RECIIS
Revista Eletrônica de Comunicação Informação & Inovação em Saúde
[www.reciis.cict.fiocruz.br]
ISSN 1981-6278

SUPLEMENTO – BIOINFORMÁTICA E SAÚDE

Artigos originais

A plataforma PDTIS de bioinformática: da seqüência à função

DOI: 10.3395/reciis.v1i2.Sup.98pt

Thomas Dan Otto
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brasil
otto@fiocruz.br

Marcos Catanho
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brasil
mcatanho@fiocruz.br

Wim Degrave
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro - RJ, Brasil
wdegrave@fiocruz.br

Antonio Basílio de Miranda
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro - RJ, Brasil
antonio@fiocruz.br

Resumo

Desde os anos 1980, a Bioinformática tem adquirido uma importância crescente nas ciências biológicas. Devido à complexidade de ferramentas tanto de *hardware* quanto de *software*, e a necessidade de ter-se um gerenciamento especializado de infra-estruturas de rede, processamento e armazenamento de dados, muitas instituições de pesquisa organizaram um núcleo de bioinformática (*core facility*, CPD, Centro de (Bio)informática etc.). Neste artigo, descrevemos uma evolução típica de atividades na área de bioinformática e a organização de um núcleo mínimo com suporte institucional. Desde o final da década de 1980, a Fiocruz tem contribuído para o estabelecimento e desenvolvimento da Bioinformática no Rio de Janeiro, através de diferentes iniciativas. Iniciando com cursos de análise de seqüências ministrados na Fiocruz, hoje a instituição conta com um serviço permanente de Bioinformática, integrado a outras atividades estratégicas desenvolvidas, através de uma rede temática de plataformas com suporte de um Programa de Desenvolvimento Tecnológico em Insumos para a Saúde (PDTIS), cujas principais atribuições concentram-se no suporte, no estabelecimento de novas metodologias e na participação ativa em projetos de desenvolvimento tecnológico.

Neste artigo, descrevemos a Plataforma PDTIS de Bioinformática, o seu escopo, os seus recursos e as suas atividades de suporte, ensino e pesquisa, considerando que muitos outros centros de pesquisa encontram-se em situação similar O surgimento da Bioinformática como uma nova área do conhecimento e um histórico de seu desenvolvimento na Fiocruz também são apresentados, de forma resumida.

Palavras-chave

Bioinformática, biologia computacional, banco de dados, genoma, análise de seqüências

A bioinformática

Nesta seção apresentaremos um resumo da história da Bioinformática, destacando os primeiros programas de análise de seqüências, as primeiras bases de dados, o surgimento de métodos automáticos de seqüenciamento, os projetos genoma, a genômica comparativa e a metagenômica, citando os pontos principais de deste processo.

É possível situar as origens da Bioinformática e da Biologia Computacional na década de 1960, quando avanços tecnológicos permitiram que os computadores emergissem como ferramentas importantes na Biologia Molecular (assim como em todas as outras áreas). Três teriam sido os principais fatores motivadores deste surgimento: (i) o crescente número de seqüências protéicas disponíveis, ao mesmo tempo uma fonte de dados e de questões importantes, porém intratáveis sem o auxílio de um computador; (ii) a idéia de que as macromoléculas carregam informação ter se tornado parte fundamental do modelo conceitual da Biologia Molecular e (iii) a disponibilidade de computadores mais velozes nas principais universidades e centros de pesquisa. Entre os pioneiros da Bioinformática podemos citar a Dra. Margaret Dayhoff, por coletar, organizar e disponibilizar o primeiro atlas de seqüências protéicas em 1965 (DAYHOFF et al., 1965). Outro importante pioneiro é o Dr. Temple Smith, co-autor do algoritmo de programação dinâmica conhecido como Smith-Waterman (SMITH e WATERMAN, 1981) utilizado em ferramentas de buscas por similaridade.

Até o final dos anos 60, diversas técnicas computacionais (algoritmos e programas de computador) para a análise da estrutura, da função e da evolução de seqüências nucleotídicas e protéicas, bem como bancos de dados rudimentares de proteínas, já haviam sido desenvolvidos (HAGEN, 2000; OUZOUNIS e VALENCIA, 2003). Novas técnicas e abordagens foram desenvolvidas nas décadas seguintes, destacando-se: (i) os algoritmos para alinhamento de seqüências; (ii) a criação de bancos de dados de acesso público; (iii) a implementação de sistemas rápidos de busca em bancos de dados; (iv) desenvolvimento de sistemas mais sofisticados para a predição de estrutura de proteínas e (v) ferramentas para a anotação e comparação de genomas e sistemas para análise funcional de genomas (OUZOUNIS, 2002).

