Francisco Prosdocimi

Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais

Fabrício R. Santos

Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais

Sobre bioinformática, genoma e ciência

As notícias sobre o seqüenciamento do genoma do homem e de outras espécies, e sobre a identificação de genes e de suas funções, tornaram-se freqüentes nos últimos anos. As informações genômicas acumulam-se com tal rapidez que, para sua análise e detalhamento, são necessários processos cada vez mais poderosos e criativos. Essa é a tarefa da bioinformática, ciência que trouxe uma abordagem científica aos dados basicamente descritivos obtidos pelas máquinas seqüenciadoras.

A bioinformática consiste principalmente na análise computacional de seqüências de DNA, RNA e proteínas. Essa nova ciência surgiu na última década em função da necessidade de ferramentas sofisticadas para analisar o crescente volume de dados gerado em biologia molecular. O GenBank, criado no centro norte-americano para informação biotecnológica (NCBI, na sigla em inglês), foi um dos primeiros e ainda é o mais popular banco de dados para o depósito de seqüências de DNA. É lá que pesquisadores de todo o mundo depositam as seqüências de As, Cs, Gs e Ts (iniciais das 'peças' básicas da molécula de DNA, os nucleotídeos adenina, citosina, guanina e timina) que obtêm ao seqüenciar o genoma dos mais diversos organismos.

No final dos anos 90 observou-se um crescimento exponencial do número de seqüências de biomoléculas depositadas no GenBank (figura 1). Esse aumento teve início a partir de 1990, quando surgiram os seqüenciadores de DNA a *laser*, totalmente automatizados. Tais máquinas têm com freqüência 96 capilares (tubos minúsculos por onde passam os fragmentos de DNA) e podem 'ler', em média, 550 letras (A, C, G e T) por capilar em cada análise. Há cerca de 3 bilhões dessas letras no genoma humano. Seqüenciadores ainda mais potentes, com 384 capilares, podem 'ler' mais de um milhão de letras do DNA por dia!

Existem no Brasil dezenas de seqüenciadores, grande parte deles distribuída entre laboratórios de todo o país quando do início do Projeto Genoma da Fundação de Amparo à Pesquisa do Estado de São Paulo, que seqüenciou o DNA da bactéria *Xylella fastidiosa*, praga da laranja (http://aeg.lbi.ic.unicamp.br/xf/), e o Projeto Genoma Brasileiro (http://www.brgene.lncc.br), que já permitiu o seqüenciamento dos DNAs das bactérias *Chromobacterium violaceum* e *Mycoplasma synoviae*.

A grande maioria das seqüências publicadas em bancos de dados internacionais vem de pro-

jetos genoma e transcriptoma (ou de genoma funcional). O primeiro genoma seqüenciado foi o da bactéria *Haemophilus influenzae*, em meados de 1995. Hoje, o NCBI já contém 1.628 genomas de vírus, 174 genomas de procariotos (bactérias e arqueobactérias) e 20 genomas de organismos eucarióticos. Essa imensa quantidade de informação vem se tornando cada vez mais complexa com o estudo das interações entre biomoléculas e das variações existentes entre os indivíduos de uma população (figura 2). Mas, afinal, que informações cientificamente relevantes o genoma trouxe para os pesquisadores, para as pessoas e para a sociedade? Será que projetos genoma são pesquisas meramente descritivas? Então, qual a relevância da genômica e qual o papel da bioinformática na consolidação dessa ciência?

À primeira vista os estudos de genoma não parecem ser pesquisas científicas clássicas, pois não se baseiam em hipóteses elaboradas *a priori* sobre a biologia de um dado organismo. No máximo, a pergunta que se poderia fazer antes de seqüenciar um genoma seria "esse organismo tem algum gene de potencial biotecnológico?" ou "o que há nos genes desse organismo que o faz conseguir viver nessa condição, ou que gera uma doença?". Tais perguntas, porém, dificilmente serão respondidas diretamente pelo seqüenciamento, e certamente exigirão estudos posteriores. E mais: é possível que alguma investigação não-genômica mais minuciosa sobre esse ou aquele aspecto em particular possa esclarecer de modo mais direto essas questões.

Mas isso não tira o mérito dos estudos genômicos. Acreditamos que a ciência vive hoje a era da anatomia molecular. No século 19, quando pouco se conhecia – de forma sistemática – do mundo biológico, os naturalistas, que exploravam o mundo em busca de informação taxonômica, encontrando e classificando animais e plantas então desconhecidos, podiam ser considerados os cientistas da biologia. A descrição e a documentação de novas espécies era especialmente necessária naquela época, já que pouco ou nada se conhecia sobre a grande diversidade da vida na Terra. O mesmo ocorreu quando surgiram os anatomistas, que pela primeira vez documentaram em detalhe o interior do corpo do homem e de outros animais. Eles apenas descreviam, da melhor maneira possível à época, a localização dos órgãos e tecidos.

