GENÉTICA

Descobertas em ritmo acelerado

Genoma Humano do Câncer chega a 45 mil seqüências

Respira-se contentamento nos corredores do Instituto Ludwig, onde trabalham os cinco coordenadores principais do projeto Genoma Humano do Câncer: Andrew Simpson, coordenador de DNA, Emmanuel Dias Neto, coordenador de bibliotecas, Luis Fernando Lima Reis, à frente dos trabalhos de RNA, Sandro José de Souza, em bioinformática, e Juçara de Carvalho Parra, gerente do projeto. O discreto entusiasmo do íntimo e bem azeitado grupo de pesquisadores aparece nos sorrisos com que mostram os gráficos de produtividade por centro de seqüenciamento. É só olhar: desde julho, mês a mês, todos os centros aumentam consistentemente o volume de produção.

Destaca-se, de agosto para setembro, um entre os seis envolvidos no trabalho – é o do professor Marco Antonio Zago, na Faculdade de Medicina de Ribeirão Preto, que, de 500 seqüências em um mês, disparou a produzir 4.200. Destaca-se também um único centro de seqüenciamento – ironicamente, o dos ilustres coorde-

O mecanismo de expressão do gene

O organismo de todo ser vivo é formado por células. É nelas que ocorrem todos os processos bioquímicos que levam à produção de energia e ao crescimento e manutenção do organismo. Nos organismos eucariotos, é o núcleo, uma das estruturas básicas da célula, que governa todas as funções da célula e contém a informação genética, na forma de ácido desoxirribonucléico, DNA. O DNA está contido em estruturas chamadas cromossomos.

Na descrição construída principalmente pela bioquímica das décadas de 70 e 80, a palavra gene designa, materialmente, determinados trechos das moléculas de DNA. O que dá a um trecho da molécula de DNA o status de gene é o fato de a célula encontrar ali as informações de que precisa para criar proteínas – os compostos que animam o movimentado interior da célula. Os cientistas dizem que um gene se expressou quando produziu sua proteína.

Em média, um gene estende-se por 10 mil nucleotídeos – os blocos que formam a molécula de DNA e que diferem entre si pela base nitrogenada, apenas quatro: adenina, timina, citosina e guanina, designadas por A, T, C, G. Os biólogos moleculares entendem que a informação está na ordem em que se apresentam os nucleotídeos do gene, na sua seqüência.

No espaço, a molécula de DNA apresenta-se como uma dupla hélice, duas fitas que se mantêm enroladas uma na outra graças às forças que aproximam e mantêm ligados os As de uma fita com os Ts da outra; e os Cs de uma dos Gs da outra. A regra de formação é conhecida – a cada A liga-se um T, a cada T um A; a cada C um G, a cada G um C. Ao longo de uma cadeia de DNA, os cientistas já sabem relacionar certos conjuntos de bases na seqüência, que se arranjam de determinada maneira, com o processo de transcrição da informação contida no gene.

Esses arranjos são sinais; caminhando sobre a seqüência de bases da esquerda para a direita (sentido 5'-3'), *downstream*, como se diz no jargão, os cientistas sabem, por exemplo, que o gene "começou" quando encontram uma região promotora, *promoter*, reconhecível por apresentar um certo arranjo das bases do DNA. O sinal que ela dá é: "Logo adiante na seqüência vem um gene que deverá expressar sua proteína". Ainda *downstream*, para marcar o fim do gene, há uma região terminadora, *terminator*, caracterizada por outro arranjo de sinais.

Na região promotora liga-se o complexo de proteínas e enzimas envolvidas na transcrição, que inicia e vai regular o processo de cópia do DNA em RNA – o outro tipo de ácido nucléico, passo intermediário entre o gene e a proteína. A diferença de composição entre o RNA (ácido ribonucléico) e o DNA, na composição, é a base uracil, U, que substitui a base timina do DNA. Este RNA é chamado de RNA mensageiro porque carregará a mensagem que a célula vai animar para fabricar a proteína. Há genes que expressam não proteínas como seu produto final, mas outros tipos de ácido ribonucléico, como o RNA transportador ou o RNA ribossômico.