Foi somente na década de 1980, no entanto, que a Bioinformática e a Biologia Computacional tomaram forma de disciplinas independentes, com seus próprios problemas e conquistas, sendo a primeira vez em que algoritmos eficientes foram desenvolvidos para lidar com o volume crescente de informação e que implementações destes algoritmos (programas) foram disponibilizadas para toda a comunidade científica (OUZOUNIS e VALENCIA, 2003). Foi durante esta década que surgiram pacotes de programas para a análise de seqüências nucleotídicas e protéicas, como Staden (STADEN, 1977), Pustell (PUSTELL e KAFATOS, 1982) e GCG (Devereux et al., 1984), assim como bases de dados públicas servindo de repositórios para as seqüências e resultados de análises das mesmas, como o GenBank (GENBANK, 2007), EMBL (EMBL, 2007), DDBJ (DDBJ, 2007) e Swiss-Prot (SWISS-PROT, 2007) (hoje parte do UniProt [UNIPROT, 2007]). A afirmação definitiva destas novas disciplinas aconteceu na década de noventa, com o surgimento dos projetos genoma, transcriptoma e proteoma (sustentados por avanços importantes nos métodos de seqüenciamento de DNA, no desenvolvimento de *microarrays* e *biochips* e na espectrometria de massas), das redes de computadores em escala mundial (*Internet*), de bancos de dados biológicos imensos, supercomputadores e computadores pessoais bastante robustos.

Nesse contexto, além do suporte à análise de dados gerados pelas tecnologias de alto desempenho citadas anteriormente, a Bioinformática encontra sua utilidade na solução de problemas relacionados principalmente à execução de determinadas tarefas, tais como análise de seqüências nucleotídicas e protéicas, buscas por similaridade, anotação de genes e genomas, seqüenciamento e montagem de genomas, assim como nas áreas de genômica comparativa, filogenia molecular e modelagem molecular.

Grandes avanços na tecnologia de seqüenciamento de DNA fizeram com que os projetos genoma, iniciativas voltadas para a obtenção da seqüência completa do genoma de um determinado organismo, se tornassem mais acessíveis à comunidade científica. Esta, por sua vez, os vê como uma excelente oportunidade de obter um panorama geral do metabolismo, da bioquímica e da genética do organismo em estudo (maiores informações sobre os projetos genoma podem ser encontradas no banco de dados GOLD [GOLD, 2007]). Inicialmente voltados para o estudo de organismos importantes do ponto de vista médico, industrial ou experimental (organismos-modelos para pesquisa), tais como o seqüenciamento do genoma humano (VENTER et al., 2001; LANDER et al., 2001), do camundongo(MOUSE GENOME SEQUENCING CONSORTIUM et al., 2002), de diversas bactérias como a *Escherichia coli* (BLATTNER et al., 1997) entre outros, hoje, o número de projetos genoma cresce rapidamente com a inclusão de novos organismos, importantes para os pesquisadores inclusive sob outros aspectos como, por exemplo, evolutivos, caso do projeto genoma do mamífero conhecido como ornitorrinco (*Ornithorhynchus anatitus)*. De fato, uma compilação dos projetos genoma no mundo mostra que, no momento, existem 660 genomas completamente seqüenciados, com mais 2.258 projetos em andamento, sendo 60 de arqueobactérias, 1.344 de eubactérias e 854 de eucariotos (GOLD, 2007). Adicionalmente, o uso dessas novas tecnologias de seqüenciamento em amostras ambientais deu origem à chamada metagenômica, onde todo o conteúdo genético presente nestas amostras é seqüenciado. Ao contrário dos projetos genoma usuais, nos quais o DNA de um único organismo é estudado, na metagenômica se procura estudar toda uma comunidade de organismos. Atualmente, 114 projetos metagenômicos estão em andamento (GOLD, 2007).

A plataforma PDTIS de bioinformática

Nesta seção apresentaremos um breve histórico da Plataforma de Bioinformática a partir dos primeiros cursos de análise de seqüência ministrados na Fiocruz. Seguiremos, então,

com a apresentação da Plataforma, na seqüência: (i) ***recursos computacionais*** *– os equipamentos disponíveis aos usuários; (ii)* ***atividades de suporte*** *– os serviços que a Plataforma oferece (bases de dados e programas para análise de seqüências, genômica, modelagem, etc); (iii)* ***atividades de ensino*** *– como a plataforma pode auxiliar o usuário a obter mais informações e melhorar seu desempenho em seu trabalho (cursos, tutoriais, manuais, links para outras fontes, etc.); (iv)* ***atividades de pesquisa*** *– o que a Plataforma já produziu.*

No Rio de Janeiro, os primeiros cursos de Bioinformática começaram a ser oferecidos no Instituto Oswaldo Cruz em 1989, como disciplinas nos cursos de pós-graduação *stricto sensu*, sob a coordenação do Dr. Wim Degrave, do Departamento de Bioquímica e Biologia Molecular, tendo como tema a análise computacional de seqüências nucleotídicas e protéicas. Outros cursos se seguiram, como o curso *Seqüenciamento Automatizado de DNA* (1992) e o *International Training Course on Computer Analysis of Nucleic Acid and Protein Sequences and the Use of Networks* (1992, com apoio da OMS/TDR), utilizando um VAX MX850 baseado no antigo modelo PDP 11. Posteriormente, foram organizados outros cursos nacionais e internacionais, financiados pela Fiocruz,OMS/TDR e FAO. Em 1994, o primeiro servidor de grande porte foi adquirido: *gene.dbbm.fiocruz.br*, um servidor Silicon Graphics Challenge com 4 CPUs, e uma estação gráfica *SGI Indigo2*, substituídos em 1999 pelo servidor SGI Origin2000, chamado *amoeba*. Em 2003, foi criada oficialmente a Plataforma PDTIS de Bioinformática na Fiocruz (PLATAFORMA DE PDTIS BIOINFORMÁTICA, 2007) como parte do Programa de Desenvolvimento Tecnológico em Insumos para a Saúde (PDTIS, 2007).