Portanto, se a genômica não pode ser considerada, classicamente, uma ciência, também não podem ser vistas assim a taxonomia e a anatomia, já que todas elas são empreendimentos principalmente descritivos, e não investigativos. Mas isso, mais uma vez, não lhes tira o mérito. Quanto conhecimento científico foi construído com base nas informações geradas por naturalistas e anatomistas? A ciência biomédica foi montada a partir do trabalho dos anatomistas, e a teoria mais importante e unificadora de toda a biologia – a evolução – surgiu diretamente das observações e estudos descritivos dos naturalistas Charles Darwin (1809-1882) e Alfred Wallace (1823-1913).

E a genômica? O genoma, na verdade, pode ser descrito como a 'anatomia molecular' de uma espécie. E só agora, neste início de século 21, somos capazes de desvendar e descrever como as espécies são constituídas em seu nível mais básico, o da informação molecular. A genômica, portanto, é a 'ciência descritiva' atual. E assim como as ciências biomédicas trouxeram o método científico ao estudo da anatomia, a bioinformática veio trazer cientificidade aos dados genômicos.

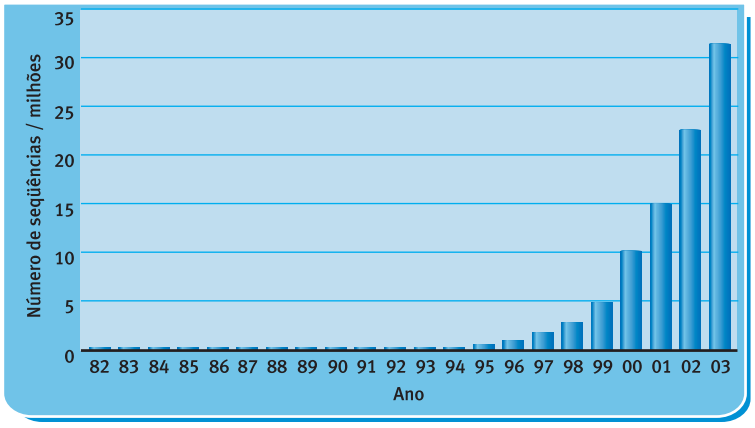

Figura 1. Crescimento do número de seqüências de biomoléculas depositadas no GenBank, o mais popular banco de dados de DNA

Se a genômica não pode ser considerada, classicamente, uma ciência, também não podem ser vistas assim a taxonomia e a anatomia, já que todas elas são empreendimentos principalmente descritivos, e não investigativos

A bioinformática traz uma abordagem científica aos dados gerados em projetos genoma, como já fazem outras ciências bem estabelecidas, como a biologia molecular, a genética e a bioquímica

É importante definirmos bem o que é a bioinformática e em que contexto esse conceito é usado neste ensaio. Muitos crêem que essa ciência consista em qualquer análise computacional de problemas biológicos, mas isso não está de acordo com sua origem. A bioinformática clássica surgiu com o seqüenciamento de biomoléculas e destas permanece inseparável. É possível propor uma definição razoavelmente clara: a bioinformática consiste em 'todo tipo de estudo ou de ferramenta computacional que se pode realizar e/ou produzir de forma a organizar ou obter informação biológica a partir de seqüências de biomoléculas'. Se o estudo envolve seqüências de biomoléculas (DNA, RNA ou proteínas), direta ou indiretamente, trata-se de bioinformática. Se não, trata-se de computação aplicada à biologia, que é extremamente importante em várias áreas e já existia bem antes do início dos seqüenciamentos de biomoléculas.

Definido o conceito de bioinformática que utilizamos, podemos enquadrar muitos estudos nessa área em três princípios paradigmáticos, aos quais daremos os nomes metafóricos de 'tijolo', 'peneira' e 'lupa'.

Estudos de bioinformática 'tijolo' são os relacionados à execução de projetos genoma e normalmente produzem processos para analisar seqüências e interpretar genomas. Algumas dessas ferramentas já são clássicas. Podemos citar o *base calling*, onde as bases do DNA são lidas no seqüenciador a partir dos cromatogramas (perfis de emissão fluorescente que variam entre os nucleotídeos A, C, G e T). Cada cromatograma é transformado em uma seqüência, e um índice de confiabilidade é associado a cada letra do DNA. Em seguida analisam-se as seqüências que têm parte das letras em comum, para eliminar as sobreposições, alinhar os trechos corretos e com isso gerar o 'texto' completo do genoma da espécie estudada (que pode ter milhões ou bilhões de letras). Novas ferramentas para 'conferir' seqüências, alinhá-las (na montagem de um genoma), identificar genes e padronizar processos de *base calling* são alguns exemplos de projetos de bioinformática 'tijolo', sem os quais é impossível a análise sistemática dos 'edifícios genômicos'.

Vale observar que as ferramentas de comparação de seqüências de DNA têm permitido um grande avanço na identificação das funções de genes. Nesse caso, a seqüência de um novo gene é comparada com aquelas armazenadas em um banco de dados de genes de função conhecida, permitindo a rápida dedução da possível função desse gene recém-seqüenciado. Testes experimentais para descobrir a função de cada novo gene descoberto possivelmente exigiriam várias décadas de pesquisa.