No caso de organismos procariotos – cujas células não têm núcleo –, a transcrição do gene em RNA men-

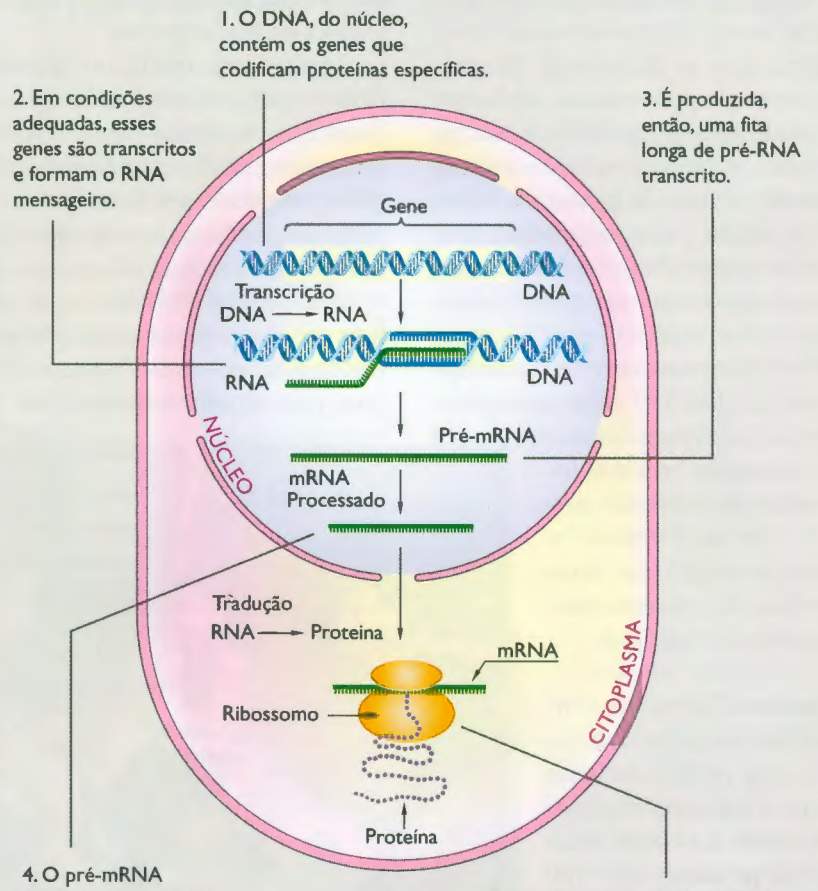
nadores – que diminuiu a produção. A subida de um tem a ver com a queda do outro e não tolda a evidência: os centros de seqüenciamento, em geral, já lidam bem com os novos seqüenciadores capilares Megabace 1000 e obtêm deles bom rendimento.

Também é evidente que quem está puxando o bom ritmo de produção do projeto como um todo é o centro de seqüenciamento do professor Zago. O aperfeiçoamento das rotinas criadas e desenvolvidas pelo paraense Wilson Araújo Júnior, na Faculdade de Medicina de Ribeirão Preto, foi o santo que fez o milagre da multiplicação das seqüências – o crescimento foi de 350% de um mês para o outro.

A performance de Ribeirão Preto mostra que não deverá haver problemas para cumprir o cronograma de

O gene e a sua proteína em organismos eucariotos

SÍRIO CANÇADO



sageiro se dá em um só passo. Não é assim (conforme a descoberta de Phil Sharp em 1977) nos organismos eucariotos – entre os quais nos encontramos. Neles, há duas etapas: num primeiro passo, a célula gera a partir do gene uma molécula "bruta" de RNA , chamada pré-RNA mensageiro. Este pré-RNA mensageiro é a transcrição linear de toda a seqüência do gene; ele se tornará um RNA mensageiro "maduro" depois da segunda etapa de processamento, que se dá ainda dentro do núcleo: a célula remove partes da molécula do pré-RNA – esta parte do processo chama-se *splicing* –, acrescenta a ela determinadas seqüências nas extremidades que lhe conferem maior estabilidade e só então a envia, como RNA maduro, para o citoplasma, onde se dará sua tradução em proteína.