O PDTIS possui como principal meta o estímulo ao desenvolvimento tecnológico na Fiocruz, através da articulação de redes cooperativas multidisciplinares, visando o desenvolvimento de produtos, processos e serviços com impacto na Saúde Pública e no desenvolvimento econômico e social do Brasil. Simultaneamente, pretende promover e estimular mudanças culturais na própria instituição, através do estabelecimento de pontes entre a pesquisa aplicada, a produção de insumos para a saúde e a gestão tecnológica institucional. O modelo adotado para essa estruturação, o modelo em Redes Cooperativas, visa motivar os pesquisadores a trabalharem de forma organizada e colaborativa em torno de objetivos comuns e de tecnologias similares, obtendo assim resultados significativos quanto à otimização de recursos humanos e financeiros. Este programa é gerenciado através de um Núcleo Gestor, o qual é composto pelos coordenadores do programa, pelos coordenadores das respectivas redes cooperativas e pelas gerências de qualidade, gestão tecnológica e gestão financeira.

Equipadas com modernos aparelhos e operadas por pessoal especializado, as plataformas tecnológicas oferecem não somente serviços de suporte, participando ativamente de projetos de pesquisa, projetos de desenvolvimento tecnológico e redes de plataformas. São caracterizadas por seu alto valor estratégico, o que as torna indispensáveis para os setores públicos e privados. Suas principais atividades concentram-se no suporte, no estabelecimento de novas metodologias e na participação ativa em projetos de pesquisa.

A Plataforma PDTIS de Bioinformática (Figura 1) tem como objetivo principal disponibilizar localmente as principais ferramentas de bioinformática e os principais bancos de dados de seqüências nucleotídicas e protéicas para toda a Fiocruz. Esta disponibilização tem sido feita de duas formas: (i) via *Internet*, oferecendo acesso à maior parte dos programas e das bases de dados disponíveis publicamente no mundo, e (ii) através de acesso ao servidor *bioinfo*, no qual diversos programas e bases de dados estão localmente instalados, destinado principalmente à realização de tarefas computacionalmente mais intensas. Portanto, para usuários com necessidade de analisar uma ou poucas seqüências por vez, recomendamos a utilização de interfaces *web*, as quais proporcionam acesso a diversos programas que podem ser utilizados via *Internet*. Já para os usuários com um volume maior de dados, ou com necessidades especiais, sugerimos a abertura de conta para a utilização dos serviços. As regras e normas para a abertura de contas e utilização da plataforma podem ser acessadas em sua página principal (PDTIS, 2007). É possível a instalação de outras bases de dados e programas; nestes casos o usuário deverá solicitar o serviço diretamente aos administradores da plataforma.

Atualmente, a Plataforma de Bionformática conta com seis servidores para uso da comunidade da Fiocruz e diversos computadores pessoais para alunos e pesquisadores visitantes (Tabela 1). O servidor *bioinfo* hospeda diversos bancos de dados e programas, e é aberto para os usuários interessados, mediante cadastro e abertura de conta; o servidor *genoma*, dedicado à montagem de genomas; o servidor *tremendao*, dedicado ao armazenamento e processamento de seqüências oriundas da Plataforma de Seqüenciamento, e também à hospedagem de banco de dados; o servidor *magneto*, dedicado à execução de serviços *web* e programas *online* como o PISE (LETONDAL, 2001); e os servidores *neptune* e *pluto*, reservados para o processamento de dados e serviços *web*.

Os principais bancos de dados de seqüências nucleotídicas e protéicas, como o GenBank (BENSON et al., 2007), InterPro (MULDER et al., 2007), Swiss-Prot (BOECKMANN et al., 2003), Blocks (HENIKOFF et al., 1999; HENIKOFF et al., 2000), Prosite (HULO et al., 2006), TrEMBL (BOECKMANN et al., 2003), além de diversos outros, encontram-se instalados, sofrendo atualizações semanais, e estão disponíveis para consultas pela comunidade da Fiocruz (Tabela 2). De forma similar, os principais programas para a análise de seqüências e outras tarefas como a montagem de genomas também estão disponíveis, *e.g.* EMBOSS (*European Molecular Biology Open Software Suíte*), um pacote de programas de código livre desenvolvido para atender às necessidades da comunidade de bioinformatas e biólogos moleculares (RICE et al., 2000); BLAST (*Basic Local Alignment and Search Tool*), um algoritmo de alinhamento local para a comparação de seqüências nucleotídicas ou protéicas (ALTSCHUL et al., 1997),

