

Mais bits a serviço do DNA

Bioinformatas brasileiros criam ferramentas para estudar genomas

Marcos Pivetta

Há pouco mais de uma década quase não havia genomas completos para serem analisados. Hoje faltam programas e mão de obra especializada para dar conta da quantidade de seqüências de DNA já depositadas em bases públicas de dados e que saem diariamente de uma nova geração de sequenciadores. Extremamente velozes, essas máquinas determinam os pares de bases do material genético, as chamadas letras químicas, a um preço milhares de vezes menor do que no início dos anos 2000, quando chegou ao fim a epopeia de sequenciar o primeiro genoma humano. De olho nesse desafio, o matemático João Meidanis, sócio-fundador da empresa Scylla Bioinformática e professor da Universidade Estadual de Campinas (Unicamp), investiu numa linha de pesquisa: criar métodos mais simples e eficientes para comparar dois ou mais genomas.

Ao lado de Pedro Feijão, ex-aluno seu de doutorado, formulou em 2009 a base teórica de uma técnica para comparar genomas inteiros denominada *single-cut-or-join* (SCJ) e, no ano passado, testou-a na prática em genomas de alguns organismos, como plantas e bactérias. “Com nosso método, podemos comparar facilmente dois ou mais genomas sem aumentarmos exponencialmente o número de cálculos, como ocorre com outras técnicas”, afirma Meidanis. “Assim podemos construir árvores genealógicas e ver quais são os genomas mais próximos e os mais distantes do ponto de vista evolutivo.” O matemático foi um dos coordenadores de bioinformática do projeto que sequenciou, no ano 2000, o genoma da bactéria da *Xylella fastidiosa*, causadora da doença do amarelinho nos laranjais. O trabalho rendeu a primeira capa da revista científica *Nature* a uma pesquisa brasileira.

