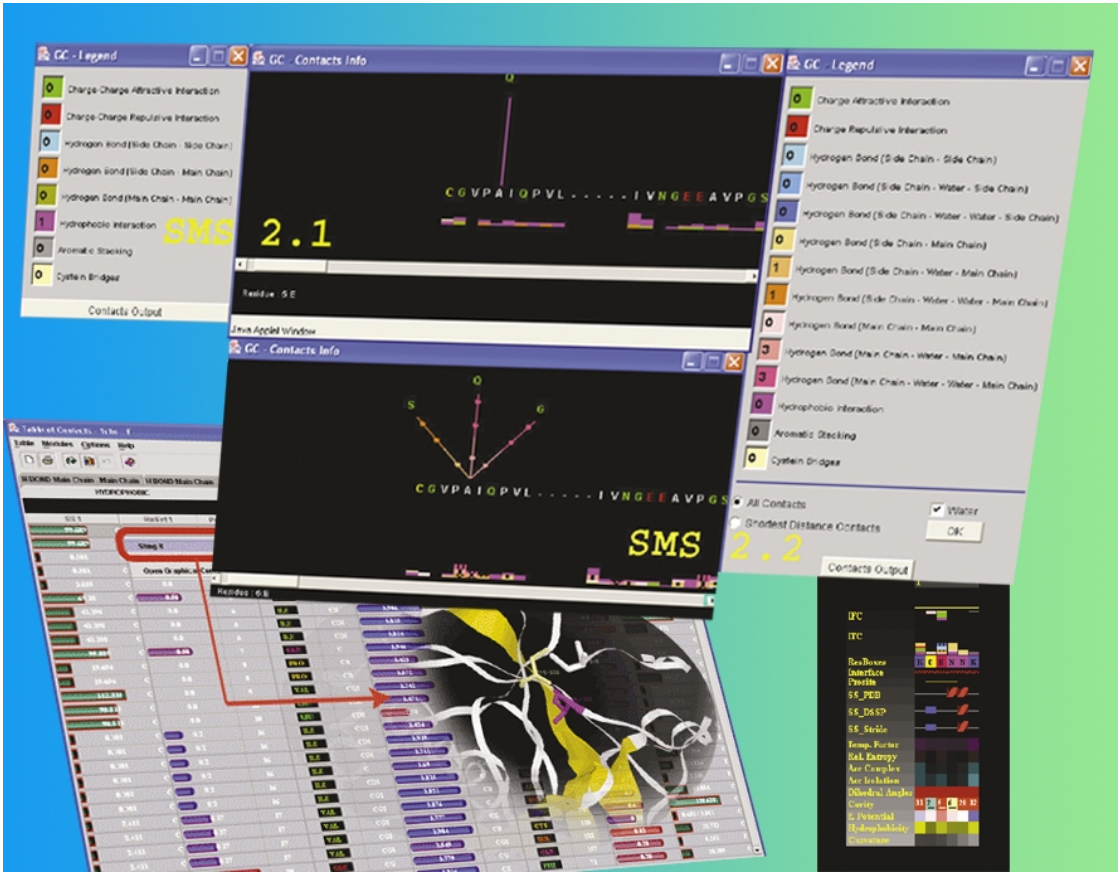
Ministério da Agricultura,
Pecuária e Abastecimento

Documentos

Dezembro, 2002 25

ISSN 1677-9274

Criação de um Núcleo para a Pesquisa e Descrição dos Serviços Oferecidos na Área de Bioinformática Estrutural



Embrapa

ISSN 1677-9274
Dezembro, 2002

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 25

Criação de um Núcleo para a Pesquisa e Descrição dos Serviços Oferecidos na Área de Bioinformática Estrutural

Paula Kuser Falcão
Roberto Hiroshi Higa
Adauto Luiz Mancini
Goran Neshich

Campinas, SP
2002

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Av. André Tosello, 209
Cidade Universitária "Zeferino Vaz" – Barão Geraldo
Caixa Postal 6041
13083-970 – Campinas, SP
Telefone (19) 3789-5743 - Fax (19) 3289-9594
URL: http://www.cnptia.embrapa.br
e-mail: sac@cnptia.embrapa.br

Supervisor editorial: *Ivanilde Dispato*
Normalização bibliográfica: *Marcia Izabel Fugisawa Souza*
Capa: *Intermídia Produções Gráficas*
Editoração eletrônica: *Intermídia Produções Gráficas*

1ª. edição
on-line - 2002

Criação de um núcleo para a pesquisa e descrição dos serviços oferecidos na área de bioinformática estrutural / Paula Kuser Falcão... [et al.]. – Campinas : Embrapa Informática Agropecuária, 2002.
39 p. : il. – (Documentos / Embrapa Informática Agropecuária ; 25).

ISSN 1677- 9274

1. Bioinformática. I. Falcão, Paula Kuser. II. Série.

CDD – 570.285 (21st ed.)

Autores

Paula Kuser Falcão
Ph.D. em Física Aplicada, Cristalografia de Proteínas, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
e-mail: paula@cnptia.embrapa.br

Roberto Hiroshi Higa
M.Sc. em Engenharia Elétrica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
e-mail: roberto@cnptia.embrapa.br

Adauto Luiz Mancini
Bacharel em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
e-mail: adauto@cnptia.embrapa.br

Goran Neshich
Ph.D. em Biofísica,Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
e-mail: neshich@cnptia.embrapa.br

Apresentação

Este documento relata as atividades realizadas no primeiro ano do projeto de bioinformática que visa a criação de um núcleo para pesquisa e oferta de serviços de bioinformática estrutural, para análise de estruturas de proteínas. O objetivo principal deste projeto consiste em instalar e oferecer ferramentas computacionais através de uma interface Web. Esse ambiente visa atender à crescente demanda nesta área de pesquisa originada principalmente nos projetos de seqüenciamento de genoma.

Estão descritas aqui as ferramentas do software Sting Millennium Suite (SMS) desenvolvidas durante o projeto, as máquinas utilizadas para o desenvolvimento dos trabalhos, e as possíveis aplicações deste trabalho.

Atendendo às necessidades de formação de recursos humanos e divulgação de pesquisa, também descreve-se aqui como foi desenvolvido o primeiro curso oferecido para pesquisadores da área de bioinformática.

O SMS é uma ferramenta importante para o pesquisador que trabalha com estruturas de proteínas. As análises obtidas com a utilização do programa podem dar subsídios para entender questões relativas ao funcionamento dessas moléculas, levando em conta a relação estrutura/função.

José Gilberto Jardine
Chefe-Geral

Sumário

Criação de um Núcleo para a Pesquisa e Descrição dos Serviços Oferecidos na área de Bioinformática Estrutural

Paula Kuser Falcão
Roberto Hiroshi Higa
Adauto Luiz Mancini
Goran Neshich

Introdução

Este relatório descreve as atividades e resultados obtidos no período de 01/01/02 a 20/12/02 referente ao auxílio individual à pesquisa "Criação de um Núcleo para a pesquisa e oferta de serviços em BioInformática" (Proc. nº 01/08895-0). O auxílio foi pedido para Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), para a compra de material permanente para a formação do laboratório de bioinformática visando os seguintes objetivos:

1. instalação e oferta de ferramentas computacionais *Sting Millennium Suite* (SMS) através da interface web;
2. criação de novos algoritmos e programas para análise estrutural das proteínas;
3. oferta de bancos de dados públicos armazenados localmente;
4. estabelecimento de um ambiente (Núcleo de Bioinformática) para a pesquisa e oferta de serviços na área de bioinformática;
5. formação de recursos humanos na área de bioinformática estrutural.

Instalação e Oferta de Ferramentas Computacionais *Sting Millennium Suite* (SMS) através da Interface Web

STING Millennium (SMS) é uma suíte de programas com ferramentas para análise estrutural de proteínas. Estes programas estão concentrados em um pacote com o objetivo de oferecer um instrumento completo para estudos de macromoléculas. Informações como posição dos aminoácidos na seqüência e na estrutura, busca de padrões, identificação de vizinhança, ligações de hidrogênio, ângulos e distâncias entre átomos, são facilmente obtidas. Além disso, dados sobre natureza e volume dos contatos atômicos inter e intracadeias, análise da qualidade da estrutura, conservação e relação entre os contatos intracadeia, parâmetros funcionais, são questões que o usuário pode responder sobre sua proteína utilizando o SMS.

A palavra STING é um acrônimo de *Sequence To and withIN Graphics*. O programa STING original (Neshich et al, 1998) foi desenvolvido para permitir um acoplamento bidirecional das informações de seqüência e estrutura de uma proteína e também para oferecer uma maneira simples e fácil de mapear um único aminoácido (ou nucleotídeo) na sua posição tridimensional, e vice-versa.

O *STING Millennium* (SMS) expandiu sua lista de propriedades e agora é uma ferramenta utilizada tanto para uso didático como para pesquisa na área de biologia molecular estrutural. A análise da interface macromolecular foi especialmente caracterizada. Os programas são fáceis de serem usados e praticamente não requerem nenhum treinamento.

Arquitetura e Organização do SMS

A versão 2.1 do *STING Millennium* foi lançada em novembro de 2001. Desde então, o SMS passou por uma reformulação na sua organização que agora se apresenta da maneira representada na Fig. 1.