Figura 1 – Página de entrada da Plataforma PDTIS de Bioinformática

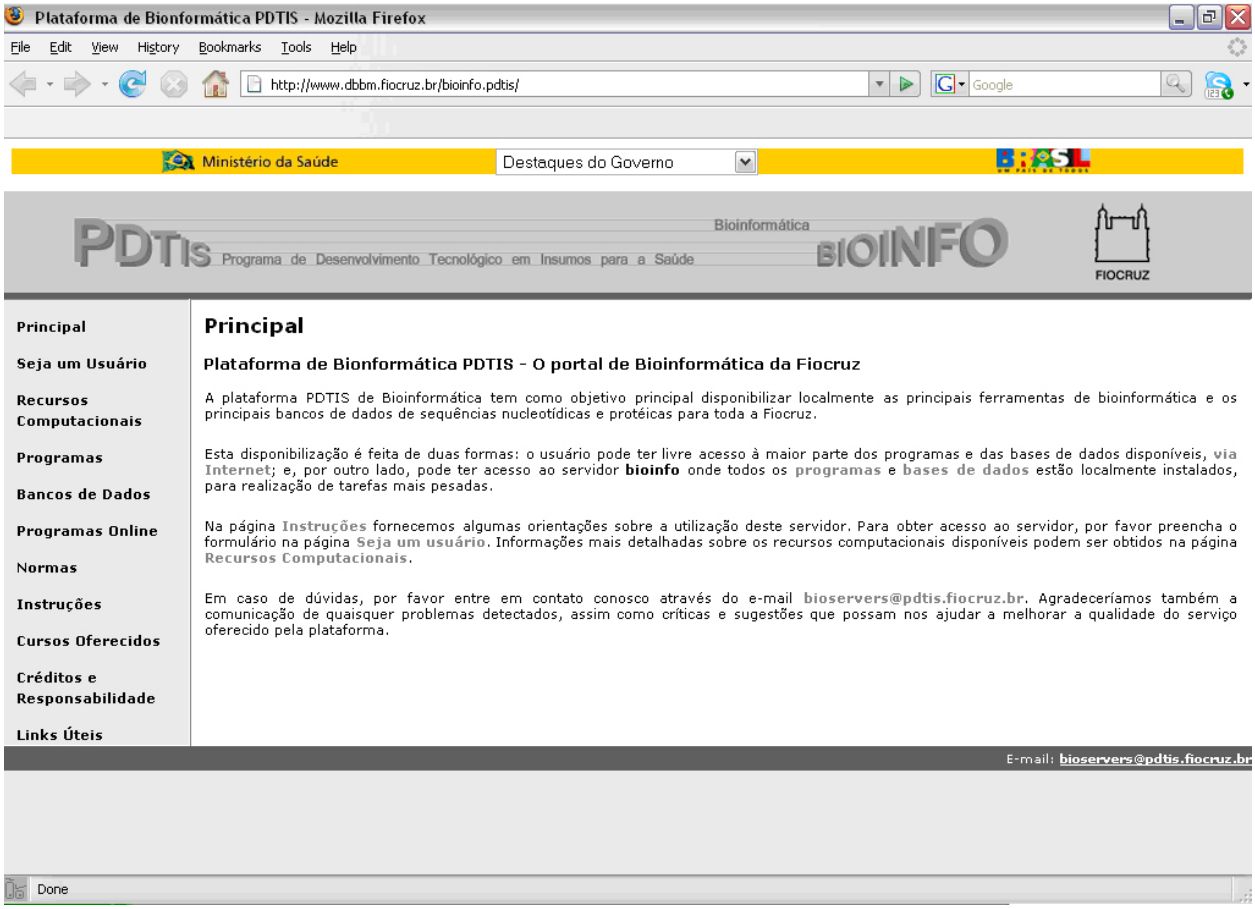

Tabela 1 – Recursos computacionais disponíveis na Plataforma PDTIS de Bioinformática

Computador	Nome	TCP/IP	Serviços	Modelo
Bioinfo	bioinfo.pdtis.fiocruz.br	157.86.152.23	Bancos de dados e programas abertos para todos os usuários	A
Genoma	genoma.pdtis.fiocruz.br	157.86.152.20	Montagem de genomas	A
Tremendao	-	157.86.44.10	Bancos de dados	D
Magneto	-	157.86.44.10:808	Servidor *web*, programas *online* (PISE)	B
SeqAnalist	-	157.86.44.10	SeqScape, análise de seqüências	C
Proteoma	proteoma.pdtis.fiocruz.br	157.86.44.32	Programas para a análise de proteomas	B
Neptune	-	157.86.176.101	Processamento de dados	B
Pluto	-	157.86.176.102	Processamento de dados	B

Configuração dos computadores

A: SunFire v65-X com 2 processadores Intel Xeon de 3,2 gigahertz (GHz) e 5 gigabytes (GB) de RAM com 1 disco de 36 GB e 3 discos de 72 GB.