Os segmentos removidos da molécula do RNA mensageiro correspondem a trechos do gene chamados íntrons; os segmentos que restam se chamam exons. Os íntrons são, então, trechos internos ao gene que *não* expressam proteína; exons, os que sim, se expressam como proteína. Essa estrutura de exons e íntrons não existe nos genes da maior parte dos procariotos: toda a seqüência do DNA do gene das bactérias, por exemplo, é regularmente usada para comandar a formação das proteínas. Uma característica do processo nos eucariotos é a possibilidade que a célula tem de combinar diferentemente os exons; o que resulta em diferentes moléculas de RNA mensageiro maduro; e, por conseqüência, em diferentes proteínas traduzidas no citoplasma.

Este RNA mensageiro é a matéria-prima de projetos como o Genoma Câncer, projetos de seqüenciamento de *Expressed Sequence Tags*: obtém-se dele uma molécula de DNA que só contém, do gene como um todo, os exons escolhidos pela célula. Esse tipo de molécula de DNA, sintetizada nos laboratórios, chama-se *complementary* DNA – cDNA. Os ESTs, então, são segmentos pequenos de cDNA – variam de 200 pares de bases a 700 pares de bases.

A EST é uma etiqueta (*tag*), uma espécie de bandeirinha, que marca um lugar do cromossomo que contém as informações para a síntese de proteína – isto é, que é gene. No entanto, a técnica tem limitações importantes. Entre elas: da maneira como o cDNA é obtido hoje na maior parte dos laboratórios, resulta uma concentração muito grande de amostras das extremidades dos genes que se expressam em maior quantidade. Nos bancos de dados de todo o mundo, há poucas ESTs que marquem o centro de genes pouco expressos. As Orestes da parceria FAPESP/Ludwig vêm preencher esta lacuna.

produção de seqüências. Já a análise dos dados que estão sendo depositados todos os dias, e sua comparação com os bancos internacionais de DNA, mostra que a tecnologia Orestes cumpre o que prometia ser, no lançamento do projeto, no final de março. Também as últimas notícias importantes do mundo da genômica reforçam o acerto e a oportunidade dos projetos que seqüenciam não diretamente o gene, mas sua transcrição a partir de uma molécula de RNA mensageiro – como ocorre nesta primeira parceria da FAPESP com um instituto privado de financiamento de pesquisa. Será esta informação já elaborada pela célula que vai iluminar os dados da seqüência completa dos 23 cromossomos humanos (*ver quadro*).

MIGUEL BOYAYAN

Souza, Juçara, Dias Neto, Simpson e Reis: novas metodologias

As notícias: É Andrew Simpson quem descreve o cenário. Há muitos indícios de que o número de genes do genoma humano esteja mesmo próximo dos 140 mil, como a Incyte – uma das companhias privadas que dão contribuição significativa nesta área da pesquisa de ponta – afirmou na badaladíssima 11ª Conferência de Seqüenciamento e Análise de Genomas, promovida pela Tigr, em Miami, no final de setembro.

O número não era mais surpresa, e corrobora os achados de outro importante parceiro privado – a Celera, parceria de Craig Venter e da Perkin Elmer, fabricante de equipamentos. O cientista – que anunciou o término do seqüenciamento do genoma da *Drosofila melanogaster* na reunião de Miami – já percebera que o tamanho da seqüência da mosca seria 40% maior do que previsto. Imagina-se o mesmo do tamanho do genoma humano.

Então, no que toca ao DNA do *Homo sapiens*, há mais genes do que o previsto (os cientistas já sabem da existência de cerca de 85 mil genes; mas nem 10% estão completamente seqüenciados) e devem ser localizados numa cadeia maior – talvez 4 bilhões de nucleotídeos.