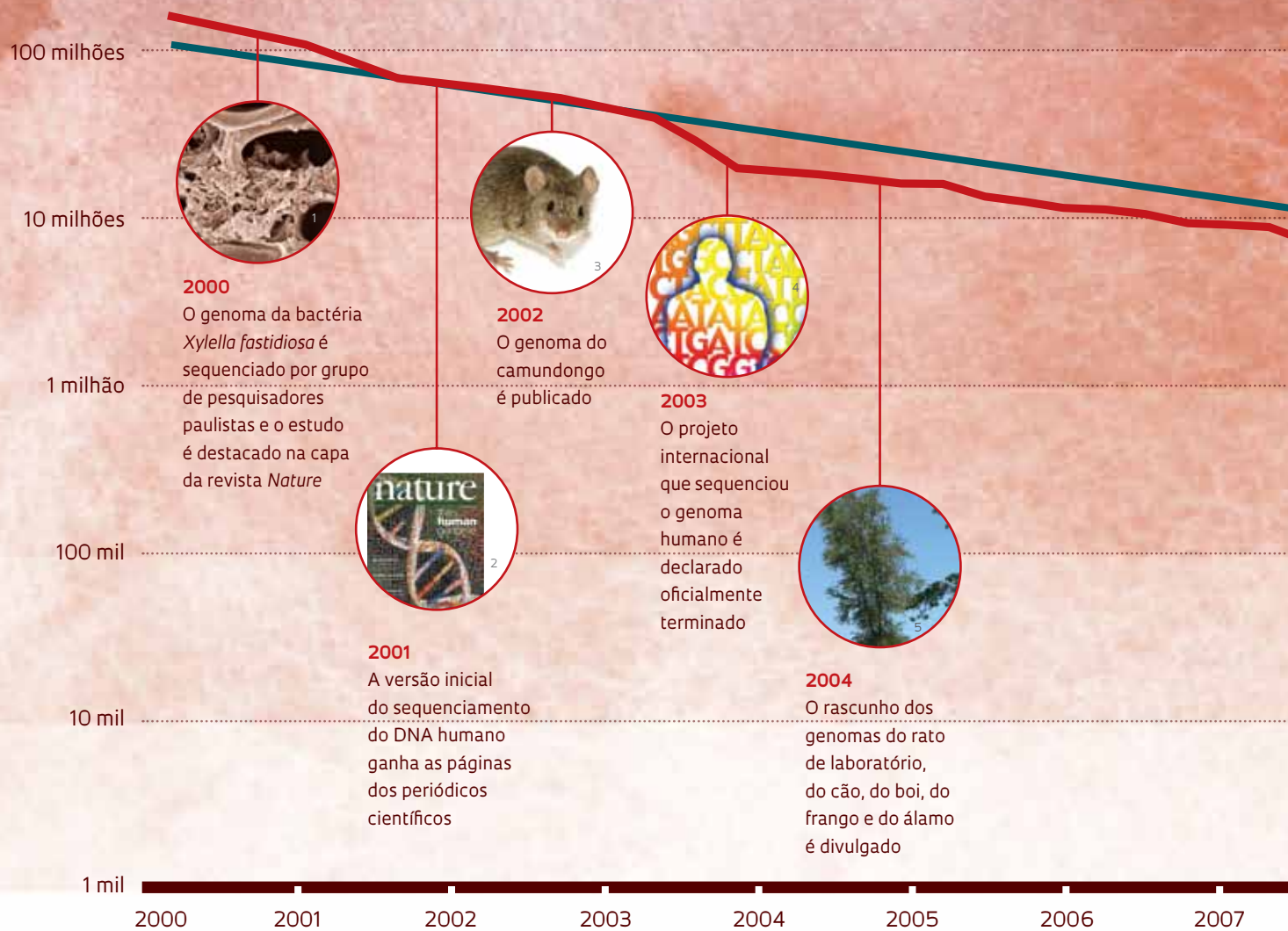


0 39 233 09660 273 650 96

O tombo do DNA

Em uma década, o custo de sequenciar o genoma humano caiu de US\$ 100 milhões para menos de US\$ 10 mil

* eixo y em escala logarítmica (em US\$)



Para confrontar todo o material genético de uma espécie com o de outra, os pesquisadores têm de recorrer a simplificações. A principal delas é considerar que os genes presentes nos genomas comparados são exatamente os mesmos, embora estejam ordenados de forma diferente na sequência específica de cada organismo. Partindo desse raciocínio, os métodos para comparar genomas contabilizam o número de rearranjos que teriam sido necessários para que um genoma se transformasse no outro. Esses rearranjos decorrem da movimentação de grandes segmentos de DNA ocorrida ao longo do tempo na sequência original. Dessa forma, quanto menor for o número de rearranjos que separam dois genomas, mais próximos eles se encontram na árvore evolutiva.

Em seu método, Meidanis e Feijão formularam uma definição alternativa para o conceito de ponto de quebra (*breakpoint*), parâmetro importante para encontrar rearranjos numa sequência e, assim, calcular a proximidade de dois genomas. Ponto de quebra é o local em que há uma interrupção num longo segmento conservado nos genomas que estão sendo comparados.

No ano passado, a dupla ainda refinou outro método de comparação de genomas, mais elaborado que o SCJ. Inicialmente proposta no ano 2000, essa segunda técnica permitia confrontar apenas genomas circulares. Com o aperfeiçoamento, passou também a ser útil para comparar o material genético de cromossomos lineares. “Essa era uma das limitações da técnica origi-

LEI DE MOORE

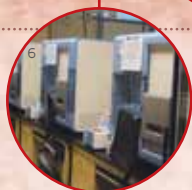
Até meados dos anos 2000, a queda no custo de sequenciar seguiu o barateamento tecnológico decorrente da Lei de Moore. Depois o preço despencou



2011 Cientistas chineses do BGI anunciam projeto para sequenciar 3 milhões de genomas de plantas, animais, seres humanos e microrganismos



2012 O genoma completo de 1.092 pessoas, de diferentes continentes, é publicado na revista *Nature*



2008

As máquinas de segunda ou última geração chegam ao mercado e derrubam o custo de sequenciar genomas

FONTE: NHGRI GENOME SEQUENCING PROGRAM

2008 2009 2010 2011 2012

nal”, comenta Feijão, hoje funcionário da Scylla. O novo método, baseado no que os matemáticos denominam formalismo algébrico por adjacências, ainda não foi testado em genomas reais. Existe, por ora, na forma de teoria.