O SMS está organizado em duas camadas lógicas: o *sms servidor* e o *sms cliente*. O *sms servidor* é responsável pela atualização regular de todos os bancos de dados de domínio público utilizados pelo sms – pdb, hssp e prosite e, calcular, utilizando estes bancos de dados, uma série de propriedades macromoleculares para cada estrutura do pdb – potencial eletrostático, curvatura, área acessível para solvente para cada cadeia e para todo o complexo, as estruturas secundárias calculadas de acordo como dssp (Kabsch & Sander, 1983) e com o stride (Frishman & Argos,

1995), contatos intra e intercadeia, interações proteina/dna, hidrofobicidade, ângulos dihedros, e padrões definidos pelo prosite (descrição completa na Fig. 1). Ele também é responsável por prover acesso ao banco de dados sms através do protocolo HTTP. O *sms cliente* provê uma interface gráfica e amigável com o usuário, envia as requisições do usuário para o *sms servidor* e apresenta as respostas formuladas pelo sms ao usuário.

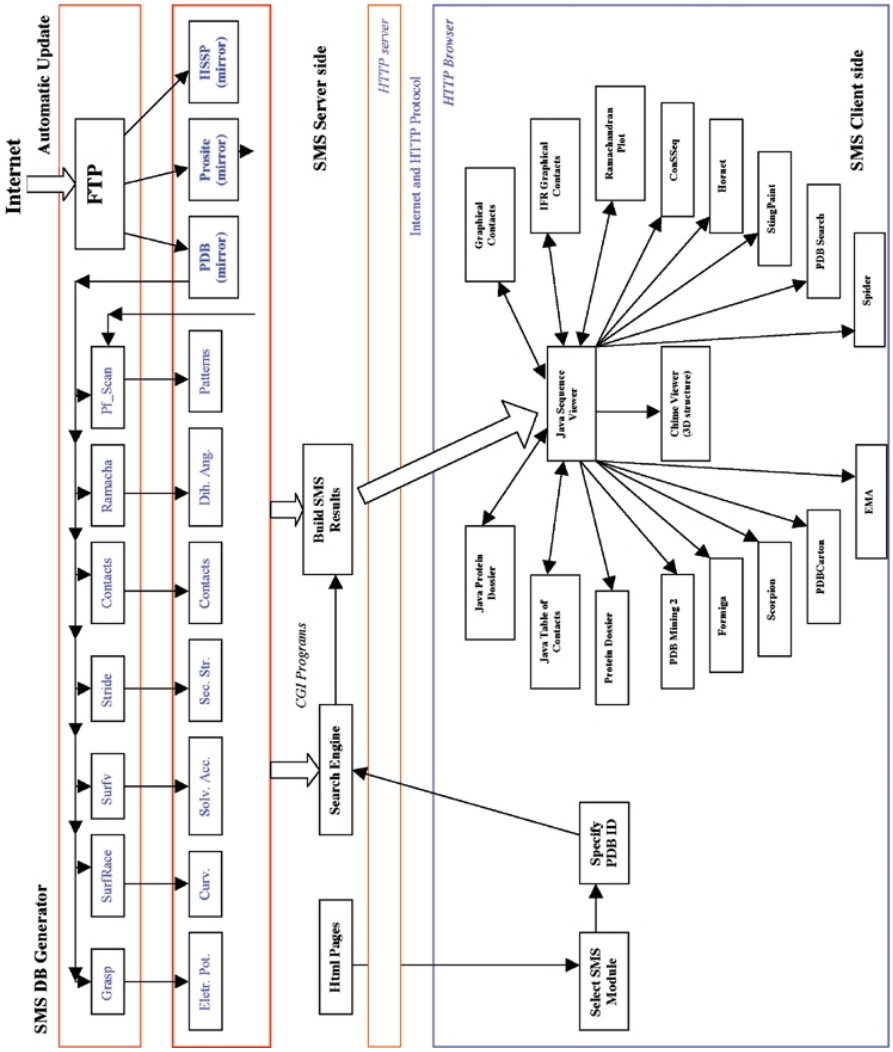


Fig. 1. Organização do SMS a partir da versão 2.2.

A interação entre o sms cliente e o sms servidor acontece da seguinte forma: um cliente HTTP (*browser*) permite que o usuário selecione um módulo sms e especifique um pdb id. No lado do servidor, um servidor HTTP atende às requisições dos clientes sms transmitidas através da Internet. Quando ele recebe uma requisição de um cliente, ele a redireciona para programas *Common Gateway Interface* (CGI) apropriados. Estes programas CGI realizam a busca pelo arquivo pdb apropriado e, acessando os bancos de dados do sms, criam a resposta apropriada para a requisição. Esta resposta é, então, transmitida para o cliente para ser apresentada ao usuário.

A resposta é apresentada para o cliente utilizando uma série de recursos como HTML 3.2 – *Hyper Text Markup Language*, java script 1.2 e java 2 e o MDL chime plugin, dependendo do módulo requisitado. Todos estes recursos constituem uma interface viva e que pode ser usada pelo usuário para explorar os resultados.

Criação de Novos Algoritmos e Programas para Análise Estrutural das Proteínas

O SMS é um programa que está em contínuo desenvolvimento. Atualmente está disponível a versão 2.2 do programa, que foi instalada no final do ano de 2002 para uso externo. Na versão 2.2 foram instroduzidas muitas novidades em módulos já existentes ou ainda novos módulos que serão descritos nesta seção.

Está em desenvolvimento e já se encontra instalada, uma versão beta da versão 3.0 do SMS. Na versão 3.0 está sendo implementado o algoritmo *Java Protein Dossier e Java Table of Contacts* que são os módulos *Protein Dossier e Contacts* da versão anterior agora aproveitando-se das facilidades da linguagem Java. Esta versão deve ser instalada para uso externo no segundo bimestre de 2003.

A versão seguinte a ser lançada será a versão 3.1, que terá um módulo com informações sobre os contatos entre proteínas e DNA. A previsão é que esta versão fique pronta para acesso externo ainda no primeiro semestre de 2003.

Versão 2.2

A versão 2.2 do software SMS foi liberada em 1º de outubro de 2002 e está instalada nas máquinas servidoras do Núcleo de Bioinformática Estrutural da Embrapa Informática Agropecuária de Campinas, Barry Honig Laboratory/Columbia University/EUA e Cenargen/Embrapa/Brasília. As novidades introduzidas na versão 2.2,em relação a versão anterior estão listadas e descritas na Tabela 1.

Tabela 1. Descrição das novas características da versão 2.2 do SMS em relação a versão 2.1.

Nova característica	*SMS 2.1/2.2*	*Descrição*
1 Distâncias Ca-Ca em gráfico Java (Fig. 2)		O *Java Ca-Ca Distance Plot* é um gráfico que mostra as distâncias entre os carbonos alfa de um resíduo e os carbonos alfa de todos os outros resíduos da molécula.
2 Distâncias Cb-Cb em gráfico Java (Fig. 3)		O *Java Cb-Cb Distance Plot* é um gráfico que mostra as distâncias entre os carbonos beta de um resíduo e os carbonos beta de todos os outros resíduos da molécula.
3 *Contacts* totalmente revisado com melhores definições	Fig. 4	*Contacts* tem agora uma lista completa de todos os possíveis contatos entre dois aminoácidos quaisquer dentro de uma mesma cadeia polipeptídica ou entre cadeias. O usuário pode escolher entre mostrar todos os contatos ou somente os contatos mais curtos.
4 Contatos formados na interface (*IFR Contacts*) redefinidos com apresentação adequada no Java Protein Dossier (*JPD*)	Fig. 5	*IFR Contacts* inclui todos os resíduos que podem fazer contatos mas não estão na interface. Isto acontece porque os resíduos que formam a interface (*interface forming residues*) são definidos estritamente baseados na acessibilidade ao solvente. No entanto, mesmo aqueles resíduos da superfície de cadeias próximas que não perderam a acessibilidade ao solvente, ainda podem interagirfazendo, por exemplo, ligações de hidrogênio, uma vez que a distância entre doador e receptor é maior que o diâmetro da molécula de água utilizada para definir os resíduos formadores da interface (IFRs).
5 Contatos dos anéis aromáticos (*aromatic stacking contacts*)	Fig. 4	As interações entre aminoácidos com anéis aromáticos são agora identificadas e representadas no *Graphical Contacts* (GC), *Contacts* (IFRC) e *Protein Dossier* (PD)
6 Ligações de hidrogênio com até duas moléculas de água intermediárias incluídas no GC, IFRC e PD.	Fig. 6	A lista de potenciais doadores e receptores de hidrogênios foi extendida. O usuário pode optar entre exibir ou não as moléculas de água intermediárias.
7 Definição de estrutura secundária baseada no *STRIDE* foi incluída no PD	Fig. 6	Definição de estrutura calculada com o algoritmo STRIDE para complementar as definições já existentes do PDB e DSSP.
8 Inclusão dos ângulos diedros no PD		O *Protein Dossier* agora tem os valores dos ângulos *phi* e *psi* para cada amino ácido, colorido de acordo com as áreas do Diagrama de *Ramachandran*.
9 Estatísticas no Diagrama de *Ramachandran*		Inclusão de estatísticas no SMS *Java Ramachandran Plot*.
10 Revisão e melhora do *PDB Mining*		Arquivos PDB identificados pelo número de águas cristalinas e pelo número de ligantes
11 Ajuda indexada		
12 Nova entrada para web page	Fig. 7	SMS 2.2 oferece uma página inicial onde o usuário pode escolher entre três tipos de páginas: a) uma página para conexões lentas; b) página gráfica com ilustrações artísticas; c) página gráfica com ilustrações de moléculas.

