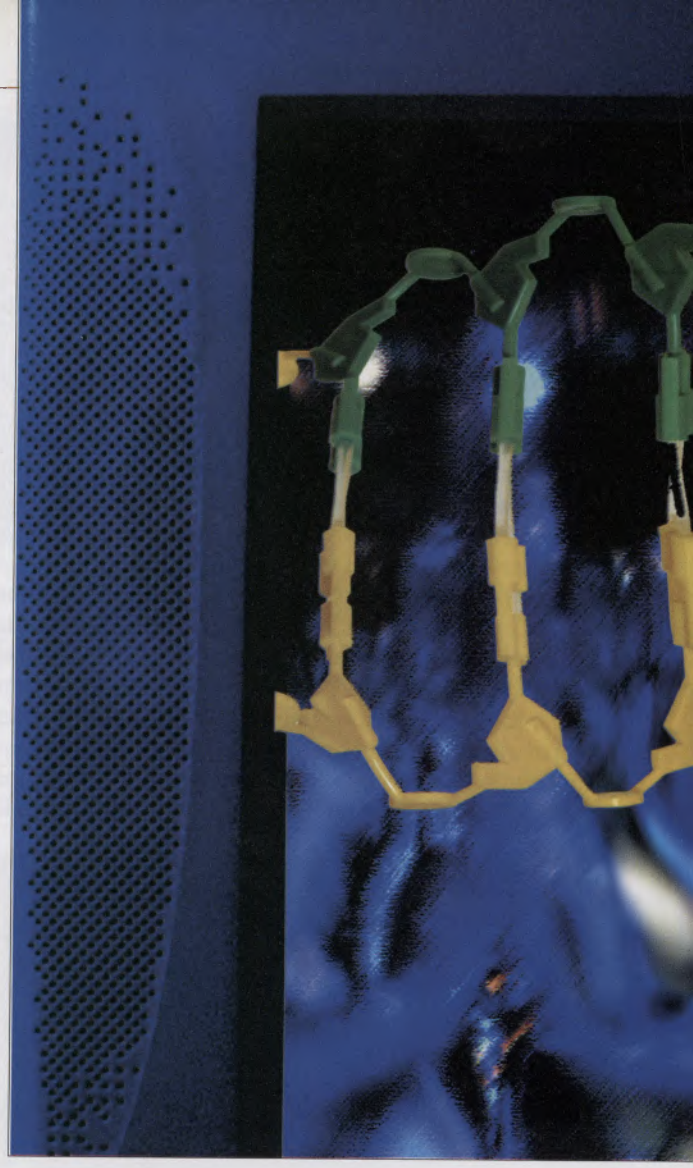
BIOINFORMÁTICA

Bits, bytes e genes

Brasileiros criam programas que simplificam a tarefa de montar e analisar genomas



ESTweb, ZERG e Sabiá: programas nascidos dos projetos de seqüenciamento

Em setembro, duas equipes de pesquisadores brasileiros publicaram artigos científicos em revistas internacionais sobre o genoma (conjunto de genes) de dois organismos, o parasita *Schistosoma mansoni*, causador da esquistossomose no Brasil, e a bactéria *Chromobacterium violaceum*, abundante no rio Negro e com potencial de uso biotecnológico. Embora tenham trabalhado de forma independente, com organismos e metodologias distintos, ambos os times desenvolveram programas de computador que organizaram e facilitaram a obtenção dos dados divulgados em seus escritos.

Do laboratório de Bioinformática do Instituto de Química da Universidade de São Paulo (IQ-USP), que participou do projeto sobre o verme da esquistossomose, saíram dois programas, o ESTweb e o ZERG, já disponíveis para download gratuito no endereço eletrônico *http://verjo19.iq.usp.br/tools.php*. Uma terceira ferramenta, o Sabiá, foi concebida no Laboratório Nacional de Computação Científica (LNCC), de Petrópolis, onde funcionou o coração da bioinformática da iniciativa que estudou os genes da *C. violaceum*. Por ora, o uso do sistema se restringe aos 25 laboratórios da rede nacional que seqüenciou o genoma da bactéria. Mas, em breve, sua utilização deverá ser aberta a todos os interessados.

Seqüências limpas - Cada um desses programas executa tarefas bastante específicas e servem a propósitos particulares. O ESTweb, que rendeu um artigo científico em 12 de agosto de 2003 na revista *Bioinformatics*, recebe e processa os fragmentos de genes ativos gerados a partir de tecidos de um organismo e os aloca num banco de dados. Esses pedaços de genes são chamados ESTs, sigla de etiquetas de seqüências expressas, que serviram de inspiração para batizar o programa. O ESTweb retira dos fragmentos de genes todos os elementos desnecessários para a análise da seqüência e, dessa forma, obtêm-se ESTs mais limpas. "O programa gera em tempo real gráficos que mostram a qualidade e o grau de redundância das seqüências produzidas pelos laboratórios", comenta Sergio Verjovski-Almeida, do Instituto de Química da USP, coordenador da iniciativa financiada pela FAPESP, que identificou 92% dos genes expressos do *S. mansoni*.

