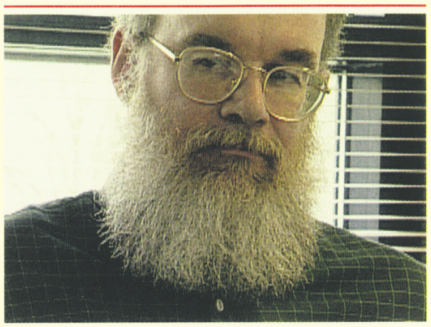
Os computadores estão revolucionando a Biologia

Phil Green, matemático, 48 anos, trabalha em Seattle, na Universidade de Washington, como Leroy Hood; e é o nome mais importante entre os pesquisadores de bioinformática – a nova especialidade inerente à transformação da Biologia numa ciência que lida com grandes quantidades de dados, produzidos em massa por seqüenciadores automáticos de grande capacidade. Doutor Green e seus colaboradores criaram a maior parte das ferramentas em uso nos programas genoma. Das soluções que conseguirem encontrar, daqui para a frente, para tornar automáticas mais e mais tarefas típicas do seqüenciamento de moléculas de DNA, depende a rapidez com que os resultados serão alcançados e sua precisão.
A profundidade da nova ligação entre as ciências da computação e a Biologia, e a forma pela qual ela se dá, são o assunto da conversa com o professor norte-americano.



TERCIO PISANI

Phil Green

■ *Um dos desafios da biologia molecular contemporânea é ligar duas linguagens – a da biologia e a das ciências da computação. O senhor pensa que este é realmente um dos problemas importantes da área?*

— Sim. Vou exemplificar com o meu caso. Fui treinado como matemático, mas desde a escola secundária me interesso por genética – um dos aspectos mais matemáticos da Biologia. O que está acontecendo agora na Biologia é que ela está se tornando um disciplina mais quantitativa, como a Química e a Física. Esta tendência já se delineava há algum tempo, mas acelerou-se bastante nos últimos dez anos, quando seqüenciar o DNA tornou-se mais e mais importante. Muito mais informação biológica tem sido gerada, e de muitos tipos. Um novo problema emergiu: como analisar os dados, qual o método quantitativo adequado para fazê-lo. Se olharmos para o futuro desde esse ponto de vista, nós estamos no começo do caminho que vai tornar a Biologia, de fato, uma ciência quantitativa. Nos projetos genoma, tentamos identificar diferentes componentes moleculares presentes nas células – as proteínas em particular. Quando dispusermos dessa lista de componentes, então o desafio – aliás, muito maior do que o desafio de seqüenciar o DNA – será entender como esses componentes interagem entre si para fazer um organismo. Para isso, será preciso usar não só as idéias da ciência da computação, mas modelamento matemático, modelamento estatístico, e desenvolver métodos inteiramente novos para entender como funciona a interação entre as moléculas. Até agora, muitos biólogos chegavam à Biologia porque queriam ser cientistas, mas sentiam-se um tanto desconfortáveis com os métodos quantitativos. A Biologia foi o campo ideal para alguém que não queria trabalhar com números ou computadores, mas agora, tudo mudou. É um dado cultural.

■ *Mas o senhor também não é um matemático no sentido clássico...*

— É verdade. Há uma transição a ser feita também para os matemáticos. Nós tendemos a idealizar os problemas

muitas observações já feitas que não compreendemos. Novos experimentos devem ser tentados para esclarecer o assunto.

■ *O senhor publicou um paper em que afirma seu ceticismo em relaçao à estratégia proposto pela Celera e por Craig Venter para sequenciar o genoma humano. O senhor poderia nos falar disso?*

— A questão central são as repetições, as seqüências repetidas do genoma humano. Na estratégia em uso atualmente, os grupos tomam clones de 150 mil pares de bases, quebram esses clones em pedaços, seqüenciam, e depois montam. Pois mesmo nessa escala, com clones desse tamanho, os grupos encontram problemas com as seqüências repetidas. Na maior parte das vezes, os programas de montagem conseguem lidar com os *repeats*; mas há casos realmente difíceis, especialmente quando você tem repetições relativamente longas, que ocorrem em vários lugares com seqüências quase idênticas. Isso quando você vê o problema em pequena escala, na escala desses clones de 150 mil pares de bases. Quando você aumenta a escala, o problema simplesmente cresce de magnitude. Quanto maior o pedaço de DNA, maior a probabilidade de você ter repetições nele. Portanto, eu sou cético: não é factível realizar a montagem na escala do genoma humano inteiro. Por outro lado, acho que a Celera na verdade não vai fazer a montagem de todo o genoma a partir dos seus próprios dados. Eles não precisam fazer isto, porque o projeto público está gerando dados que não são finalizados imediatamente, o que quer dizer que os dados parciais ficam disponíveis. O que a Celera fará, acho, é combinar os dados que eles vão obter com o *shot-gun* com os dados do projeto público. Isto vai permitir que a Celera localize as seqüências dentro do genoma e tornará o problema de montagem muito mais fácil. Ainda assim, é um grande desafio. Apesar de estarmos sempre tentando aperfeiçoar os softwares, mesmo assim, há regiões que são extremamente difíceis, há muitas seqüências repetidas e elas são muito similares umas às outras. Não haverá como realizar a finalização de forma completamente automática.

> "Presumivelmente, há outro tipo de informação na seqüência, além dos genes. Outros detalhes biológicos"

■ *Mas o senhor não está cético em relação ao cronograma do projeto Genoma Humano...*

— Também acho um desafio. O que me preocupa é que os prazos pressionam as pessoas a baixar o padrão de qualidade das seqüências. Acho que haverá pressão para que se gere menor quantidade de dados, um número menor de *reads* em cada região, para permitir que se avance mais rapidamente. Se isto acontecer, não haverá dados suficientes para obter a seqüência toda corretamente. Talvez se conseguirmos automatizar a fase de finalização, possamos ajudar. Mas, com menos dados, é provável que haja regiões em que a seqüência não será precisa, pois a pressão vai desestimular as pessoas a buscar em mais dados. Então, o que me preocupa é que o produto, a seqüência final que vai emergir daqui a cinco anos, ou daqui a dois, no projeto da Celera, poderá conter muitos erros. Haverá regiões montadas de maneira errada; haverá outras regiões em que a montagem estará certa, mas com trechos errados na seqüência, o que pode comprometer o trabalho dos biólogos. Não é realista pensar que vamos obter seqüências perfeitas. O objetivo com o qual concordamos no Genoma Humano é de admitir um erro a cada 10 mil pares de bases. Quando os biólogos ouvem esse número, eles sentem que é excessivo. Para mim, não soa excessivo, porque o comprimento de uma região codificadora dentro de um gene é talvez mil ou mil e quinhentos pares de bases, o que garante que apenas uma minoria dos genes contenha erros em sua seqüência. Queremos que pelo menos as regiões codificadoras tenham alto grau de precisão, porque haverá um grande número de estudos biológicos das proteínas criadas por elas, e será necessário também comparar as seqüências codificadoras de proteínas entre diversos organismos, para analisar aspectos relacionados à evolução. Então, a precisão é necessária, quase sempre. Se viermos a admitir um erro a cada mil pares de bases, isto significa que praticamente todo gene terá pelo menos um erro. Uma parcela importante deles vai resultar em conclusões erradas a respeito do aminoácido codificado, o que é muito sério. Assim, um erro em 10 mil bases é razoável. Seria possível aumentar a precisão, mas, então, o custo se tornaria talvez alto demais.

> "O que me preocupa é que os prazos pressionam as pessoas a baixar o padrão de qualidade das seqüências"