METAGENÔMICA

Meidanis, logicamente, não é o único a sentir os efeitos da nova realidade em seu campo de atuação. De volta ao Brasil desde meados de 2011, depois de oito anos no Virginia Bioinformatics Institute, nos Estados Unidos, João Carlos Setubal, hoje professor titular do Instituto de Química da Universidade de São Paulo (USP), nota que a demanda por serviços e pesquisa em sua área cresceu e se sofisticou nos últimos tempos. Desde

que regressou a São Paulo, já recebeu, por exemplo, 16 propostas para colaborar com iniciativas de outros pesquisadores. “Os sequenciadores de última geração geram uma quantidade astronômica de dados genômicos, de proteômica e do metabolismo dos organismos”, afirma Setubal, que também foi um dos coordenadores de bioinformática do projeto da *Xylella*. “Com o barateamento da tecnologia, hoje qualquer projeto de pesquisa com um mínimo de recursos pode sequenciar o genoma de um organismo.”

Na última década, um campo que se abriu para o pessoal da biologia e da bionformática foi o da metagenômica, em que se estuda a microbiota de um nicho ecológico. O principal projeto de Setubal, um temático da FAPESP sobre microrganismos presentes no Zoológico de São Paulo, é nessa área. Por essa abordagem, em vez de isolar e cultivar os microrganismos a fim de se extrair em separado o DNA de cada espécie, o pesquisador retira uma amostra diretamente do ambiente a ser estudado. Nessa amostra, o DNA de várias espécies se apresenta “misturado” e cabe ao bioinformata encontrar técnicas para separar e caracterizar o material genético de cada uma delas. “Estamos estudando três microbiomas no Zoológico: a compostagem feita pelos funcionários do parque, a água dos lagos e as fezes dos macacos bugios”, afirma Setubal.

A metagenômica é também uma forma de garimpar organismos desconhecidos num hábitat específico. A equipe de Ana Tereza Ribeiro de Vasconcelos, coordenadora do centro de bioinformática do Laboratório Nacional de Computação Científica (LNCC), em Petrópolis, participou da descoberta de bactérias magnéticas encontradas na lagoa de Araruama, no litoral do Rio de Janeiro, uma das mais salinas do mundo. Uma dessas bactérias encontradas foi a *Candidatus magnetoglobus multicellularis*. Identificada por Ulysses Lins, da Universidade Federal do Rio de Janeiro (UFRJ), a bactéria é difícil de ser isolada do ambiente e mantida em meio de cultura. “Hoje estamos envolvidos em uma dezena de projetos de metagenômica”, diz Ana Tereza, que conta com três sequenciadores em seu laboratório e um time de aproximadamente 25 pessoas.

A escala de tempo e dinheiro envolvidos nos projetos dedicados a analisar o DNA de organismos mudou radicalmente na última década. Nos primeiros anos da era genômica, apenas grandes empreitadas ousavam se aventurar nessa nova seara. Em abril de 2003, quando o consórcio público internacional que sequenciou pela primeira vez um genoma humano chegou oficialmente ao fim, a megainiciativa tinha consumido 13 anos de trabalho de centenas de cientistas de ao menos 18 países (Brasil inclusive) e estimados US\$ 2,7

bilhões. Numa proporção muito menor, mas ainda assim grandiosa, o sequenciamento da *Xylella* custara US\$ 12 milhões à FAPESP e envolveu a contribuição de 192 pesquisadores por três anos.

Agora sequenciar genomas se tornou uma tarefa entre 10 mil e 20 mil vezes mais em conta do que há pouco mais de uma década, segundo dados do National Human Genome Research Institute (NHGRI) dos Estados Unidos. A chegada em massa ao mercado no início de 2008 dos sequenciadores de segunda geração, que empregam uma tecnologia diferente das primeiras máquinas do tipo Sanger, tem feito o preço do sequenciamento despencar num ritmo acelerado que supera de longe os ganhos de desempenho decorrentes da Lei de Moore na informática. Hoje, em dois ou três dias, a um custo de uns poucos milhares de dólares, é possível determinar todos os 3 bilhões de letras químicas do DNA de uma pessoa. “A bioinformática é uma nova ferramenta, uma lupa, para entendermos melhor o fenômeno biológico, que não mudou, mas agora pode ser visto de outra forma”, diz Gonçalo Pereira, do Instituto de Biologia (IB) da Unicamp.

