabelecer contatos com outros pesquisadores da área de Biologia Computacional e até mesmo com biólogos que trabalhem com análise de seqüências;
- estreitar e fortalecer o relacionamento com a pesquisadora Marie-France Sagot e com outros pesquisadores do laboratório LBBE (*Laboratoire de Biométrie et Biologie Évolutive*) da Universidade Claude Bernardde - Lyon I.

3 Trabalhos a serem realizados

Já foram identificados alguns tópicos que são de interesse de ambas as partes e que iremos investigar durante o projeto. Estes tópicos são:

1. Busca de pontos de quebra (*breakpoints*) ocorridos durante o processo evolutivo;
2. Buscas de inversões e repetições ocorridas durante a evolução de espécies próximas com genomas sequenciados;
3. Buscas de motivos (*motifs*);
4. Técnicas para alinhamento múltiplo (de várias seqüências);
5. Comparação de genomas de espécies próximas;

Pretendemos implementar os algoritmos que vierem a ser elaborados para estes tópicos e realizar testes com dados reais.

Iremos realizar também algumas melhorias e mudanças, já identificadas nos algoritmos para alinhamento com inversões não sobrepostas e para alinhamento com duplicações desenvolvidos durante o doutorado. Este trabalho deverá ser feito no momento em que estivermos trabalhando nas buscas de inversões e repetições, ou seja, deverá ocorrer junto com o item 2.

Eventualmente, poderemos também investigar a possibilidade de inclusão de técnicas aceleradoras como o uso de árvores de sufixos ou de filtragem [14].

Estaremos receptivos a trabalhar também com problemas novos que vierem a surgir durante o projeto.

Para os resultados relevantes obtidos redigiremos os respectivos artigos e sempre que possível, enviaremo-os para publicações de reconhecimento internacional na área.

4 Material e Métodos

As implementações dos algoritmos serão desenvolvidos em microcomputadores, e serão processados dados reais obtidos do sequencimento de seqüências biológicas. Estas seqüências serão obtidas através de bancos de dados públicos ou de algum contacto com biólogos locais.

Quaisquer informações e artigos a serem utilizados na pesqusa serão obtidos preferencialmente através de meios eletrônicos em formato digital. Caso

isto não seja possível então serão utilizados os periódicos e publicações arquivados na universidade.

5 Forma de análise dos resultados

Os resultados obtidos com os algoritmos aqui propostos serão comparados com aqueles obtidos de eventuais programas já disponíveis.

Também pretendemos obter retorno da relevância e acuracidade dos resultados obtidos a partir de diálogos com biólogos e pesquisadores locais.

Referências

[1] A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell*, 72(6):971–983, Mar 1993.

[2] Carlos E. R. Alves, Alair Pereira do Lago, and Augusto F. Vellozo. Alignment with non-overlapping inversions in $O(n^3 \log n)$-time. In *Proceedings of GRACO2005*, volume 19 of *Electron. Notes Discrete Math.*, pages 365–371 (electronic), Amsterdam, 2005. Elsevier.

[3] Armour, Anttinen, May, Vega, Sajantila, Kidd, Kidd, Bertranpetit, Pääbo, and Jeffreys. Minisatellite diversity supports a recent African origin for modern humans. *Nat Genet*, 13(2):154–160, Jun 1996.

[4] David A. Bader, Bernard M. E. Moret, and Mi Yan. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. In *Algorithms and data structures (Providence, RI, 2001)*, volume 2125 of *Lecture Notes in Comput. Sci.*, pages 365–376. Springer, Berlin, 2001.

[5] Jeffrey A. Bailey, Robert Baertsch, W. Kent, David Haussler, and Evan E. Eichler. Hotspots of mammalian chromosomal evolution. *Genome Biology*, 5:R23, 2004.

[6] S. Batzoglou, L. Pachter, J. P. Mesirov, B. Berger, and E. S. Lander. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research*, 10(7):950–958, Jul 2000.

[7] Gary Benson. Sequence alignment with tandem duplication. In *RECOMB '97: Proceedings of the first annual international conference on Computational molecular biology*, pages 27–36, New York, NY, USA, 1997. ACM Press.

[8] V. Campuzano, L. Montermini, M. D. Molto, L. Pianese, M. Cossee, F. Cavalcanti, E. Monros, F. Rodius, F. Duclos, A. Monticelli, F. Zara, J. Canizares, H. Koutnikova, S. I. Bidichandani, C. Gellera, A. Brice, P. Trouillas, G. de Michele, A. Filla, R. de Frutos, F. Palau, P. I. Patel, S. di Donato, J.-L. Mandel, S. Cocozza, M. Koenig, and M. Pandolfo. Friedreich's Ataxia: Autosomal Recessive Disease Caused by an Intronic GAA Triplet Repeat Expansion. *Science*, 271:1423–1427, March 1996.

[9] Alberto Caprara. Sorting permutations by reversals and Eulerian cycle decompositions. *SIAM J. Discrete Math.*, 12(1):91–110 (electronic), 1999.

