Programa CAPES/COFECUB
Projeto de Cooperação Internacional – Tipo II

Curso de Pós-graduação ao qual a proposta está vinculada: Ciência da Computação do Instituto de Matemática e Estatística da USP
Avaliação pela CAPES: em andamento
Área: Ciências Exatas e da Terra
Duração prevista: 2 anos *Data de Início:* Fevereiro de 1999

1 Título do projeto

Algoritmos para Análise de Genomas

2 Identificação das equipes

2.1 Equipe brasileira

- Nome e função dos coordenadores do projeto:
  - Profa. Dra. Yoshiko Wakabayashi (livre-docente) – coordenadora
    Documento de identificação: RG 4.335.204 – SSP/SP.
  - Prof. Dr. José Augusto Ramos Soares (doutor) – vice-coordenador
- Instituição à qual pertencem os coordenadores:

Instituto de Matemática e Estatística – Universidade de São Paulo (IME-USP)
Departamento de Ciência da Computação
Rua do Matão, 1010 – Cidade Universitária
CEP 05508-900 – São Paulo, SP – Brasil
Tel : +55 11 818 6135 Fax: +55 11 818 6134
E-mail : yw@ime.usp.br, jose@ime.usp.br

- Integrantes:
  - Prof. Dr. Carlos Eduardo Ferreira, doutor, IME-USP
  - Profa. Dra. Cristina Gomes Fernandes, doutora, IME-USP
  - Prof. Dr. João Carlos Setubal, livre-docente, IC-UNICAMP
  - Prof. Dr. José Coelho de Pina Júnior, doutor, IME-USP
  - Prof. Dr. José Augusto Ramos Soares, doutor, IME-USP – vice-coordenador

- Profa. Dra. Nami Kobayashi, doutora, IME-USP
- Profa. Dra. Yoshiko Wakabayashi, livre-docente, IME-USP – coordenadora
- Estela Maris Rodrigues, doutoranda, IME-USP
- Francisco Elói, doutorando, IME-USP
- Liliane Rose Benning Salgado, doutoranda, IME-USP

2.2 Equipe francesa

- Nome e função dos coordenadores do projeto:
  - Prof. Dr. Maxime Crochmore (titular) – coordenador
    Documento de identificação: Passaporte No. 93-11-9324214
  - Dra. Marie-France Sagot (pesquisadora) – vice-coordenadora
- Instituição à qual pertence o coordenador:

  Institut d'électronique et d'informatique Gaspard-Monge (IGM)
  Université de Marne-la-Vallée
  Laboratoire d'Informatique
  5, bd Descartes
  Champs-sur-Marne
  77454 Marne-la-Vallée Cedex 2 – France
  Tel : +33 1 60 95 75 63 Fax : +33 1 60 95 75 57
  E-mail : mac@univ-mlv.fr

- Instituição à qual pertence a vice-coordenadora:

  Service d'Informatique Scientifique (SIS)
  Institut Pasteur
  28, rue du Dr. Roux
  75724 Paris Cedex 15 – France
  Tel : +33 1 40 61 34 61 Fax : +33 1 40 61 30 80
  E-mail : sagot@pasteur.fr

- Integrantes:
  - Equipe do Institut d'électronique et d'informatique Gaspard-Monge (IGM)
    * Prof. Dr. Jean Berstel, Universidade de Marne-la-Vallée, responsável pelo curso de pos-graduação
    * Prof. Dr. Maxime Crochemore, Universidade de Marne-la-Vallée, Diretor do IGM – coordenador

    * Prof. Dr. Eugene W. Myers, Universidade do Arizona, Tucson, USA, colaborador do IGM
    * Profa. Dra. Giuseppina Rindone, Universidade de Marne-la-Vallée
    * Dra. Marie-France Sagot, pesquisadora do Instituto Pasteur, colaboradora do IGM – vice-coordenadora
    * Cyril Allauzen, doutorando, Universidade de Marne-la-Vallée
    * Mathieu Raffinot, doutorando, Universidade de Marne-la-Vallée
  – Equipe do Laboratório de Informática de Rouen (LIR)
    * Prof. Dr. Saïd Abdeddaïm, Universidade de Rouen
    * Prof. Dr. Christian Charras, Universidade de Rouen
    * Prof. Dr. Christophe Hancart, Universidade de Rouen
    * Prof. Dr. Thierry Lecrocq, Universidade de Rouen
    * Laurent Mouchard, doutorando, Universidade de Rouen

3 Detalhamento do projeto

3.1 Justificativa

A recente aceleração dos projetos de seqüenciamento de genomas, nos quais estão envolvidos diversos países no mundo, entre eles o Brasil e a França, fez com que a área de biologia computacional se tornasse uma área de pesquisa de reconhecida importância para os biólogos e os cientistas da computação.