Mesmo com o próximo lançamento da seqüência completa do genoma humano – o esboço dos 23 cromossomos, prometido até o final de junho pelos pesquisadores financiados com dinheiro público; ou o genoma finalizado, que a Celera afirma que termina até o final de 2000 –, qual a melhor estratégia para localizar os genes nas cadeias dos cromossomos e, depois, para estudar o que a célula busca de informação neles? A estratégia de ESTs – *Expressed Sequence Tags*, das quais a tecnologia Orestes, criada principalmente por Emmanuel Dias Neto e por Andrew Simpson – é um caso particular e um desenvolvimento.

Razões: Randy Scott, cientista e presidente da Incyte, palestrante na conferência, comparou a competição em torno da seqüência completa do DNA humano, e em torno da primazia na obtenção das informações relevantes, não a uma corrida de 100 metros rasos, mas a uma maratona – o que não é uma novidade para especialistas. A observação aponta para o fato de que o conhecimento da seqüência ordenada de todas as bases das moléculas de DNA presente no núcleo de cada uma dos trilhões de células do corpo humano – o produto ao qual os biólogos moleculares terão acesso até o final de 2000 – é o começo e não o fim da investigação sobre o patrimônio genético humano.

Nesta altura do desenvolvimento da genética molecular, não há ferramentas de bioinformática tão sensíveis e precisas a ponto de detectar sem erro todos os genes, que são os trechos relevantes da cadeia – aqueles que têm a informação que permite a construção das proteínas. Além da questão topológica, há outra: o *splicing* alternativo.

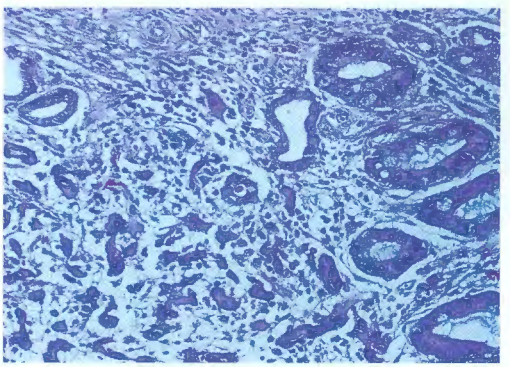
Os cientistas sabem desde a descoberta de Phillip Sharp, em 1977 (*ver Notícias FAPESP Nº 46*), que os organismos eucariotos cujas células têm núcleo usam a informação bruta contida na seqüência de DNA dos genes de muitas maneiras diferentes – ao contrário do que acontece entre as bactérias.

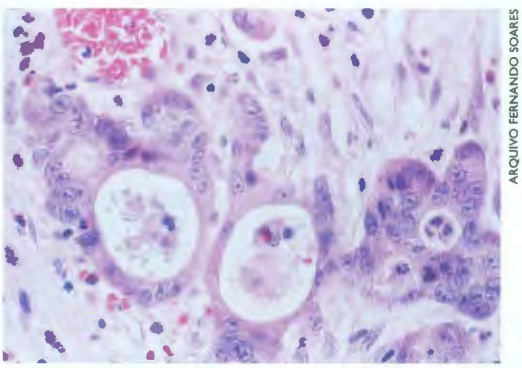
No caso da nossa espécie, estimativas apresentadas na mesma conferência em Miami indicam a possibilidade de cada gene humano "se apresentar", em média, de 50 maneiras diferentes – quer dizer, ser lido de 20, ou 60, ou 100 combinações de seus

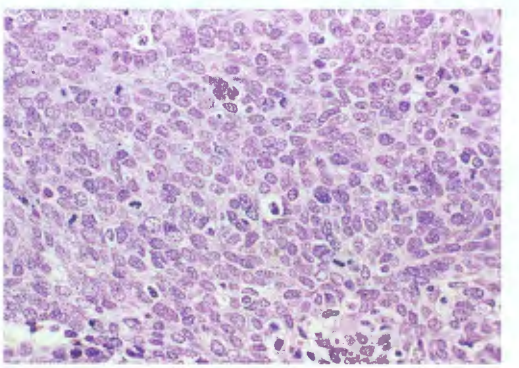
trechos internos, segundo momentos e necessidades diferentes na vida de uma célula.