[10] Xin Chen, Jie Zheng, Zheng Fu, Peng Nan, Yang Zhong, Stefano Lonardi, and Tao Jiang. Assignment of orthologous genes via genome rearrangement. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2(4):302–315, 2005.

[11] David A. Christie. A 3/2-approximation algorithm for sorting by reversals. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998)*, pages 244–252, New York, 1998. ACM.

[12] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376, Jun 1999.

[13] Alair Pereira do Lago, Ilya Muchnik, and Casimir Kulikowski. A sparse dynamic programming algorithm for alignment with non-overlapping inversions. *Theor. Inform. Appl.*, 39(1):175–189, 2005.

[14] Alair Pereira do Lago and Imre Simon. *Tópicos em Algoritmos sobre Seqüências*. IMPA, Rio de Janeiro, 2003. ISBN 85-244-0202-4.

[15] I. Dubchak, M. Brudno, G. G. Loots, L. Pachter, C. Mayor, E. M. Rubin, and K. A. Frazer. Active conservation of noncoding sequences revealed

by three-way species comparisons. *Genome Research*, 10(9):1304–1306, Sep 2000.

[16] Edwards, Hammond, Jin, Caskey, and Chakraborty. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, 12(2):241–253, Feb 1992.

[17] Nadia El-Mabrouk. Genome rearrangement by reversals and insertions/deletions of contiguous segments. In *Combinatorial pattern matching (Montreal, QC, 2000)*, volume 1848 of *Lecture Notes in Comput. Sci.*, pages 222–234. Springer, Berlin, 2000.

[18] Nadia El-Mabrouk. Sorting signed permutations by reversals and insertions/deletions of contiguous segments. *J. Discrete Algorithms*, 1(1):105–121, 2000.

[19] Nadia El-Mabrouk. Reconstructing an ancestral genome using minimum segments duplications and reversals. *J. Comput. System Sci.*, 65(3):442–464, 2002. Special issue on computational biology 2002.

[20] Y. H. Fu, A. Pizzuti, R. G. Fenwick, Jr., J. King, S. Rajnarayan, P. W. Dunne, J. Dubel, G. A. Nasser, T. Ashizawa, P. de Jong, B. Wieringa, R. Korneluk, M. B. Perryman, H. F. Epstein, and C. T. Caskey. An Unstable Triplet Repeat in a Gene Related to Myotonic Muscular Dystrophy. *Science*, 255:1256–1258, March 1992.

[21] H Hamada, M Seidman, B H Howard, and C M Gorman. Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence. *Mol. Cell. Biol.*, 4(12):2622–2630, 1984.

[22] Sridhar Hannenhalli and Pavel A. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. In *ACM Symposium on Theory of Computing*, pages 178–189. Association for Computing Machinery, 1995.

[23] Sridhar Hannenhalli and Pavel A. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *J. ACM*, 46(1):1–27, 1999.

[24] Hellman, Steen, Sundvall, and Pettersson. A rapidly evolving region in the immunoglobulin heavy chain loci of rat and mouse: postulated role of (dC-dA)n.(dG-dT)n sequences. *Gene*, 68(1):93–9100, Aug 1988.

[25] Haim Kaplan, Ron Shamir, and Robert E. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversals. In *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (New Orleans, LA, 1997)*, pages 344–351, New York, 1997. ACM.

[26] Haim Kaplan, Ron Shamir, and Robert E. Tarjan. A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM J. Comput.*, 29(3):880–892 (electronic), 2000.

[27] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica*, 13(1-2):180–210, 1995.

[28] W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Science*, 100:11484–11489, September 2003.

[29] E. V. Koonin. The emerging paradigm and open problems in comparative genomics. *Bioinformatics*, 15(4):265–266, Apr 1999. Editorial.

[30] Q Lu, L L Wallrath, H Granok, and S C Elgin. (CT)n (GA)n repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila hsp26 gene. *Mol. Cell. Biol.*, 13(5):2802–2814, 1993.

[31] C. Mayor, M. Brudno, J. R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, L. S. Pachter, and I. Dubchak. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16(11):1046–1047, Nov 2000.

[32] W. Messier, S.-H. Li, and C.-B. Stewart. The birth of microsatellites. , 381:483–+, June 1996.

[33] G. Moore, K. M. Devos, Z. Wang, and M. D. Gale. Cereal genome evolution. Grasses, line up and form a circle. *Current Biology*, 5(7):737–739, Jul 1995. Review.

[34] Pardue, Lowenhaupt, Rich, and Nordheim. (dC-dA)n.(dG-dT)n sequences have evolutionarily conserved chromosomal locations in Drosophila with implications for roles in chromosome structure and function. *EMBO J*, 6(6):1781–1789, Jun 1987.

[35] Qian Peng, Pavel A. Pevzner, and Glenn Tesler. The Fragile Breakage versus Random Breakage Models of Chromosome Evolution. *PLoS Computational Biology*, 2:e14, February 2006.

[36] P. Pevzner and G. Tesler. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Science*, 100:7672–7677, June 2003.