Os problemas dessa área apresentam essencialmente dois aspectos. O primeiro é o de obter uma formalização das questões levantadas por biólogos que atuam nessa área, que permita dar a elas um tratamento matemático rigoroso que seja ao mesmo tempo pertinente em termos da biologia. O segundo aspecto diz respeito ao desenvolvimento de métodos computacionais (algoritmos) que permitam resolver de uma maneira eficaz os problemas tais como foram formalizados. Neste aspecto, incluímos ainda a necessidade de incorporar esses algoritmos aos programas e ferramentas utilizados pela comunidade dos biólogos, de modo a torná-los úteis e acessíveis a essa comunidade.

Esse projeto objetiva trazer contribuições à parte formal assim como à parte prática desse campo de pesquisa. Ele reúne pesquisadores das áreas de algoritmos e combinatória, e se apóia na experiência na área de biologia computacional (acumulada ao longo dos anos por membros das equipes de ambos os países), assim como na experiência do pesquisador americano Eugene Meyers que colaborará no projeto do lado francês. Esse projeto se beneficia também dos estreitos laços que unem a equipe brasileira com a vice-coordenadora da equipe francesa, Marie-France Sagot, que realizou os seus estudos de graduação (Bacharelado em Ciência da Computação) no IME-USP, onde se destacou pelo seu excelente desempenho.

Cabe aqui destacar a participação do pesquisador João Carlos Setubal do Instituto de Computação da UNICAMP, um dos responsáveis junto ao Laboratório de Bioinformática (LBI) por toda a parte de bioinformática do projeto GENOMA, financiado pela FAPESP. Este pesquisador coordena esse laboratório (desde dezembro de 1997), conduzindo pesquisas na montagem de fragmentos de DNA e na análise das seqüências resultantes, adquirindo dessa forma, uma importante experiência em projetos deste tipo. O projeto GENOMA tem como objetivo seqüenciar e analisar o genoma da bactéria *Xylella Fastidiosa*, e nele estão envolvidos 34 laboratórios de biologia molecular no Estado de São Paulo.

O presente projeto deveria contar, do lado brasileiro, com a coordenação desse pesquisador da UNICAMP, porém devido à sua participação no projeto GENOMA, cuja vigência vai até o início do ano 2000, essa coordenação recaiu sobre a equipe da USP que vem demonstrando interesse nessa área desde 1995. A coordenadora do projeto, Y. Wakabayashi e o pesquisador C.E. Ferreira têm atuado na área de grafos e combinatória poliédrica, mas tiveram seus interesses despertados por problemas de montagem de DNA, propostos por J. Meidanis, onde vislumbraram a possibilidade de usar técnicas dessa área no tratamento de tais problemas (essa interação resultou em dois artigos sobre esse assunto, em co-autoria com C.C. de Souza, também da UNICAMP). Essa equipe conta com pesquisadores de formações diversas mas afins que certamente poderão trazer grandes contribuições a este projeto: C.G. Fernandes atua na área de grafos e algoritmos de aproximação; J.C. de Pina Jr. e J.A.R. Soares atuam na área de complexidade e algoritmos combinatórios; e N. Kobayashi atua em autômatos, área essa que tem trazido grandes contribuições à biologia computacional. J.A.R. Soares orientou um mestrado e está orientando um doutorado em biologia computacional; além disso, alguns pesquisadores deste projeto também orientam programas de iniciação científica nessa área.

3.2 Objetivos

Esse projeto possui dois objetivos principais.

O primeiro objetivo é investigar problemas surgidos diretamente dos programas de seqüenciamento de genomas, os quais se encontram em desenvolvimento atualmente no Brasil (http://watson.fapesp.br/Genoma.htm et http://www.dcc.unicamp.br/genoma/) e cuja parte computacional, como já mencionamos, é dirigida por um dos membros da equipe brasileira.