A segunda criação da equipe paulista é uma ferramenta de caráter mais analítico. "O ZERG interpreta a saída do BLAST", diz o biólogo Eduardo Reis, um dos inventores do software, recorrendo ao jargão da bioinformática. O BLAST é um programa de domínio público, popular entre biólogos moleculares e outros profissionais que trabalham com genes e proteínas. Sua

EDUARDO CÉSAR

função é comparar qualquer EST com as seqüências genéticas depositadas nos bancos de dados públicos. Assim, o pesquisador descobre se suas ESTs são iguais ou semelhantes a outras já conhecidas e, em muitos casos, consegue associar essas seqüências a genes com funções definidas. Embora muito útil, o BLAST tem um probleminha: em grandes empreitadas, como no projeto do *S. mansoni*, gera um relatório quilométrico, de difícil compreensão, com muitos dados a serem checados.

Destrinchar esse balanço não é uma tarefa para seres humanos, mas para outro software. "Existem programas comerciais que lêem a resposta do BLAST, mas não com a mesma precisão e velocidade do ZERG", diz o programador Apuã Paquola, do Laboratório de Bioinformática do IQ/USP. Num artigo publicado em 22 de maio de 2003 na *Bioinformatics*, os autores do ZERG, cujo nome foi emprestado de um jogo para computador, mostraram que seu invento é até 250 vezes mais rápido do que seus concorrentes.

Apesar do nome brasileiríssimo, o terceiro programa é um acrônimo de uma expressão em inglês: Sabiá quer dizer *System for Automated Bacterial Integrated Annotation*. O programa, que serve para montar e anotar apenas genomas de bactérias, foi concebido no LNCC e usado pela primeira vez durante o trabalho de seqüenciamento da *C. violaceum*, projeto financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

É preciso alguma noção de certos procedimentos básicos do mundo da genômica para ter uma idéia do que faz o programa. Seqüenciar um genoma é determinar a ordem em que aparecem os seus pares de bases nitrogenadas, as unidades químicas primordiais que formam a molécula de ácido desoxirribonucléico, o DNA, e costumam ser representadas pela letras A (adenina), C (citosina), G (guanina) e T (timina). Como o genoma de um organismo pode ser muito grande para ser seqüenciado de uma única vez – o da *C. violaceum* tem, por exemplo, 4,7 milhões de pares de bases –, os pesquisadores têm de quebrá-lo em pedaços pequenos. A exemplo do que se faz com as peças de um quebra-cabeça, sua montagem consiste em juntar, de forma correta, essas partes menores, devidamente seqüenciadas. "Durante o processo de montagem, o Sabiá aponta as regiões do genoma em que os dados gerados pelo seqüenciamento são de boa ou má qualidade", diz Ana Tereza Vasconcelos, do LNNC, coordenadora do projeto com a *C. violaceum* e uma das autoras do software.

Macacos e homens - Concluído o quebra-cabeça da montagem, o dispositivo inventado pela equipe de pesquisadores que trabalhou com o material genético da *C. violaceum* inicia a anotação do genoma. Em linhas gerais, essa tarefa equivale a descobrir que proteínas são produzidas a partir das receitas químicas contidas nos genes de um genoma. Dessa forma, chega-se à função (ou funções) de um gene. Grande parte dos dados de anotação deriva de comparações. Com o auxílio de programas, como o Sabiá e outros de uso gratuito ou pago, os cientistas confrontam o material genético recém-identificado num organismo com seqüências já conhecidas, com função definida, que se encontram arquivadas em bancos de dados públicos. Se, no macaco, uma dada seqüência leva à produção de uma proteína qualquer, chamada, digamos, X, é provável que uma seqüência parecida, se presente no homem, também leve à síntese dessa mesma proteína X.

Claro que as coisas não são tão simples assim, mas esse é o espírito da anotação. "O Sabiá funciona num ambiente computacional que permite cruzar as informações de oito bancos de dados públicos", diz Ana Tereza. "Dá até para comparar genomas inteiros." Para aumentar sua autonomia de vôo, o Sabiá, que deverá ser alvo de um artigo científico neste ano, será aprimorado. A idéia é produzir uma versão do sistema que sirva também para a montagem e anotação de genomas de outros organismos, além das bactérias. •