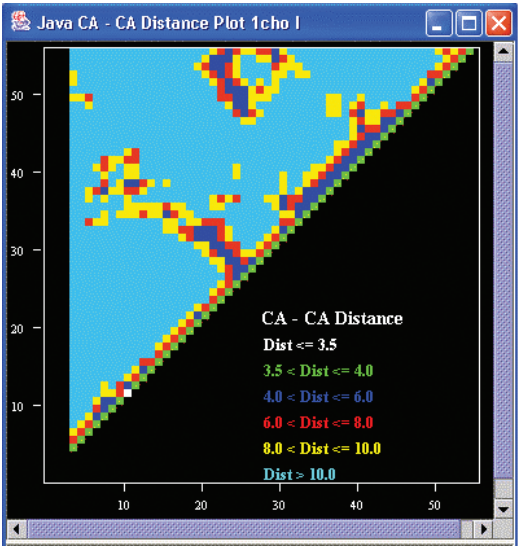


Fig. 2. *Java Ca-Ca Distance Plot.*

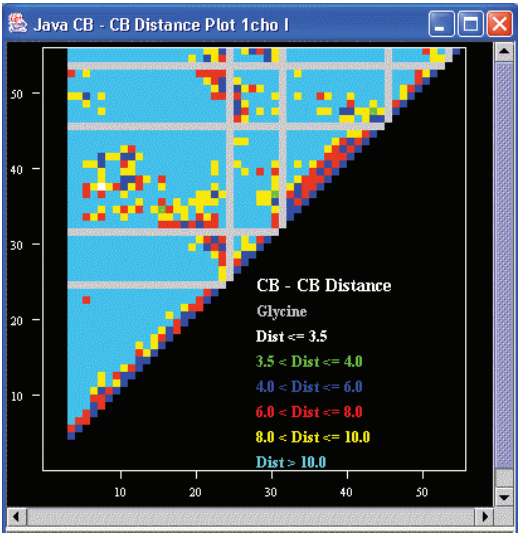


Fig. 3. *Java Cb-Cb Distance Plot.*

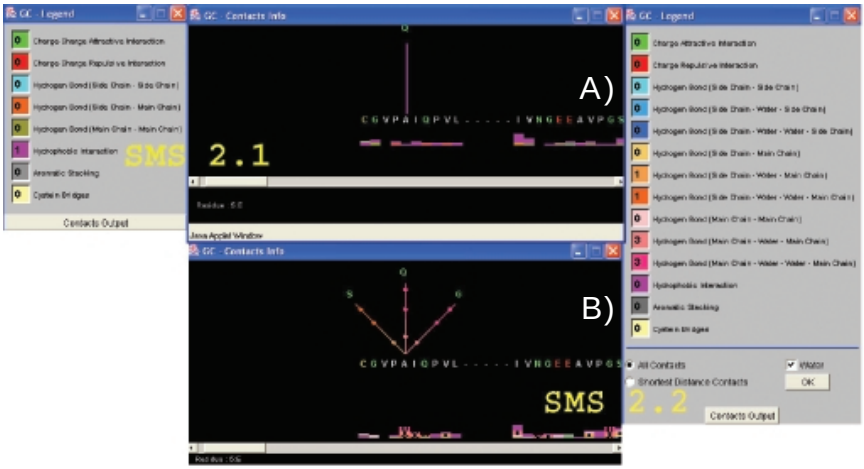


Fig. 4. Contacts em A) SMS 2.1, B) SMS 2.2. Em B é possível ter uma lista completa de todos os possíveis contatos entre dois aminoácidos quaisquer dentro de uma mesma cadeia polipeptídica ou entre cadeias.

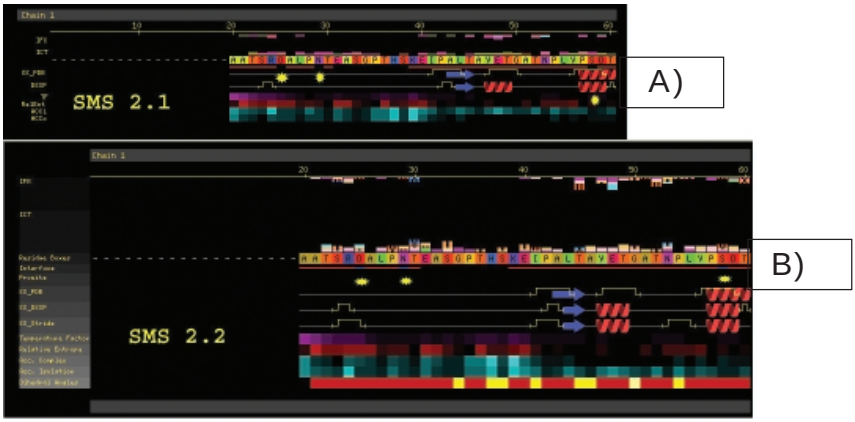


Fig. 5. IFR Contacts em A) SMS 2.1, B) SMS 2.2. IFR Contacts na versão 2.2 inclui mesmo aqueles resíduos da superfície de cadeias próximas que não perderam a acessibilidade ao solvente e ainda podem interagir através de interações como, por exemplo, ligação de hidrogênio.

Fig. 6. *Protein Dossier* em A) SMS 2.2 , B) SMS 2.1. Na versão 2.2, o *Protein Dossier* têm novos parâmetros associados.

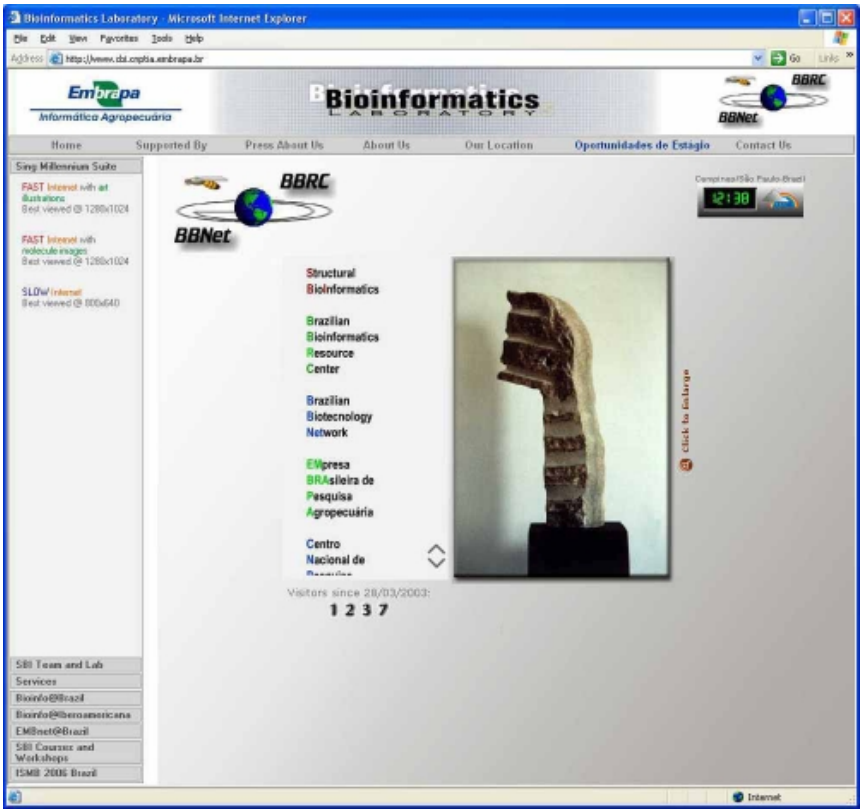


Fig. 7. SMS 2.2 oferece uma página inicial onde o usuário pode escolher entre três tipos de páginas: a) uma página para conexões lentas, b) página gráfica com ilustrações artísticas, c) página com ilustrações de moléculas.

Versão 3.0

Esta versão tem como novidade o aplicativo *Protein Dossier* usando a linguagem de programação Java (*Java Protein Dossier* – JPD), com vários parâmetros novos adicionados, e também um aplicativo Java para a Tabela de Contatos (internos e na superfície). A versão 3.0 está instalada na servidora beta do NBI (http://bsgi.nbi.cnptia.embrapa.br/SMS).

O *Java Protein Dossier* é uma atualização do já existente *Protein Dossier,* onde foram acrescentados os parâmetros potencial eletrostático, curvatura, hidrofobicidade, dupla ocupância e acessibilidade relativa (Fig. 8).