Mesmo que essa nova estimativa da freqüência de combinações diferentes seja alta demais, as diferenças de combinação continuarão importantes. Esses trechos internos ao gene dos organismos superiores, e que se emendam de maneira alternativa para codificar esta ou aquela proteína, chamam-se exons; de novo, não há ferramentas refinadas a ponto de localizá-los sem erros significativos diretamente da seqüência; novamente nos bancos de dados, são informações novas, que aparecem entre as seqüências brasileiras por causa das peculiaridades da metodologia usada aqui, e que tornam os resultados únicos no mundo da genômica.

À medida que a produção aumenta, aumenta também a possibilidade de descoberta de novos genes, relacionados aos tumores dos cânceres mais comuns no Brasil. Simpson gostaria de entrar no ano 2000 com 100 mil seqüências produzidas; se, em um ano, atingir o objetivo de 500 mil, o projeto terá contribuído







ARQUIVO FERNANDO SOARES

Material para exame microscópico de tipos variados de câncer: de estômago (*acima*), intestino (*à dir.*) e colo de útero (*ao lado*)

te, é na estratégia de ESTs que os pesquisadores depositam as esperanças de fazê-lo com rapidez e precisão. Os cientistas esperam encontrar na diversidade de combinações mais explicações sobre a biologia do câncer, por exemplo. Daí o interesse de Phillip Sharp – Nobel de 1993 – em participar do corpo de consultores do Genoma Câncer paulista.

O andar da carruagem: Até dia o 24 de outubro, os centros já haviam depositado 35 mil seqüências geradas com a metodologia patenteada pela FAPESP e pelo Instituto Ludwig. Somadas a elas as 10 mil seqüências produzidas durante o projeto piloto, são 45 mil – quase 10% das 500 mil contratadas, que deverão ser produzidas até o fim do projeto, previsto para 2001. Vinte e sete por cento delas não têm nenhum correspondente com 20% de todas as ESTs geradas internacionalmente. Com uma diferença: além de pelo menos um quarto delas ser completa novidade, a maior parte do que será seqüenciado aqui virá do centro dos genes, que só a metodologia Orestes alcança. O processo tradicional de obtenção de ESTs enfrenta a limitação técnica de serem, quase sempre, seqüências das extremidades dos genes, menos relevantes para a codificação de proteínas.

O protocolo: Animado com o salto na produção, Simpson já fala em dobrar o objetivo de produção de seqüências do projeto. Se as rotinas e os métodos desenvolvidos e consolidados em Ribeirão Preto puderem ser adaptados e estendidos aos outros centros de seqüenciamento, a velocidade de cruzeiro poderá ser atingida antes de maio do ano que vem, para quando está prevista.

O seqüenciador capilar do centro funciona de dez a doze horas por dia; são cinco "corridas" diárias, de segunda a sexta, que fornecem a leitura de qualidade, em média, de mais de 90 das 96 amostras processadas de cada vez. O protocolo testado e aperfeiçoado por eles abrevia etapas e gasta menos reagentes – um item importante no custeio das plantas genômicas.

Sete laboratórios abastecem o centro da Faculdade de Medicina de Ribeirão Preto – dois a mais do que o padrão dos centros de seqüenciamento. Estes dois laboratórios desprenderam-se do Ludwig e integram agora o grupo de Ribeirão. A mudança é uma das razões que explicam a queda, em setembro, na produção do instituto; e não a única que explica a disparada de Ribeirão. Há espaço para a produção crescer bastante, em todos os centros – daí o discreto otimismo reinante.

A continuar tudo indo bem, Simpson poderá botar em andamento um de seus planos: pedir ao coordenador de bioinformática, Sandro de Souza, e a seu grupo que analisem as seqüências do genoma humano disponibilizadas pelos cientistas do Hemisfério Norte à luz das informações vindas das Orestes.

"Você vai ver: vamos achar 50% de erro na anotação dos genes feitas por eles", prevê o coordenador, que acredita numa contribuição significativa ao conhecimento na área a ser feita pelo projeto que comanda. Para ele, não há outra metodologia capaz de fornecer dados complementares tão importantes a toda a informação que está sendo gerada no mundo. •