Tais problemas referem-se em particular à montagem (*assembly*) de fragmentos, isto é, à reconstrução da seqüência de DNA original a partir de sua quebra (em pedaços) que é feita para se investigar o seqüenciamento. Vários aspectos desse problema foram tratados anteriormente, de modo formal e prático ao mesmo tempo, em particular por E.W. Myers, mas algumas questões permanecem em aberto. Pretendemos investigar algumas delas, entre as quais, o problema de comparar os resultados fornecidos por diversos programas de montagem e o de refinar a montagem final (através da análise de traços).

Um segundo objetivo, mais amplo, diz respeito à análise de um genoma. Esse objetivo engloba uma grande variedade de problemas que apresentam como característica comum o fato de que fazem uso unicamente da informação presente nas seqüências, ou de uma informação que pode ser deduzida diretamente dessas seqüências, para analisar as macromoléculas e/ou identificar os objetos biológicos envolvidos na função e/ou estrutura do genoma. Tais problemas englobam a predição de genes e de sua função, a inferência de padrões, e a filogenia, assim como a busca e predição de estruturas de RNA. Os dois primeiros problemas apresentam pontos em comum. De fato, uma interessante área de pesquisa no caso da predição de genes explora a possibilidade de se basear unicamente na identificação de sinais presentes nas seqüências, dado um modelo —mesmo rudimentar— de como tais sinais são reconhecidos e processados. Ambos se referem a um campo de investigação que possui aplicações em outras áreas fora da biologia e que podem ser descritos como extração de informação. A filogenia bem como a busca e predição de estruturas de RNA podem ser formuladas através do uso de cadeias de caracteres ou grafos (em especial árvores) dependendo do tipo específico de problema que se está tentando resolver.

3.3 Metodologia

Com pouquíssimas exceções, todos os problemas que propomos tratar nesse projeto exigem um tratamento em três etapas.

A primeira pede que o problema (tendo em vista a solução desejada) seja definido e modelado de maneira precisa e rigorosa. A segunda etapa consiste em elaborar algoritmos que permitam resolver o problema. Será dada aqui preferência aos métodos exatos. No caso de problemas NP-difíceis, sempre que possível, procuraremos desenvolver algoritmos de aproximação (algoritmos rápidos que forneçam uma garantia quanto à qualidade dos resultados obtidos quando comparados com a solução ótima); quando isto não for possível, desenvolveremos heurísticas.

A última etapa consiste em implementar os algoritmos desenvolvidos, colocando-os à disposição de todos os participantes do projeto e pessoas ou entidades interessadas.

3.4 Texto completo

Veja texto anexo entitulado *"Algorithms for genome analysis"*.

O texto referente ao projeto de pesquisa foi elaborado em conjunto com a equipe francesa, tendo sido desde o início escrito em inglês, para facilitar o intercâmbio de idéias e informações entre todos os seus membros. Achamos mais conveniente apresentá-lo tal como foi concebido, esperando que isso não seja empecilho para o julgamento de seu mérito.

3.5 Especificação da infra-estrutura disponível

O grupo de pesquisa do IME-USP compartilha a rede local do Instituto, com cerca de 90 microcomputadores 486/Pentium, mais de 45 *workstations* SUN, um servidor Enterprise 3000 Ultra com 6 CPUs e 2 Gbytes de memória, um servidor SPARC-1000 com 4 CPUs e 704 Mbytes de memória, um servidor SPARC-20, um servidor SPARC-10, cerca de 70 X-terminais, além de uma máquina paralela Parsytec PowerXplorer com 16 nós, 32 processadores e 512 Mbytes de memória, interligados em uma rede local conectada ao backbone FDDI do campus da Cidade Universitária, que por sua vez está conectado à Internet. Outros computadores de alto desempenho da USP (como CRAY, Convex) são acessíveis através da rede do IME ou por linhas discadas. Também está acessível um laboratório de cerca de 100 microcomputadores, usado por alunos de graduação e pós-graduação.

A biblioteca "Carlos Benjamin de Lyra" do IME-USP é uma das mais completas na área de Matemática da América Latina. Seu acervo, que também é especializado em Estatística e Ciência da Computação, reúne cerca de 40.000 volumes (livros, enciclopédias e teses), 1.300 títulos de periódicos (490 correntes) com 250 coleções completas.

O IME-USP é assinante do banco de dados MathSci Disc em CD-ROM, produzido pela American Mathematical Society. Esse banco de dados, atualizado anualmente, contém aproxidamente 1.300.000 itens com mais de 700.000 *reviews* de artigos de periódicos, livros e artigos de conferências. Todos os computadores em rede do IME-USP têm acesso *on line* à MathSciNet.