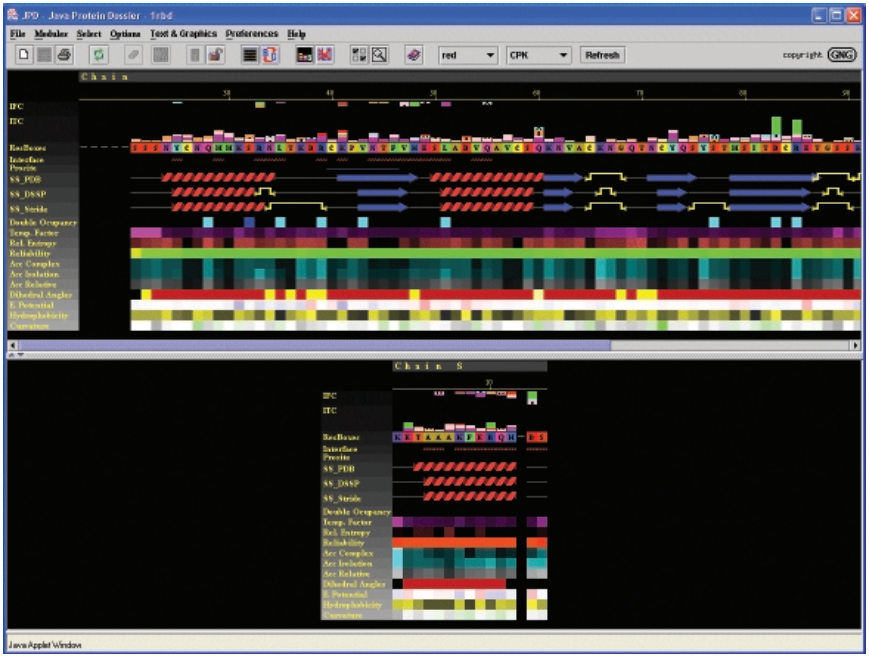


Fig. 8. *Java Protein Dossier* da proteína 1rbd (ribonuclease-s), um complexo que consiste de dois fragmentos de ribonuclease-a: proteína-s (resíduos 21-124) e peptídeo-s (resíduos 1 – 15). A parte superior da Fig. representa a cadeia polipeptídica da proteína, enquanto na parte inferior (cadeia S) estão representados os aminoácidos do peptídeo. As duas cadeias podem ser visualizadas simultaneamente no *Java Protein Dossier*, com o objetivo de proporcionar uma visualização facilitada dos resíduos que estão na interface proteína-peptídeo *(Interface Forming Residues)* e de suas interações. No JPD também é possível visualizar simultaneamente duas cadeias de proteínas distintas, visando a verificação de regiões estruturalmente conservadas.

O potencial eletrostático pode ser usado como ferramenta na análise da estrutura e função da proteína uma vez que estas interações interatômicas governam as interações entre proteínas e ligantes. Os valores de potencial eletrostático são obtidos com o programa *GRASP* (Nicholls et al., 1991). Além de precisar ser energeticamente favoráveis, as interações entre macromoléculas dependem fortemente das características da propriedade de curvatura das superfícies que interagem. Os valores de curvatura são calculados com o programa *SurfRace* (Tsodikov et al., 2002). O efeito hidrofóbico, considerado uma das mais importantes das várias forças que determinam a estrutura tri- dimensional de uma proteína, está mapeado no JPD de acordo com a escala de hidrofobicidade desenvolvida Radzicka & Wolfenden (1988).

O parâmetro introduzido para sinalizar a dupla ocupância ocorre em vários arquivos PDB com coordenadas quando é possível uma segunda opção de ocupância dos átomos nos mapas de densidade eletrônica de estruturas resolvidas através de difração de raios X. Resíduos que possuem ocupância dupla são agora marcados no *Java Protein Dossier*.

A maior contribuição em JPD (comparado com o *Protein Dossier*) é a possibilidade de consultar qualquer dado dos parâmetros mapeados através do clique do mouse. Esta flexibilidade oferece uma riqueza de informações disponíveis apenas clicando o mouse por cima do parâmetro e aminoácido em questão.

O Java Table of *Contacts* produz um perfil completo das interações que podem ocorrer em uma estrutura tridimensional e cria um gráfico destes contatos (Fig. 9). Este gráfico lista todas as possíveis interações na proteína, incluindo interações com as moléculas de água, além de outros parâmetros representados quantitativamente com valores, e qualitativamente com cores. Os parâmetros listados aqui incluem o valor da acessibilidade relativa de cada resíduo em isolamento e em complexo com outra cadeia, a entropia relativa de cada resíduo, o tipo de estrutura secundária a qual o resíduo pertence, a distância dos átomos interagindo, com código de cores adequado.

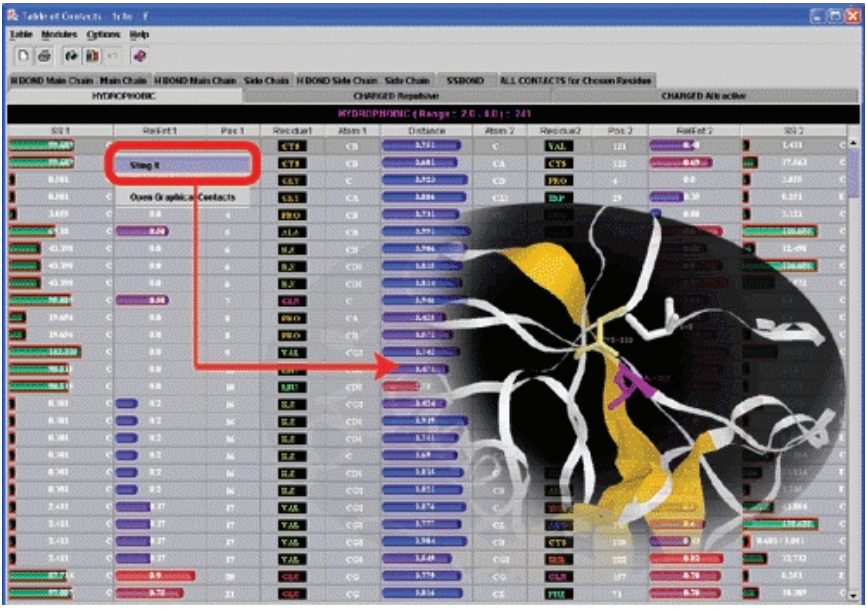


Fig. 9. Exemplo de análise obtida com o aplicativo *Java Table of Contacts*.

Versão 3.1

Adiciona o aplicativo *Protein DNA Contacts*, para a oferta de conhecimento nos estudos estruturais para famílias de proteínas que fazem interação com DNA, como por exemplo as proteínas reguladoras de expressão gênica. A definição dos contatos utilizados pelo SMS estão apresentados a seguir, juntamente com a descrição dos contatos proteínas/DNA calculados e apresentados somente a partir da versão 3.1. Apesar deste aplicativo se encontrar em fase de teste, já está instalado na servidora alfa do NBI (http://asgi.nbi.cnptia.embrapa.br/SMS)

Exemplo de aplicação do *STING Millennium*

Apresenta-se a seguir um exemplo do uso do software SMS. A Fig. 10 mostra uma colagem de gravuras de diferentes módulos de uma sessão do *STING Milllenium*, que analisa e oferece uma descrição completa do aminoácido His57 da proteína alfa-quimotripsina complexada com um ovomucóide (1cho.pdb).

Este aminoácido foi escolhido numa análise preliminar com o módulo *Java Protein Dossier* (Fig. 11a a 11f) que rapidamente pode mostrar os resíduos que se encaixam dentro de critérios físico-químicos e estruturalmente importantes como: estar estereoquimicamente bem posicionado, pertencer a interface da enzima com o inibidor, ter um alto índice de conservação evolutiva, e ser um resíduo hidrofílico da superfície da cadeia. Com apenas 5 cliques de mouse pode-se obter a listagem de todos os aminoácidos que satisfazem estes critérios a partir do *Java Protein Dossier.*

A análise ilustrada na Fig. 10 permite responder a várias questões relevantes para a estrutura da proteína sobre este aminoácido específico com apenas alguns cliques de mouse. Exemplos de questões que podem ser respondidas são:

1. Que tipo de contatos internos o resíduo His57 faz? Quantos contatos?
2. Quais são os resíduos que estão em contato com a His57 dentro da estrutura tridimensional?
3. Como estão os ângulos dihedros deste resíduo?
4. Quanto este resíduo está exposto a uma pressão evolucionária e que outros resíduos ocupam esta posição em outras seqüências homólogas?
5. Mostrar todos os parâmetros que podem ser calculados e mostrados no SMS para esta seqüência específica.

As questões 1 e 2 são respondidas com o módulo *Graphical Contacts,* (imagens A e B). Existem interações com os resíduos Ala55, Asp102, e Ser195, descritas como uma interação hidrofóbica com a Alanina 55, uma ligação iônica entre o átomo OD1 do aminoácido Asp102 e o átomo ND1 da His57, uma ligação de hidrogênio entre o átomo N da cadeia principal da His57 e o átomo OD2 da cadeia lateral do aminoácido Asp102, e ainda uma ligação de hidrogênio entre os átomos NE2 da His e OG do aminoácido Ser195. A questão 3 é verificada com o módulo *Ramachandran Plot* (imagem C), ou seja, o resíduo His57 está numa região favorável para resíduos que fazem parte de uma hélice alfa. A questão 4, é resolvida com a utilização do módulo *ConSSeq* (imagem D). O aminoácido His57 é altamente conservado, com um grau de conservação de 97%. Em 3% dos casos existe uma Valina nesta posição. A questão 5 pode ser verificada com o módulo *Java Protein Dossier* (imagem E). Alguns parâmetros mostrados no dossier indicam que: His57 é um aminoácido localizado na

interface da proteína, tem fator de temperatura baixo, está localizado no meio de uma hélice alfa, não está acessível quando complexado com o ligante, mas sua acessibilidade aumenta em isolamento.