Mas sequenciar é uma coisa, extrair informação útil dos bilhões de dados que os computadores despejam cotidianamente nas mãos dos cientistas é outra, bem mais complexa. “O sequenciamento em si hoje é barato, virou uma *commodity*,

mas a análise dos dados é cara”, diz o cientista da computação João Paulo Kitajima, da Mendelics, empresa recém-criada que trabalha com diagnósticos genômicos personalizados. “A procura pelo trabalho de bioinformática cresceu exponencialmente e há um *gap* entre a demanda e a oferta de especialistas no Brasil e no exterior.”

É difícil estimar com precisão o tamanho da comunidade de bioinformatas no país. Segundo Guilherme Oliveira, presidente da Associação Brasileira de Bioinformática e Biologia Computacional (AB3C), cerca de 300 pessoas, entre professores, alunos e pesquisadores, mantêm vínculos com a entidade. “Antigamente o bioinformata era um autodidata”, diz Oliveira, coordenador do centro de bioinformática da Fiocruz de Minas Gerais. “Hoje boa parte sai dos cursos de pós-graduação e em cada estado há algum especialista no setor. Uma novidade é que agora empresas também atuam na área.” Grandes universidades do país, como a USP, a UFRJ e a Universidade Federal de Minas Gerais (UFMG), e a Fiocruz mantêm programas de pós-graduação específicos sobre bioinformática. Outras universi-

Barateamento do sequenciamento tornou as técnicas genômicas acessíveis a projetos de todos os orçamentos

China tem maior centro de sequenciamento

Em menos de 15 anos, um centro chinês de bioinformática saiu da condição de sócio menor do consórcio internacional que mapeou o primeiro genoma humano para se tornar a maior potência global em termos de sequenciamento de DNA. Fundado em 1999, o Beijing Genomics Institute (BGI, na sigla em inglês) é dono hoje de 180 máquinas sequenciadoras, a maior parte delas de última geração, que podem produzir diariamente 6 terabytes de dados, equivalentes aos genomas completos de 2 mil indivíduos. Tem 4 mil funcionários e filiais nos Estados Unidos, Europa e Japão. As ações dos chineses, que trabalham numa escala gigantesca, criam a expectativa de que o preço de sequenciar um genoma humano caia em breve para US\$ 1.000 e os colocam na condição de protagonistas em projetos de ponta, que vão muito além

da determinação da sequência genética de um símbolo nacional, o panda-gigante, feito alcançado há três anos.

Em 2010, por exemplo, o BGI sequenciou o primeiro genoma completo de um ancestral do ser humano, o DNA de um esquimó que viveu há 4 mil anos. Em 2012 forneceu o DNA de 100 chineses para um esforço internacional que estuda o genoma de cerca de mil indivíduos de diferentes regiões do planeta. Ainda no ano passado o centro se propôs a sequenciar nos próximos anos 3 milhões de genomas, de seres humanos, plantas, animais e microrganismos.

A política dos chineses é agressiva em todos os sentidos, não só no campo científico, mas também do ponto de vista comercial. Além de vender seus préstimos em bioinformática, o BGI tenta garantir para si o acesso aos mais recentes avanços do setor. No início

deste ano, o centro asiático recebeu o sinal verde dos americanos para comprar por US\$ 177 milhões uma companhia da Califórnia, a Complete Genomics, que desenvolveu uma nova tecnologia de sequenciamento, cujos resultados seriam mais precisos do que os obtidos com os métodos atualmente em uso em todo o mundo.



Panda-gigante: um dos genomas que os chineses do Beijing Genomics Institute (BGI) sequenciaram por completo



Águas hipersalinas da lagoa de Araruama, no Rio de Janeiro: estudos de metagenômica encontraram bactérias magnéticas no local

dades trabalham o tema como uma linha de pesquisa no âmbito da pós de uma área mais ampla, como biologia ou computação.