A biblioteca do IME-USP possui ainda uma videoteca e diversos CD-ROM (INSPEC, Computing Surveys and Abstracts, Mathematical Reviews). Além disso, os docentes e alunos têm acesso a um banco de dados contendo informações sobre os acervos bibliográficos de todas as bibliotecas da USP, através do sistema de consultas DEDALUS. O acesso a esse banco de dados pode ser feito através das estações de trabalho ou diretamente nos terminais existentes na biblioteca.

O Instituto de Computação da UNICAMP dispõe do seguinte parque computacional.

Servidoras de Rede: *Plataforma Solaris:* 4 Sparc 10 (2 com 1 CPU, 2 com 2 CPU's), 4 SPARC 20 (1, 2, 4 e 4 CPU's), 3 Sparc 1000 (1 Sunsite) (2, 2 e 8 CPUs), totalizando cerca de 100 Gb de espaço em disco, além de várias estações servidoras (Sparc 2 e 4) de menor porte. *Plataforma Windows e MacIntosh:* 7 servidoras. Além disso, em breve contará com 6 servidoras Enterprise que devem adicionar cerca de 120 Gb de espaço em disco, além de servidores para plataforma Windows mais potentes. O IC conta ainda com 1 roteador CISCO, 6 switches ATM, e uma servidora Silicon Graphics Indigo-2, com cerca de 5 Gb de espaço em disco.

3.6 Metas de formação e linhas de atuação futuras

Através do presente projeto planejamos formar alunos, capacitando-os a atuarem em instituições de ensino e/ou pesquisa no país. Consideramos de suma importância a oportunidade que um projeto como este proporciona tanto aos alunos quanto aos docentes, permitindo que estes conheçam outros centros de pesquisa e trabalhem durante certo período com pesquisadores que atuam nessa área na França.

Consideramos também muito importante receber visitas de pesquisadores franceses para desenvolver pesquisa conjunta e para promover uma maior divulgação dessa área de pesquisa.

Essencialmente, o que esperamos é que a pesquisa conjunta que planejamos desenvolver resulte na formação de pessoal e publicação de artigos em revistas internacionais de alta qualidade. Esperamos também que a experiência adquirida neste tema possa ser utilizada em projetos futuros (como o projeto GENOMA da FAPESP) que agências governamentais do país poderão eventualmente priorizar, tendo em vista a solução de problemas de interesse nacional.

Esperamos que durante o desenvolvimento deste projeto possamos agregar novos alunos que o IME-USP ou o IC-UNICAMP futuramente aceitarão para os seus programas de mestrado e doutorado.

3.7 Cronograma de implementação de bolsas, estágios e missões

- Missões de estudo (fluxo de doutorandos previstos)
  - **Primeiro ano:** Um doutorando do Brasil para a França
  - **Segundo ano:** Um doutorando do Brasil para a França
- Missões de trabalho (intercâmbios previstos)
  - **Missões França/Brasil:**
    1o ano: Duas
    2o ano: Duas
  - **Missões Brasil/França:**
    1o ano: Duas
    2o ano: Duas

3.8 Orçamento

- Missões de estudos (França/Brasil)
  - **Primeiro ano:** 1 passagem aérea Brasil/França/Brasil e um estágio linguístico na França. Pretendemos solicitar uma Bolsa de Doutorado com duração de 3 anos.
  - **Segundo ano:** 1 passagem aérea Brasil/França/Brasil.

- Missões de trabalho
  - **Missões França/Brasil (a serem financiadas pelo COFECUB):**

    1o ano: 2 passagens áereas (França/Brasil/França) e 42 diárias correspondentes a duas missões de 3 semanas cada.

    2o ano: 2 passagens áereas (França/Brasil/França) e 42 diárias correspondentes a duas missões de 3 semanas cada.
  - **Missões Brasil/França (a serem financiadas pela CAPES):**

    1o ano: 2 passagens áereas (Brasil/França/Brasil) e 42 diárias correspondentes a duas missões de 3 semanas cada.

    2o ano: 2 passagens áereas (Brasil/França/Brasil) e 42 diárias correspondentes a duas missões de 3 semanas cada.

- Custeio de atividades
  - R$ 4.000,00 correspondentes a R$ 2.000,00 por ano, para compra de material bibliográfico e despesas com xerocópias, etc.

4 Currículos das equipes

Seguem anexos os currículos de todos os membros das equipes brasileira e estrangeira.