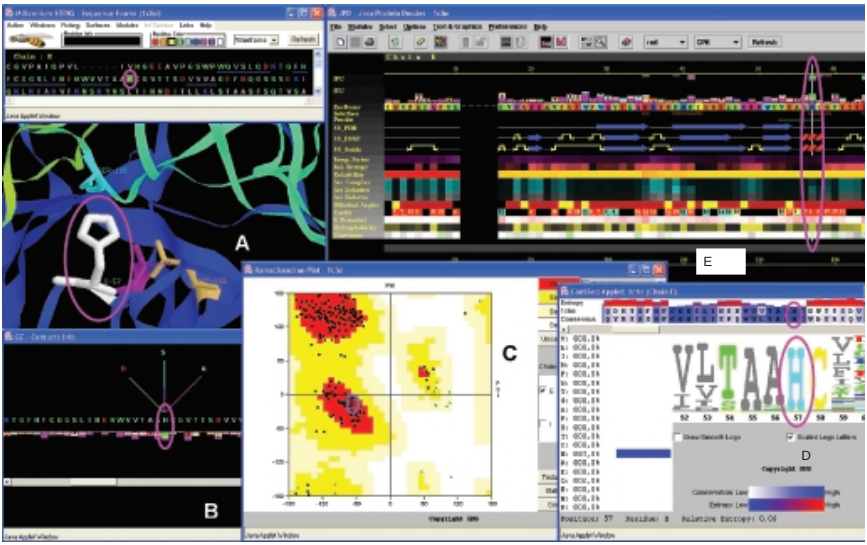


Fig.10. Colagem de figuras que representam a análise do resíduo His57 do arquivo pdb 1cho. A) vizinhança da His57 em 3D, B)*Graphical Contacts*, C) Diagrama de *Ramachandran*, D) *ConSSeq*, E) *Protein Dossier*.

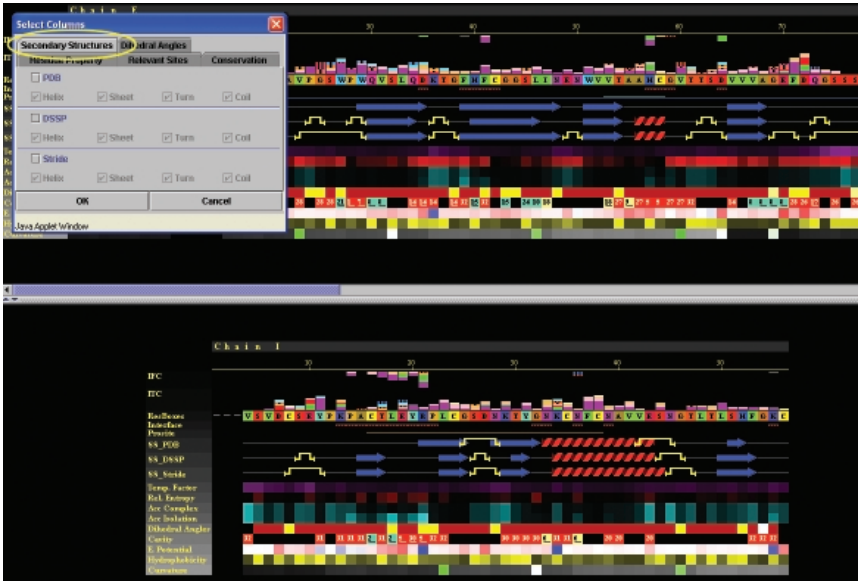


Fig. 11a. JPD selecionando resíduos sem especificar estrutura secundária.

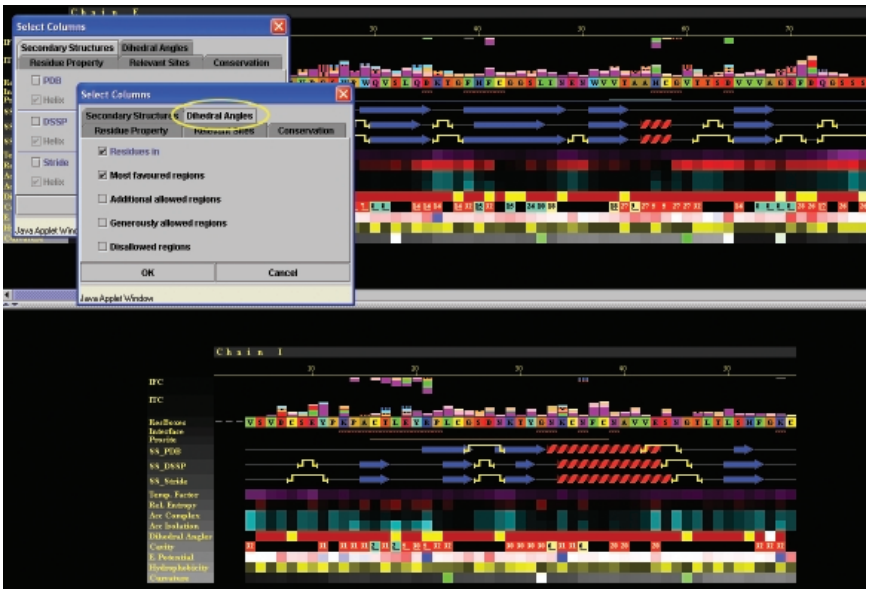


Fig.11b. JPD selecionando resíduos com ângulos diedros localizados na região mais favorável do Gráfico de *Ramachandran.*

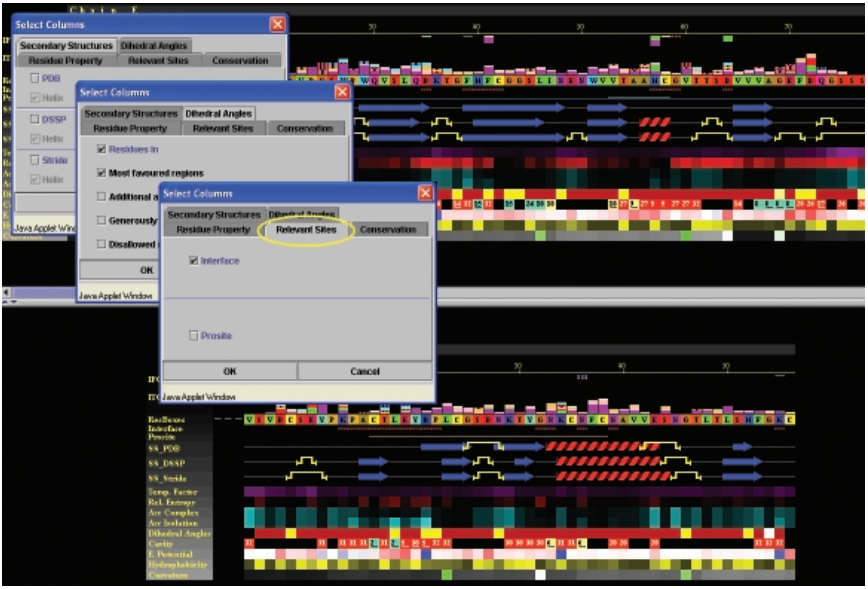


Fig. 11c. JPD selecionando resíduos que estão localizados na interface.

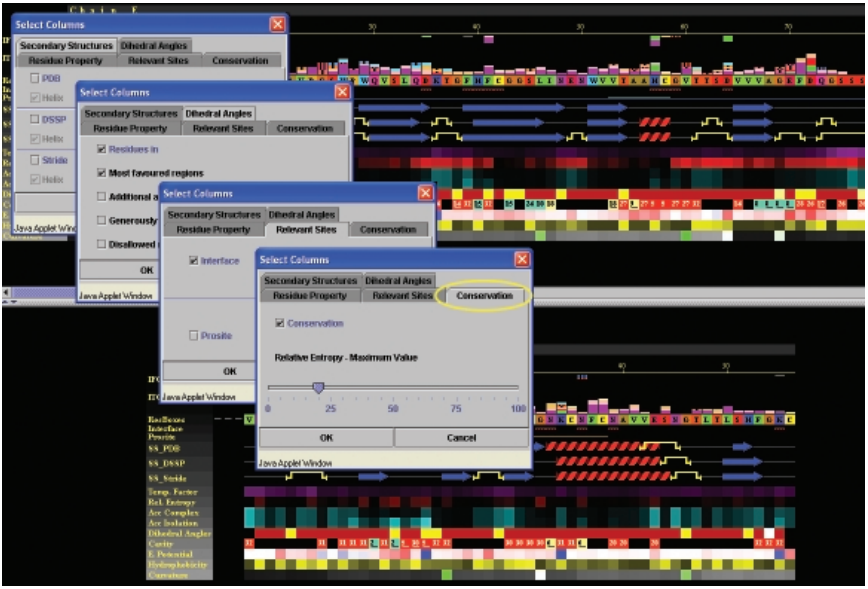


Fig.11d. JPD selecionando resíduos altamente conservados.

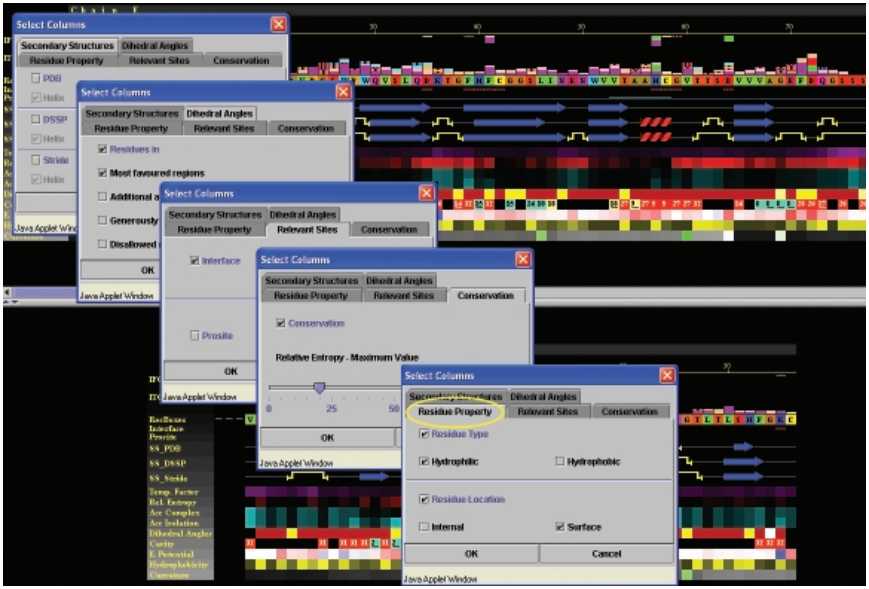


Fig. 11e. JPD selecionando resíduos hidrofílicos localizados na superfície.