B: Intel Pentium-4 3 GHz, 1 GB de RAM, 80 GB de disco.

C: Intel Pentium-4 3 GHz, 256 MB de RAM, 30 GB de disco.

D: Athlon 3 GHz com processador de 64 bits, 1 GB de RAM, 200 GB de disco.

Tabela 2 – Bancos de dados instalados e disponíveis para consultas na Plataforma PDTIS de Bioinformática

Nome	Descrição	Tamanho*
AAindex	Índices de aminoácidos e matrizes de similaridade.	1.008 KB
BIOCATAL	Catálogo de softwares para biologia molecular.	2,1 MB
BLAST	Bancos de dados para a execução do BLAST.	32 GB
Blocks	Banco de dados de alinhamentos múltiplos, sem *gaps*, correspondendo às regiões mais conservadas em um grupo de proteínas relacionadas.	1,1 GB
ChEBI	*Chemical Entities of Biological Interest* (ChEBI): dicionário de pequenas entidades moleculares como átomos, moléculas, íons, radicais, etc., identificáveis como entidades distintas. Podem ser naturais ou produtos sintéticos.	201 MB
CPGISLE	Banco de dados de ilhas CpG.	21 MB
CUTG	Compilação da utilização de códons para cada organismo representado no GenBank, a partir das seqüências disponíveis no mesmo.	689 MB
GO	Vocabulário controlado de termos para a descrição de atributos de produtos gênicos.	2,4 GB
InterPro	Banco de dados de famílias protéicas, domínios e sítios funcionais onde características identificáveis encontradas em proteínas conhecidas podem ser aplicadas a seqüências protéicas desconhecidas.	4,2 GB
KEGG	Banco de dados de enzimas e de vias metabólicas.	29 GB
MassSpecDB	Banco de dados de seqüências protéicas não-idênticas, compilado a partir de diversas fontes primárias.	2,5 GB
PDB	Repositório para o processamento e distribuição de dados sobre a estrutura tri-dimensional de proteínas e ácidos nucléicos.	148 MB
PeptideSearch	Dados de espectrometria de massa.	53 MB
Pfam	Coleção de alinhamentos múltiplos e *Hidden Markov Models* cobrindo várias famílias e domínios protéicos.	22 GB
Prints	Banco de dados de famílias protéicas e domínios.	59 MB
PROSITE	Banco de dados de famílias protéicas e domínios.	192 MB
RefSeq	Banco de dados de referência do NCBI.	12 GB
RESID	Banco de dados de modificações em proteínas. É uma coleção de anotações e estruturas para modificações protéicas pós-traducionais.	3,5 MB
Swiss-Prot	Banco de dados de proteínas.	6,8 GB
TRANSFAC	Banco de dados de fatores de transcrição eucarióticos e seus sítios de ligação ao DNA.	9,5 MB
TrEMBL	Suplemento do Swiss-Prot, de anotação automática, contendo as traduções conceituais de todas as seqüências nucleotídicas do EMBL não integradas ao Swiss-Prot.	1,8 GB
UniGene	Sistema experimental para a partição automática do GenBank em um conjunto não-redundante de *clusters* orientados por gene.	8,6 GB
UTR	Banco de dados de regiões não-traduzidas de mRNAs eucarióticos.	872 MB

* Espaço ocupado por cada uma das bases de dados em 08 de outubro de 2007. O tamanho dos bancos varia de acordo com suas versões, aumentando a cada novo lançamento. KB: kilobytes; MB: megabytes; GB: gigabytes.

Tabela 3 – Programas disponíveis na Plataforma PDTIS de Bioinformática

Nome	*Descrição*	*Comandos básicos*
Busca por similaridade		
BLAST	O programa BLAST executa buscas por similaridade local de uma ou mais seqüências em bancos de dados de ácidos nucléicos ou proteínas.	blastall
WU-BLAST	O programa WU-BLAST executa buscas por similaridade local de uma ou mais seqüências em bancos de dados de ácidos nucléicos ou proteínas.	blastn blastp blastx tblastn tblastx

cont.

Tabela 3 – Programas disponíveis na Plataforma PDTIS de Bioinformática (cont.)