[37] R.I. Richards, K. Holman, S. Yu, and G.R. Sutherland. Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. *Hum. Mol. Genet.*, 2(9):1429–1435, 1993.

[38] David Sankoff. Chromosomal breakpoint reuse in genome sequence rearrangement. *Journal of Computational Biology*, 12(6):812–821, 2005.

[39] M. Schöniger and M. S. Waterman. A local algorithm for DNA sequence alignment with inversions. *Bulletin of Mathematical Biology*, 54(4):521–536, Jul 1992.

[40] S. Schwartz, Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. PipMaker–a web server for aligning two genomic DNA sequences. *Genome Research*, 10(4):577–586, Apr 2000.

[41] T. A. Tatusova and T. L. Madden. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, 174(2):247–250, May 1999.

[42] S. A. Tishkoff, E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd, K. Cheung, B. Bonne-Tamir, A. S. Santachiara-Benerecetti, P. Moral, M. Krings, S. Paabo, E. Watson, N. Risch, T. Jenkins, and K. K. Kidd. Global Patterns of Linkage Disequilibrium at the CD4 Locus and Modern Human Origins. *Science*, 271:1380–1387, March 1996.

[43] Augusto F. Vellozo, Carlos E. R. Alves, and Alair Pereira do Lago. Alignment with non-overlapping inversions in $o(n^3)$-time. In *6th Workshop on Algorithms in Bioinformatics*. Springer, 2006. Lecture Notes in Bioinformatics, to appear.

[44] Verkerk, Pieretti, Sutcliffe, Fu, Kuhl, Pizzuti, Reiner, Richards, Victoria, and Zhang. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, 65(5):905–914, May 1991.

[45] R. Wagner. On the complexity of the extended string-to-string correction problem. In *Seventh ACM Symposium on the Theory of Computation*. Association for Computing Machinery, 1975.

[46] Waterman and Eggert. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *Journal of Molecular Biology*, 197(4):723–728, Oct 1987.

[47] Weber and May. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet*, 44(3):388–396, Mar 1989.

[48] H.A. Yee, A.K.C. Wong, J.H. van de Sande, and J.B. Rattner. Identification of novel single-stranded d(TC)n binding proteins in several mammalian species. *Nucl. Acids Res.*, 19(4):949–953, 1991.

6 Anexo

6.1 Carta com as justificativas para o pós-doutorado logo após o doutorado

Em Dezembro de 2005 fiquei durante 3 semanas em Lyon financiado pelo INRIA (França) através de um projeto que o nosso departamento (DCC do IME/USP) mantém com o INRIA e o laboratório LBBE (*Laboratoire de Biométrie et Biologie Évolutive*) da Universidade Claude Bernardde - Lyon I - França. Esta viagem foi muito proveitosa, inclusive com possibilidade de publicação de um artigo sobre um problema (alinhamento com duplicações) detectado durante esta minha visita. Porém, o tempo da visita foi muito curto para adquirir os conhecimentos decorrentes de uma convivência com outros pesquisadores de um grande centro de pesquisa no exterior.

Como a visita mostrou-se produtiva acredito que em um período mais longo eu poderia obter mais resultados, trocar conhecimentos e fortalecer a colaboração com pesquisadores daquela instituição e, principalemte, com a pesquisadora Marie-France Sagot (supervisora deste projeto de pós-doutorado) que tem comprovadamente grande competência, grandes conhecimentos na área e reconhecimento internacional (vide currículo anexo).

Portanto, um pós-doutorado com ela irá me ajudar a evoluir como pesquisador, transmitindo-me conhecimentos técnicos e possibilitando-me estabelecer contatos e conviver com outros pesquisadores estrangeiros que fazem pesquisa de ponta na área. Tudo isto é muito importante e me dá condições de prosseguir pesquisando numa instituição nacional de bom nível nesta área, visto que os grandes pesquisadores nestas instituições trabalharam em grandes centros de pesquisa no exterior, ou durante o doutorado e/ou durante o pós-doutorado.

Gostaria muito de fazer este pós-doutorado agora pois continuaria o ritmo de estudos que venho mantendo nos últimos anos, alimentado pelos bons resultados que venho obtendo. No momento não tenho vínculo empregatício nenhum, pois me dedico exclusivamente aos estudos e recebo bolsa da CAPES. Vale a pena ressaltar que sou casado, temos uma casa que construímos e minha esposa tem um bom emprego fixo aqui no Brasil. Se for possível ela irá obter uma licença não remunerada para me acompanhar por pelo menos 6 meses neste pós-doutorado no exterior. Se não for possível ela obter esta licença, ela irá me visitar durante as férias dela (Julho e Janeiro).

Existe um grande interesse no momento também por parte da instituição

francesa para que este projeto aconteça, face à colaboração proveitosa com pesquisadores do DCC - IME/USP.

Acho que agora seria um bom momento para estreitar estes laços e manter este relacionamento com essa instituição de pesquisa, pois aproveitaria o interesse atual de ambas as partes.

São Paulo, 31 de agosto de 2006.

Augusto Fernandes Vellozo