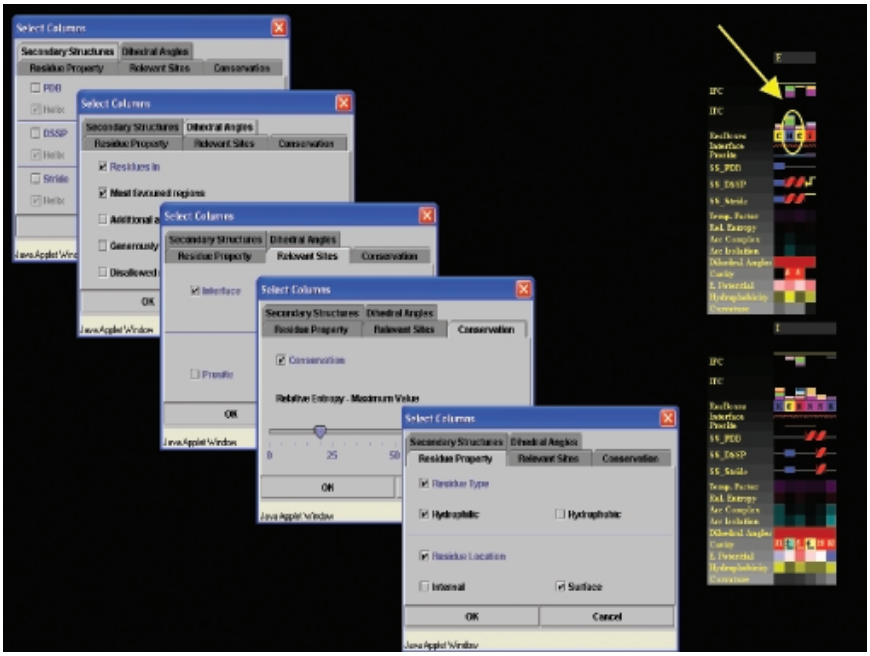


Fig. 11f. Resultado da seleção dos resíduos mostrando apenas aqueles resíduos estruturalmente importantes para o funcionamento da enzima.

Estatísticas de Acesso ao Software SMS

Número de acessos em cada núcleo

- Protein Data Bank (PDB)

Estatística:

Período 09-Jan. 2002 a 12-Dec.-2002

Número de acessos completados: 648.893

Média de acessos por dia: 2.928

Total de dados transferidos: 36.006 Mbytes

http://mirrors.rcsb.org/SMS/SMS_stat/

Gráfico de acesso mensal

- Universidade de Columbia – Nova Iorque – USA

Estatística:

Período 19-Out.-2001 a 11-Dez.-2002

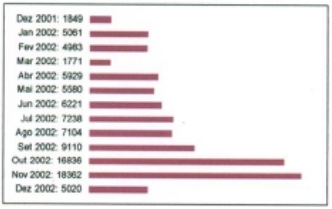
Número de acessos completados: 95.065

Média de acessos por dia: 266

Total de dados transferidos: 563.848 kbytes

http://trantor.bioc.columbia.edu/SMS/SMS_stat/

Gráfico de acesso mensal



- Cenargen/Embrapa – Brasília

Período 04-Mar.-1997 a 12-Dez.-2002

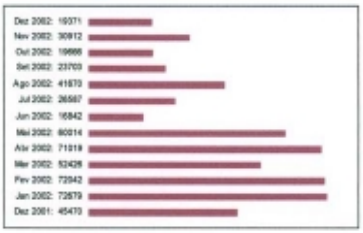
Número de acessos completados: 3.203.389

Média de acessos por dia: 2.519

Total de dados transferidos: 49.299 Mbytes

http://asparagin.cenargen.embrapa.br/statistics/

Gráfico de acesso mensal



- Núcleo de Bioinformática Estrutural (NBI) da Embrapa Informática Agropecuária – Campinas

Período 27-Jun.-2002 a 12-Dez.-2002

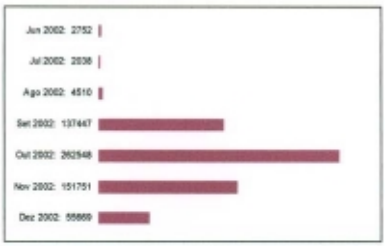
Número de acessos completados: 616.715

Média de acessos por dia: 4.682

Total de dados transferidos: 11.765 Mbytes

http://www.nbi.cnptia.embrapa.br/SMS/statistics

Gráfico de acesso mensal



Exemplo de Locais que estão Acessando o SMS

O programa SMS é acessado por laboratórios do mundo todo, por grupos de pesquisa em universidades ou companhias. A lista a seguir mostra alguns laboratórios de renome internacional que utilizaram os módulos do SMS no início de dezembro, segundo registrado nas estatísticas de acesso do SMS. Esta lista foi reduzida para representação neste documento. A informação completa pde ser encontrada na nossa webpage.

País/Local	servidor
Alemanha	
Universidade de Hamburgo	lc berlin.zbh.uni-hamburg.de
Universidade de Hannover	pc3.lci.uni-hannover.de
Universidade de Leipzig	spatz.iom.uni-leipzig.de
Universidade de Heidelberg	pc-sattler5.nmr.embl-heidelberg.de
Indústria Química e Farmacêutica Bayer	bayer-212-64-224-241.bayer.de
Austrália	
Universidade de Melbourne	grumpy.its.unimelb.edu.au
Canada	
Universidade de Alberta	artemis.biochem.ualberta.ca
Universidade de Montreal	phmc15.pharmco.umontreal.ca
EUA Comercial	
Companhia Farmacêutica Pharmacia	us1.pharmacia.com
Centro de Pesquisa Almaden , da IBM	wfp2.almaden.ibm.com
Companhia Farmacêutica Exelixis	shaker.exelixis.com
ICN Farmacêutica	zamfir.icnpharm.com
Empresa Farmacêutica Merck	spot.merck.com
Avigen Inc/Gene Terapy	avigen19.avigen.com
Farmacêutica Agouron/Pfizer	tbone.agouron.com
EUA Educacional	
Universidade de Harvard	net-55780.deas.harvard.edu
Massachusetts Institute of Technology - MIT	exon2.mit.edu
Instituto de Pesquisa Scripps	als.scripps.edu
Universidade de Stanford	oldmanbalaji.stanford.edu
California Institute of Technology – CalTech	thought.caltech.edu
Universidade da Califórnia, Los Angeles	wangjj.seas.ucla.edu
Universidade de Princeton	chm-323pc-3.princeton.edu
Universidade da Pensilvânia	cdb1108.med.upenn.edu
Universidade de Purdue	ecn198-dhcp-27.ecn.purdue.edu
Biomolecular Structure Center da Universidade de Washington	wh.bmsc.washington.edu
Universidade da California, Santa Cruz	dhcp-141-207.ucsc.edu
Michigan State University	rhea.bch.msu.edu
Finlândia	
Universidade de Helsinki	base.medchem.helsinki.fi
França	
Instituto de Ciências da Matéria – Escola Superior de Engenharia	cmtmoissan2.ismra.fr
Escola Nacional Superior de Química de Paris	ch226.enscp.jussieu.fr
Inglaterra	
Universidade de Londres-Birkbeck College	gh61.chem.bbk.ac.uk
Universidade de York	ursula.chem.york.ac.uk
Universidade de Leeds	proxy2.leeds.ac.uk
The Hutchinson/MRC Resarch Centre-Cancer Resarch UK	deathstar.hutchison-mrc.cam.ac.uk
Itália	
Universidade de Torino	sgpc54.pharm.unito.it
Japão	
Universidade de Osaka	lab8-1.cheng.es.osaka-u.ac.jp
Universidade de Shizuoka	ed2604.u-shizuoka-ken.ac.jp
Companhia Nippon Kayaku	ns.nipponkayaku.co.jp
Suécia	
Empresa Biotech	moskva.biotech.kth.se

Oferta de Bancos de Dados Públicos

Cópias locais atualizadas de vários bancos de dados foram colocadas à disposição do usuário através da página web do NBI. São os seguintes bancos de dados:

- PDB (*Protein Data Bank*) - Banco de dados com todas as estruturas de proteínas e nucleotídeos disponíveis,
- *SwissProt* - Banco de dados curado de sequências de proteínas com nível alto de anotação,
- EMBL - Banco de dados de seqüências de nucleotídeos,
- TrEMBL - Suplemento do SwissProt anotado por computador,
- HSSP - Banco de dados de estrutura secundária de proteínas derivadas por homologia,
- DSSP - Banco de dados de estrutura secundária para todas as proteínas do PDB,
- *Prosite* via SMS - Banco de dados de famílias de proteínas e domínios.

Além disso estão disponibilizados manuais de vários programas e a interface para acesso à suíte de programas *EMBOSS*, um pacote de programas com código livre de alta qualidade para análise de seqüências.