FASTA	O programa FASTA executa buscas baseadas no algoritmo de Pearson e Lipman, visando a detecção de similaridade local entre uma determinada seqüência e um grupo de seqüências do mesmo tipo (ácidos nucléicos ou proteínas).	fasta34
SSEARCH	O programa SSEARCH executa uma busca rigorosa, baseada no algoritmo de Smith-Waterman, visando a detecção de similaridade local entre uma determinada seqüência e um grupo de seqüências do mesmo tipo (ácidos nucléicos ou proteínas).	ssearch34
HMMER	Programa para buscas por similaridade em bancos de dados que utiliza descrições estatísticas do consenso obtido para uma família de seqüências (*Profile Hidden Markov Models*).	hmmalign hmmcalibrate hmmsearch hmmbuild
Alinhamentos múltiplos		
ClustalW	Programa para alinhamento (global) múltiplo de seqüências nucleotídicas ou protéicas.	clustalw
T-Coffee	Programa para alinhamento (global) múltiplo de seqüências nucleotídicas ou protéicas.	t_coffee
MUMmer	Sistema para alinhamento múltiplo e visualização de genomas inteiros, ou grandes segmentos de ADN ou proteína, de forma computacionalmente rápida e eficiente.	mummer nucmer promer run-mummer1 run-mummer3
SeaView	Editor gráfico de alinhamentos múltiplos de seqüências.	seaview
Montagem de seqüências		
AMOS	Pacote de programas de código livre que oferece uma infra-estrutura para o desenvolvimento de ferramentas para montagem de seqüências.	runAmos
PCAP	Programa para a montagem de seqüências nucleotídicas baseado na detecção de sobreposições (overlaps) entre os fragmentos.	pcap
CAP3	Programa para a montagem de seqüências nucleotídicas baseado na detecção de sobreposições (*overlaps*) entre os fragmentos.	cap3
Phred/Phrap/ Consed	Pacote para a leitura de cromatogramas, atribuição de bases, atribuição de qualidade a cada base (Phred), montagem de seqüências (Phrap), visualização, edição e finalização das montagens (Consed).	phredPhrap phredPhrap.poly consed
TIGR Assembler	Ferramenta para a montagem de grandes conjuntos de dados de seqüências como ESTs, BACs ou pequenos genomas.	TIGR_Assembler
Anotação de seqüências		
Glimmer	Sistema desenvolvido para a busca de genes em DNA microbiano, especialmente genomas de bactérias, arqueobactérias e vírus.	glimmer3
Artemis	Ferramenta gratuita de visualização e anotação de genomas.	art
ACT	*Artemis Comparison Tool*. Ferramenta para a visualização interativa de comparações entre seqüências genômicas e suas anotações.	act
Programas diversos		
EMBOSS	Pacote gratuito de programas para análise de seqüências.	wossname
REPuter	Família de programas usados para a detecção e visualização de repetições em genomas inteiros ou cromossomos.	repfind repvis repselect
PerlPrimer	Interface gráfica de código livre, escrita em linguagem Perl, desenvolvida para auxiliar o desenho de *primers* para PCR padrão, PCR *real-time* e seqüenciamento.	perlprimer.pl
CodonW	Programa desenvolvido para simplificar a análise multivariada (análise de correspondência) de códons e utilização de aminoácidos.	codonw

HMMER (*Hidden Markov Modeler*), uma implementação de código aberto de um algoritmo para a construção de PHMMs, *Profile Hidden Markov Models*, que podem ser usados em buscas por similaridade (EDDY, 1998); ClustalW, um programa para o alinhamento múltiplo de seqüências (THOMPSON et al., 1994; CHENNA et al., 2003), entre outros (Tabela 3).

Além de todas estas bases de dados e programas, a Plataforma de Bioinformática tem gerado outras bases e aplicativos novos desenvolvidos pela equipe e de utilização acadêmica livre. Podemos citar o BioParser (CATANHO et al., 2006a), programa que analisa o resultado de programas de buscas por similaridade, como o BLAST e o FASTA (PEARSON e LIPMAN, 1988; PEARSON, 1990), gerando uma base de dados local que pode ser consultada através de uma interface gráfica *web*; o GenoMycDB (CATANHO et al., 2006b), banco de dados para análise comparativa de genes e genomas de micobactérias; o LocalCOG (LOCALCOG, 2005), implementação relacional do banco de dados COG (TATUSOV et al., 1997; 2003), de instalação local, com interface desenvolvida para a elaboração de consultas complexas; o Squid (CARVALHO et al., 2005), programa para o compartilhamento de recursos computacionais, permitindo a execução do programa BLAST de forma distribuída em computadores pessoais comuns, maximizando a utilização dos recursos computacionais disponíveis; entre outros em distintos estágios de desenvolvimento (PDTIS, 2007; THE BIOINFORMATICS TEAM, 2007).

A Plataforma de Bioinformática tem atuado na formação de recursos humanos, tanto através da formação de Mestres e Doutores (até o momento já foram finalizadas três dissertações de Mestrado, e três teses de Doutorado encontram-se em andamento), como através do oferecimento de cursos. Com diferentes graus de regularidade, são oferecidos cursos *Análise de Seqüências Nucleotídicas e Protéicas*, *Introdução a Programação em PERL* e o curso *Origem, Estrutura e Evolução dos Genomas*. Também oferecemos cursos básicos e introdutórios de bioinformática, nos chamados cursos de inverno e verão (CURSOS DE FÉRIAS DO IOC, 2007), cursos de extensão promovidos pelo Programa de Pós-Graduação em Biologia Celular e Molecular (PGBCM) do Instituto Oswaldo Cruz/Fiocruz.