Os trabalhos de sequenciamento e análise do genoma do *Schistosoma mansoni*, parasita causador da esquistossomose, são o projeto de maior visibilidade a que o centro de bionformática da unidade mineira da Fiocruz tem se dedicado nos últimos anos. Mas as seis máquinas sequenciadoras e os 15 especialistas em bionformática do setor comandado por Oliveira participaram de cerca de 60 projetos distintos, que incluem estudos sobre o genoma do câncer, de agentes infecciosos, de raças de bovinos e trabalhos de metagenômica. Hoje o centro ainda gera e analisa dados para a Rede de Pesquisa de Identificação Molecular da Biodiversidade Brasileira (BR-BoL), coordenada por Cláudio Oliveira, do Instituto de Biociências da Universidade Estadual Paulista (Unesp) de Botucatu, que deverá catalogar 120 mil exemplares de 24 mil espécies da natureza em quatro anos. A BR-BoL é o braço brasileiro de um projeto internacional, o Barcode of Life (Código de Barras da Vida), que visa identificar espécies por meio da caracterização de seu DNA.

A bioinformática se espalhou pelo país e atingiu centros longínquos das grandes capitais do Sudeste. Na Universidade Federal do Pará (UFPA), Artur Silva faz pesquisas no setor e colabora com grupos de São Paulo. Desde maio do ano passado, Sandro de Souza, que por anos chefiou essa área no Instituto Ludwig de Pesquisa sobre o Câncer de São Paulo, está no Instituto do Cérebro da Universidade Federal do Rio Grande do Norte (UFRN). Ele não tem um sequenciador a seu lado na capital potiguar. Nem por isso se mostra preocupado. “Dá para fazer sequenciamento até na nuvem da internet se a gente quiser”, afirma Souza. “Estou começando

meus trabalhos em neurociências sem problemas.”

Até porque Souza ainda tem acesso a todas as máquinas do Ludwig, que, com o fechamento do instituto na capital paulista, foram transferidas para a Faculdade de Medicina de Ribeirão Preto (FMRP) da USP, onde começou a funcionar no ano passado o Centro de Medicina Genômica. “As técnicas de genômica e bionformática farão uma revolução na prática médica assim como fez a medicina baseada em imagens”, diz Wilson Araújo da Silva Júnior, um dos responsáveis pelo novo centro da FMRP.

Para tornar mais acessíveis os serviços de sequenciamento e análise de DNA e RNA, a Unicamp inaugura em 1º de março o Laboratório Central de Tecnologias de Alto Desempenho (LacTAD), que atuará nas áreas de genômica, proteômica, biologia celular e bioinformática propriamente dita. Entre os equipamentos do laboratório há dois sequenciadores de nova geração da empresa Illumina, daqueles que em poucos dias é capaz de sequenciar por completo um DNA humano, e um terceiro para sequenciar regiões específicas de genomas. A rigor, as máquinas do centro já estão em uso desde o ano passado, quando chegaram à universidade, mas espalhadas em distintas unidades. A partir do próximo mês elas passam a operar no prédio de 2.000 metros quadrados construído para o LacTAD.

“Acreditamos que haja uma demanda reprimida por esse tipo de serviço e a bionformática se tornou um gargalo de muitas pesquisas nas áreas biológicas”, afirma o químico Ronaldo Pilli, pró-reitor de Pesquisa da Unicamp, que está à frente do projeto do novo laboratório. “Acompanhamos a tendência mundial de oferecer esse tipo de serviço de forma centralizada. Assim é mais fácil comprar, operar e manter atualizados os equipamentos.” Os equipamentos do LacTAD custaram cerca de R\$ 5,5 milhões e foram adquiridos por meio do programa Multiusuários da FAPESP. Orçado em R\$ 4 milhões, o prédio foi bancado pela universidade.

O LacTAD prestará serviços a pesquisadores da Unicamp e também de outras universidades e empresas. No site do laboratório há um formulário para pesquisadores interessados em cotar os preços dos seus serviços. “Faremos trabalhos que podem custar de R\$ 100 até R\$ 100 mil”, comenta Pilli. É a democratização da bioinformática. ■

Projetos

1. Estudos da diversidade microbiana no Parque Zoológico de São Paulo – nº 11/50870-6; 2. EMU: Laboratório Central de Tecnologias de Alto Desempenho – nº 09/54129-9. **Modalidades:** 1. Programa Biota – Projeto Temático; 2. Programa de Equipamentos Multiusuários. **Coordenadores:** 1. João Setubal – USP; 2. Fernando Ferreira Costa – Unicamp. **Investimento:** 1. R\$ 1.711.698,25 (FAPESP); 2. R\$ 6.034.431,00 (FAPESP).