A partir de outubro de 2002 o NBI faz parte do nó nacional da EMBNet. EMBnet é formada por um grupo de laboratórios científicos colaboradores na Europa e alguns colaboradores fora da Europa. A EMBNet tem como missão não apenas colocar à disposição de seus membros bancos de dados biológicos atualizados, mas também de unir o trabalho de profissionais da área de bioinformática com o objetivo de dar suporte às áreas de genética e biologia molecular. O NBI entrou em um consórcio com outros três laboratórios brasileiros: Laboratório de Genômica e Expressão do Instituto de Biologia da Unicamp, Departamento de Bioquímica e Biologia Molecular do Instituto Oswaldo Cruz, e Laboratório de Bioinformática do LNCC (Fig. 12) para solicitar a autorização de ser um nó nacional da EMBnet - European Molecular Biology network.

Em meados de outubro o consórcio foi aprovado e a organização deste está em andamento. A vantagem de ser nó da EMBNet é que esta rede

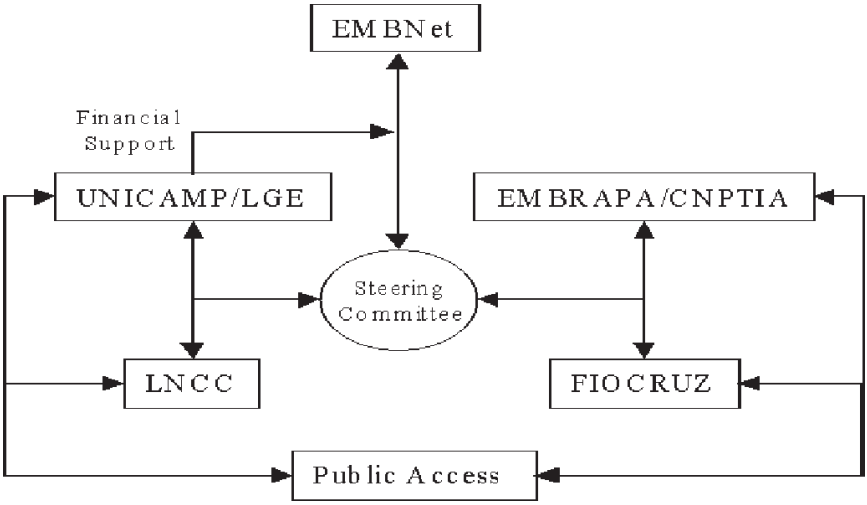


Fig. 12. Grupos participantes do consórcio para nó nacional da EMBNet.

tem uma excelente infra-estrutura para organizar cursos de treinamento, dar assistência e ajudar seus membros a interagirem respondendo rapidamente às necessidades da pesquisa biológica. Além disso fica facilitado o acesso a inúmeros bancos de dados e à logistíca da organização. A web page do consórcio está disponível no endereço: http://www.nbi.cnptia.embrapa.br/Embnet_BR/index.html

Estabelecimento de um Ambiente para Pesquisa e Oferta de Serviços na Área de Bioinformática

O Núcleo de BioInformática Estrutural foi oficialmente inaugurado no dia 21 de outubro de 2002. Durante os primeiros meses de vigência do projeto foi feita a compra e instalação dos equipamentos solicitados para o Núcleo de Bioinformática.

- Os oito PCs adquiridos estão sendo usados por pesquisadores e estagiários para o desenvolvimento de todas as ferramentas do

laboratório, geração de página do site, preparação de cursos, e pesquisa desenvolvida no laboratório.

- As estações *silicon graphics octane* estão sendo usadas para a geração de valores de potencial eletrostático para todos os arquivos de coordenadas do *Protein Data Bank* que são incorporados na ferramenta *Java Protein Dossier do* SMS. Estas estações gráficas também são usadas para visualização e construção de modelos de estruturas de proteínas e outras linhas de pesquisa.
- A SGI Origin 3400, com o disco rígido tipo RAID5 de 876 Gb, é utilizada como web server e para produção de dados dos programas do NBI.

  A web page do Núcleo de Bioinformática Estrutural pode ser acessada pelo endereço http://www.nbi.cnptia.embrapa.br/.

Formação de Recursos Humanos: Organização de Cursos e Congressos

Como a BioInformática ainda é uma área incipiente na maioria dos institutos de pesquisa, há uma necessidade urgente de treinar usuários em vários níveis. Com o objetivo de contribuir com a disseminação do conhecimento nesta área foi realizado um curso de bioinformática estrutural. Também está-se unindo esforços para trazer ao Brasil o maior evento mundial em bioinformática, o ISMB (*Intelligent Systems for Molecular Biology*) para o ano de 2005.

Infelizmente não conseguiu-se fixar alunos interessados em nosso laboratório principalmente por falta de recursos. Veja mais sobre este assunto em Considerações Finais.

ISMB

Está em julgamento, junto a Sociedade Internacional para Computação Biológica (ISCB: www.iscb.org), a proposta para a realização da Conferência Internacional em Sistemas Inteligentes para Biologia Molecular (ISMB) em 2005 no Brasil. O ISMB é o maior evento mundial para a ciência conhecida como Bioinformática. Para que a proposta fosse formalizada foi criado um comitê local formado por 14 membros de

diversas instituições de pesquisa brasileiras que têm interesse no desenvolvimento da área de bioinformática.

A realização de um evento como o ISMB no Brasil certamente trará um

Membros do comitê local

Membro	Afiliação
Goran Neshich - Presidente	Embrapa Informática Agropecuária – NBI
Ana Tereza Ribeiro de Vasconcelos –	Laboratório Nacional de Computação
Vice-presidente	Científica/ LNCC
Paula Kuser Falcão	Embrapa Informática Agropecuária – NBI
Álvaro Seixas Neto	Embrapa Informática Agropecuária – NBI
Gonçalo Amarante Guimarães Pereira	Laboratório de Genômica e Expressão-IB/ Unicamp
João Carlos Setubal	Laboratório de Bioinformática-IC/Unicamp - Allelyx
João Paulo Kitajima	Laboratório de Bioinformática-IC/Unicamp - Allelyx
José Antonio Maranho	Vertical Eventos
Rogério Meneghini	Centro de Biologia Estrutural-Laboratório Nacional de Luz Síncrotron/LNLS
Junior Barrera	Instituto de Matemática e Estatística/USP
Sérgio Verjovski	Instituto de Química/USP
Sandro Joséde Souza	Instituto Ludwig
Glaucius Oliva	Instituto de Física/USP-SC
Richard Charles Garrat	Instituto de Física/USP-SC

grande impacto para a área de Bioinformática despertando o interesse de jovens pesquisadores para as vantagens de uso das ferramentas de bioinformática.

Neste momento o NBI já está com uma página web

(http://www.nbi.cnptia.embrapa.br/LOC) configurada. Esta página tem o papel de veículo eficaz de transmição de informações pertinentes ao evento entre os membros do comitê local e também de alertar o público em geral sobre este evento.

Como a página do laboratório tem visibilidade internacional, a inclusão

do item ISMB_2005 vai criar um certo impacto para a importância da Biologia Computacional - Bioinformática no Brasil em âmbito mundial, e também vai fazer propaganda do evento.

First STING Millennium Suite (SMS) Course: Ferramentas para Analisar Estruturas Macromoleculares e Aplicações em Química-genômica

Com o objetivo de mostrar a funcionalidade do NBI e as facilidades que o laboratório e a área de bioinformática pode oferecer para os grupos envolvidos com estudos de macromoléculas foi oferecido o curso intitulado: "*First*" *STING Millennium Suite* (SMS) *Course: tools to Analyze Macromolecular Structures and Applications in Chemogenomics*" (http://www.nbi.cnptia.embrapa.br/SMScg) para membros de 8 instituições: UFPE, TECPAR, LNLS, LNCC, ESALQ, UNICAMP, Int. Butantã, Escola Paulista de Medicina (Fig. 13). Para acessar a página do curso utilizar login: SMS_cg e password 12345.

O curso aconteceu durante os dias 29 e 30 de outubro de 2002 nas dependências da Embrapa Informática Agropecuária no Núcleo de Bioinformática Estrutural (NBI), abordando os seguintes aspectos:

- Busca de similaridade de seqüências em diferentes bancos de dados;
- Análise estrutural e funcional das proteínas e seus potenciais homólogos;
- Modelagem de moléculas;
- Comparação dos descritores de estrutura e função para as proteínas modeladas e homólogas com estrutura já conhecida usando o SMS;
- Estudos dos contatos críticos na interface macromolecular;
- Estudos dos efeitos das mutações nos complexos de macromoléculas;
- Química Genômica.

Fig. 13. Participantes do curso SMScg.

O curso recebeu ótima avaliação pelos participantes (veja a seguir). Mais detalhes do curso e de sua avaliação podem ser encontrados no URL citado.

0 - péssimo 1-ruim 2 - razoável 3 - bom 4 - muito bom 5 - excelente

Questões	0	1	2	3	4	5	média
Conforto das instalações	0	0	0	4	5	8	4,2
Tempo de duração do curso	0	0	2	3	6	6	3,9
Objetividade e clareza	0	0	1	2	5	9	4,3
Didática na transmissão das informações	0	0	0	3	6	8	4,3
Conhecimento sobre os assuntos abordados	0	0	1	1	3	12	4,5
Nível de aproveitamento do conteúdo							
abordado em minhas atividades profissionais	0	0	0	7	3	6	3,9
Avaliação geral do curso	0	0	0	0	7	10	4,6

Projetos em Colaboração

Os projetos em colaboração com outros grupos são feitos com a intenção de oferecer à comunidade científica as ferramentas da Bioinformática desenvolvidas no SBI/NBI. Assim, pesquisadores de outras áreas têm a oportunidade de empregarem as ferramentas da bioinformática nos seus problemas biológicos. Esta interação é muito benéfica no momento considerando o crescimento exponencial dos dados obtidos através dos projetos genoma e da rápida diversificação das novas aplicações.