Adicionalmente, as linhas de pesquisa atuais da equipe incluem (i) classificação funcional e anotação de proteínas, vias metabólicas, enzimas análogas, pseudogenes e seqüências repetitivas; (ii) modelagem de dados biológicos; (iii) desenvolvimento de novos métodos para montagem de genomas e (iv) anotação de genes de tripanossomatídeos e micobactérias. Recentemente, foi terminada a fase de cálculo do *Genome Comparison Project* (GENOME COMPARISON PROJECT, 2007) no qual foram comparados exaustivamente todos as seqüências protéicas preditas a partir de todos os genomas completamente seqüenciados disponíveis. Os resultados destas comparações, que permitem a solução de diversos problemas de anotação assim como o desenvolvimento de diversos estudos em filogenia e evolução molecular, serão disponibilizados para a comunidade científica em breve.

Perspectivas para o futuro

Nesta seção apresentaremos as perspectivas para a Plataforma de Bioinformática.

Como foi dito, a Plataforma de Bioinformática deve atuar no suporte aos projetos de pesquisa, aos projetos de desenvolvimento tecnológico e também às redes de plataformas, oferecendo subsídios e sinergismo. Ao longo desse período em que a Plataforma de Bioinformática esteve operacional, foram identificados alguns pontos que devem ser trabalhados para que a plataforma mantenha seu padrão de qualidade.

Um ponto principal visa melhorar a interação da plataforma com o usuário. Críticas ocasionais têm como base a falta de documentação em um nível de detalhamento apropriado que permita ao usuário iniciante a plena utilização dos recursos da plataforma. Apesar de a plataforma possuir diversos tutoriais e ponteiros para outros *sites* e listas de referências descrevendo seus programas e bancos de dados, em sua maioria estes documentos não atendem à necessidade do usuário iniciante, que demanda um protocolo descrevendo a utilização dos mesmos passo a passo. Por isso, gradualmente, são elaborados e disponibilizados no *site* da plataforma protocolos simplificados descrevendo as etapas necessárias para a execução de diversas tarefas como, por exemplo, a montagem de fragmentos de seqüência, a identificação de *Single Nucleotide Polymorphisms* (SNPs), a execução e análise de buscas por similaridade etc. Isto irá facilitar em muito a utilização dos programas e bases de dados por usuários não familiarizados com os mesmos. Adicionalmente, é importante fortalecera participação nos cursos de pós-graduação, oferecendo de forma regular cursos voltados para diversos aspectos da Bioinformática, disseminando o conhecimento e atraindo mais usuários.

Também é necessário fortalecer a interação com outras plataformas, implementando, por exemplo, um *pipeline* de agrupamento (*clustering*) para a detecção de SNPs a partir do resultado da Plataforma de Genômica. A disponibilização de um servidor com placa gráfica de alto desempenho é necessária para atender aos requisitos dos interessados em modelagem molecular.

Em relação ao parque computacional, a aquisição de um *cluster* de máquinas com alto poder de processamento, ideal para processos custosos do ponto de vista computacional, e que possam ser configuradas para trabalhar em paralelo pode ser oportuno. De forma similar, estamos idealizando a construção de um *grid* na Fiocruz, o qual utilizará a capacidade ociosa dos computadores pessoais espalhados pela instituição (de forma similar ao *World Community Grid* [WORLD COMMUNITY GRID, 2007]), sem afetar o trabalho diário de cada voluntário participante. Ao aproveitar a capacidade ociosa (de centenas de máquinas), haveria ganhos imensos, pois, na verdade, a maioria dos computadores trabalha por apenas pequenos intervalos de tempo, permanecendo em repouso enquanto aguardam por uma resposta do usuário.

A Plataforma de Bioinformática seguirá proporcionando serviços de qualidade, oferecendo à instituição os principais programas e bases de dados utilizados em

Biologia Molecular e Bioinformática, sempre procurando alternativas para melhorar o seu atendimento, facilitando o desenvolvimento das pesquisas de seus usuários.

Referências bibliográficas

ALTSCHUL, S.F et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, v.25, n.17, p.3389-402, 1 set. 1997.

BENSON, D.A et al. GenBank. **Nucleic Acids Research**, v.35, p.D21-D25, jan. 2007.

BOECKMANN, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. **Nucleic Acids Research**, v.31, n.1, p.365-70, 1 jan. 2003.

CATANHO, M.; MASCARENHAS, D.; DEGRAVE, W.; MIRANDA, A.B. BioParser: a tool for processing of sequence similarity analysis reports. **Applied Bioinformatics**, v.5, n.1, p.49-53, 2006a.

CATANHO, M.; MASCARENHAS, D.; DEGRAVE, W.; MIRANDA, A.B. GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. **Genetic Molecular Research**, v.5, n.1, p.115-126, 31 mar. 2006b.

CHENNA, R. et al. Multiple sequence alignment with the Clustal series of programs. **Nucleic Acids Research**, v.31, n.13, p.3497-3500, 1 jul. 2003.

CURSOS DE FÉRIAS DO IOC. Disponível em: <http://bioinfo.pdtis.fiocruz.br/cursosdeferias_ioc/> Acesso em: 14 nov. 2007.

DAYHOFF, M.O. et al. (Orgs.). **Atlas of Protein Sequence and Structure**. Maryland, EUA: National Biomedical Research Foundation/Silver Spring, 1965.