Colaboração Nacional

Os projetos que o SBI/NBI gostaria de participar/apoiar em geral, visam a abordagem dos seguintes aspectos:

- busca de similaridade de seqüências em diferentes bancos de dados;
- alinhamento seqüencial e estrutural para determinado grupo das seqüências/estruturas com produção posterior de MSA (Alinhamento Múltiplo de Seqüências) e P_3 (árvores filogenéticas);
- produção de sequência de consenso e motivo estrutural de consenso a partir de MSA e MsTA (Alinhamento Estrutural Múltiplo), respectivamente;
- análise estrutural e funcional das proteínas homólogas em termos de seqüência e estrutura;
- modelagem de moléculas com identidade entre seqüências maior que 30%;
- comparação dos descritores de estrutura e função para as proteínas homólogas (inclusive modeladas) usando o SMS e seus componentes;
- definição das interfaces – partes da superfície das moléculas envolvidas no processo de binding;
- estudos dos contatos críticos na interface macromolecular – reconhecimento de assinatura de especificidade e geração da tabela exaustiva e completa, listando todos os parâmetros importantes para definir especificidade;

- estudos dos efeitos das mutações nos complexos de macromoléculas;
- cálculo de protein e ligand fingerprints;
- determinação dos potenciais alvos para fármacos;
- determinação dos mapas de pareamento proteína/ligante – uma tabela com descrição e listagem das moléculas ligantes (substâncias orgânicas) capazes de fazer intervenção na forma da ação físico química das proteínas/enzimas ou seja, executoras de funcionalidade. Estes são na verdade os potenciais fármacos, herbicídas, inseticidas etc.
- busca por receptores dados os ligantes e busca por ligantes dados os receptores (pesquisa básica).

Todos os procedimentos necessários para atingir os objetivos citados requerem o envolvimento de ambas as partes. O NBI oferece o conhecimento, ferramentas da bioinformática, espaço laboratorial adequado, máquinas e sistemas de rede adequados, bancos de dados e ambiente completo para fazer pesquisa e trabalho *in silico.*

O laboratório colaborador oferece o sistema biológico interessante do ponto de vista de aplicações e possível importância na decifração dos processos básicos, governando o desenvolvimento dos organismos em pauta, assim como o domínio de técnicas de biologia molecular necessárias para obtenção do material necessário para medidas experimentais. Em conjunto, esta estratégia está bem afinada para obtenção dos resultados positivos e publicáveis.

Em geral, espera-se também que o grupo colaborador já tenha o técnico/estudante/bolsista que poderia dedicar seu tempo para este trabalho sob orientação conjunta. Espera-se que as duas equipes sejam autoras de trabalhos publicados em revistas internacionais com índice de impacto acima de 1.5.

No momento está-se firmando colaborações com o Centro de Toxinas Aplicadas (CAT) do Instituto Butantã, com o Laboratório de Biologia Molecular de Plantas do Departamento de Genética da ESALQ/USP e com o Departamento de Bioquímica da UNIFESP/EPM.

O projeto do CAT tem como objetivo a identificação e caracterização de peptídeos potenciadores de bradicinina (BPPs) e/ou peptídeos homólogos expressos em tecidos “fisiológicos” de mamíferos.

Os peptídeos potenciadores de bradicinina (BPPs) presentes no veneno da *Bothrops jararaca* foram os primeiros ACEIs (inibidores da enzima conversora da angiotensina) naturais descritos. Os estudos de estrutura e atividade destes BPPs e análogos foram essenciais para o desenvolvimento dos ACEIs não peptídicos, que são amplamente empregados na medicina atual. Vários BPPs de *B. jararaca*, constituídos de 5 a 13 resíduos de aminoácidos, foram purificados e seqüenciados há muito tempo, sem que no entanto, fosse realizado um estudo genético para a determinação do mecanismo de geração e processamento destes peptídeos.

Na colaboração com o CAT propõe-se: 1) realizar buscas no genoma de mamíferos, no intuito de se identificar outras possíveis seqüências que apresentem alguma semelhança estrutural com os BPPs de serpente já caracterizados, e que possam ser identificados como possíveis correlatos endógenos destas toxinas; e 2) fazer a caracterização dos descritores de estrutura e função para os BPPs conhecidos empregando as ferramentas de bioinformática desenvolvidas no NBI.

Os projetos com os outros dois grupos estão em fase final de elaboração.

Cooperação Internacional

Formou-se um consórcio de colaboração internacional formado por 13 grupos de pesquisa para concorrer ao "The Sixth Framework Programme for Research and Technological Development"" (2002-2006). Este programa visa financiar a criação da European Research Area (ERA). Os grupos representados no consórcio são na sua maioria de países europeus (10), um grupo da China e o NBI, e estão representados pelos pesquisadores: P. Bladon, Glasgow; N. Vermeulen, Leiden; F. Jorgensen, Copenhagen; M. Rarey, Hamburg; T. Langer, Insbruck; V. Gillett e P. Willett, Sheffield; B. Maigret, e C. Chipot, Nancy; F. Cazals, Nice; D. Rognan, Strasbourg; G. Cruciani, Perugia; W. Cai and X. Shao, USTC, Hefei, China; e G. Neshich, NBI, Brasil. Além dos grupos de pesquisa também há a colaboração da empresa Americana AVAKI, Charlottlesville, que é especializada no uso de"" "GRID computing".

Dentro deste consórcio o grupo do NBI participa com o programa SMS para descrever proteínas e formar um banco de dados para fazer o casamento entre as estruturas de proteínas e possíveis grupos de ligantes que poderão ser utilizados como fármacos.

Considerações Finais

Verifica-se uma necessidade urgente de treinar usuários (pesquisadores e estudantes) desde um nível básico até um nível mais avançado do uso das ferramentas de bioinformática, promover a integração entre as áreas de bioinformática e biologia, e de treinar pessoas que consigam desenvolver ferramentas na área de bioinformática.

Planeja-se realizar cursos de bioinformática para os usuários da SmolBNet e como um módulo do curso de pós-graduação do departamento de genética da Universidade Estadual de Campinas – UNICAMP. Espera-se com estes cursos tanto oferecer a idéia da multidisciplinaridade entre as áreas de biologia e bioinformática como atrair estudantes de pós-graduação (mestrado e/ou doutorado) interessados em fazer pesquisa no NBI.

Durante o ano de 2003 espera-se suprir algumas das dificuldades que ocorridas no primeiro ano do projeto. A principal dificuldade foi ter que trabalhar com um número reduzido de pessoas no grupo em relação ao que se planejou inicialmente. Das cinco bolsas que foram solicitadas ao CNPq, somente três foram implementadas. Os pesquisadores começaram a trabalhar no grupo e após dois meses o CNPq informou que as bolsas, embora implementadas, não seriam pagas por dificuldades financeiras, e assim perdeu-se dois estudantes de mestrado e um pós-doutorado. Ao mesmo tempo, a Embrapa passou por uma séria crise financeira e cortou 2/3 dos estagiários de todos os projetos. Esses cortes causaram um grande impacto nas perspectivas de resultados para este ano. Também teve-se um pedido de bolsa de pós-doutorado para um pesquisador que havia defendido seu doutorado no Centro de Biologia Molecular e Engenharia Genética (CBMEG) do Instituto de Biologia da UNICAMP negado pela assessoria da Fapesp.

Em contrapartida a equipe aumentou com a contratação da pesquisadora Paula Regina Kuser Falcão, Ph.D. em cristalografia de proteínas, que passou em concurso realizado pelo Embrapa. A contratação da pesquisadora foi efetivada em maio de 2002. Espera-se que seja efetivada a contratação de mais um pesquisador doutor concursado na área de Matemática Aplicada, Dr. Michel Beleza.

Espera-se também para o próximo ano conseguir recursos financeiros para absorver os alunos e pesquisadores que sejam interessantes para o desenvolvimento da pesquisa feita no NBI.

Referências Bibliográficas

FRISHMAN, D.; ARGOS, P. Knowledge-based protein secondary structure assignment. Proteins: Structure, Function and Genetics, v. 23, n. 4, p. 566-579, 1995.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, v. 22, n. 12, p. 2577-2637, 1983.

NESHICH, G.; TOGAWA, R.; VILELLA, W.; HONIG, B. Sequence To and withIN Graphics PDB Viewer (STING -PDB viewer). PDB Quarterly NewsLetter - Brookhaven National Laboratory, n. 85, p. 6-7, July, 1998. Disponível em: <ftp://ftp.rcsb.org/pub/pdb/doc/newsletters/bnl/news95_jul98/newslttr.pdf>.

NICHOLLS, A.; SHARP, K.; HONIG, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins: Structure, Function and Genetics, v. 11, n. 4, p. 281-296, 1991.

RADZICKA, A.; WOLFENDEN, R. Comparing the polarities of the amino-acids - side-chain distribution coefficients between the vapor-phase, cyclohexane, 1-octanol, and neutral aqueous-solution. Biochemistry, v. 27, p. 1664-1670, 1988.

TSODIKOV, O. V.; RECORD JUNIOR, M. T.; SERGEEV, Y. V. A novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. J. Comput. Chem., v. 23, p. 600-609, 2002.


Embrapa
Informática Agropecuária