DDBJ. Disponível em: <http://www.ddbj.nig.ac.jp/> Acesso em: 08 out. 2007.

EMBL. Disponível em: <http://www.ebi.ac.uk/embl/> Acesso em: 08 out. 2007.

GENBANK. Disponível em: <http://www.ncbi.nlm.nih.gov/Genbank/> Acesso em: 08 out. 2007.

GENOME COMPARISON PROJECT. Disponível em: <http://www.worldcommunitygrid.org/projects_showcase/archives/fgc/viewFgcOverview.do> Acesso em: 08 out. 2007.

GOLD. Genomes Online Database. Disponível em: <http://www.genomesonline.org/> Acesso em: 08 out. 2007.

HAGEN, J.B. The origins of bioinformatics. **Nature Reviews Genetics**, v.1, n.3, p.231-236, dez. 2000.

LANDER, E.S. et al. Initial sequencing and analysis of the human genome. **Nature,** v.409, n.6822, p.860-921, 15 fev. 2001.

LOCALCOG. Relational database system for local use of COG data, 2005. Disponível em: http://www.dbbm.fiocruz.br/labwim/bioinfoteam/index.pl?action=softwares. Acesso em: 2005.

OUZOUNIS, C.A. Bioinformatics and the theoretical foundations of molecular biology. **Bioinformatics,** v.18, n.3, p.377-378, mar. 2002.

OUZOUNIS, C.A.; VALENCIA, A. Early bioinformatics: the birth of a discipline – a personal view. **Bioinformatics**, v.19, n.17, p.2176-2190, 22 nov. 2003.

PDTIS. Plataforma de Desenvolvimento Tecnológico em Insumos para a Saúde. Disponível em: <http://www.presidencia.fiocruz.br/vppdt1/pdtis.php> Acesso em: 08 out. 2007.

PEARSON, W.R.; LIPMAN, D.J. Improved tools for biological sequence comparison. **Proceedings of National Academy of Science**, v.85, n.8, p.2444-2448, abr. 1988.

PEARSON, W.R. Rapid and sensitive sequence comparison with FASTP and FASTA. **Methods Enzymol**, v.183, p.63-98, 1990.

PLATAFORMA PDTIS DE BIOINFORMÁTICA. Disponível em: <http://www.dbbm.fiocruz.br/bioinfo.pdtis/> Acesso em: 08 out. 2007.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: the European Molecular Biology Open Software Suite. **Trends in Genetics**, v.16, n.6, p.276-277, jun. 2000.

SMITH, T.F., WATERMAN, M.S. Comparison of Biosequences. **Advances in Applied Mathematics**, v.2, p.482-489, 1981.

SWISS-PROT. Disponível em: <http://www.expasy.ch/sprot/> Acesso em: 08 out. 2007.

TATUSOV, R.L.; KOONIN, E.V.; LIPMAN, D.J. A genomic perspective on protein families. **Science**, v.278, n.5338, p.631-637, 24 out. 1997.

TATUSOV, R.L. et al. The COG database: an updated version includes eukaryotes. **BMC Bioinformatics**, v.4, n.1, p.41, 11 set. 2003.

THE BIOINFORMATICS TEAM. Disponível em: <http://www.dbbm.fiocruz.br/labwim/bioinfoteam/> Acesso em: 08 out. 2007.

THOMPSON, J.D.; HIGGINS, D.G.; GIBSON, T.J. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Research**, v.22, n.22, p.4673-4680, 11 nov 1994.

UNIPROT. Disponível em: <http://www.expasy.uniprot.org/> Acesso em: 08 out. 2007.

VENTER, J.C. et al. The sequence of the human genome. **Science**, v.291, n.5507, p.1304-1351, 16 fev. 2001.

WORLD COMMUNITY GRID. Disponível em: <http://www.worldcommunitygrid.org/> Acesso em: 08 out. 2007.

Sobre os autores

Thomas D. Otto

Mestre em Ciências da Informática pela Universität zu Lübeck, Alemanha. Atualmente é doutorando em Biologia Celular e Molecular pelo Instituto Oswaldo Cruz (IOC/Fiocruz), onde desenvolve tese na área de Bioinformática. Tem experiência nas áreas de Informática, Bioinformática e Biologia Molecular, atuando principalmente nos seguintes temas: montagem de genomas, análise comparativa de genomas e evolução e desenvolvimento de algoritmos e aplicativos para genômica comparativa e funcional de procariotos.

Marcos Catanho

Mestre em Biologia Celular e Molecular pelo Instituto Oswaldo Cruz (IOC/Fiocruz) e Bacharel em Farmácia pela Universidade Federal do Rio de Janeiro (UFRJ). Atualmente é doutorando em Biologia Celular e Molecular pelo Instituto Oswaldo Cruz (IOC/Fiocruz), onde desenvolve tese na área de Bioinformática. Tem experiência nas áreas de Biologia Molecular e Bioinformática, atuando principalmente nos seguintes temas: análise comparativa de genomas e evolução e desenvolvimento de algoritmos e aplicativos para genômica comparativa e funcional de procariotos.