A quantidade de informações gerada por um projeto genoma torna virtualmente impossível a análise destas (ou de uma pequena parcela) pelo grupo que gerou essa seqüência completa de DNA. Assim, trabalhos posteriores, envolvendo fragmentos de diferentes genomas, serão necessários para analisar temas específicos (por exemplo, proteínas envolvidas no metabolismo de açúcares). Esses trabalhos de mineração de dados genômicos são característicos da chamada bioinformática 'peneira'.

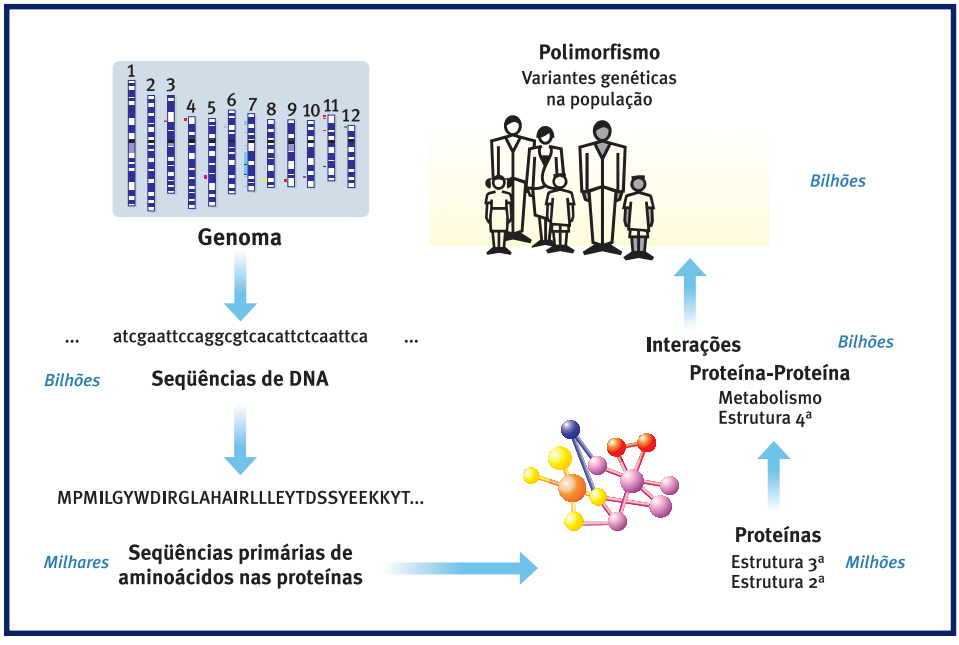

Figura 2. O estudo das interações entre diferentes biomoléculas e das variações genéticas presentes na população, torna mais complexa a imensa quantidade de informação gerada pelos projetos genoma

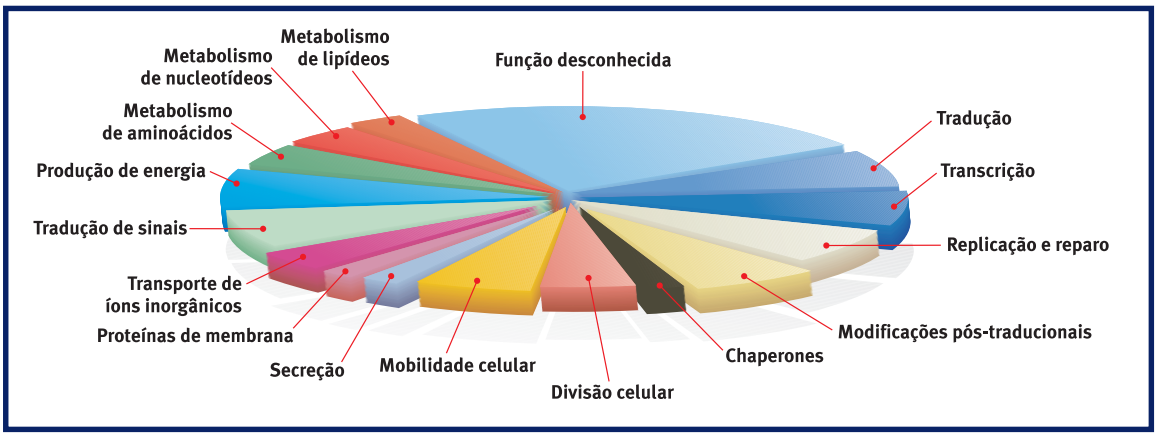

Figura 3. Funções biológicas dos genes já identificadas em estudos genéticos

Como a genômica é em essência uma disciplina descritiva, os trabalhos dessa área exibem muitos dados sem qualquer detalhamento, muitas vezes por limitação do periódico que os publica. A divisão dos genes em grupos, de acordo com sua função biológica (figura 3), é um exemplo da informação descritiva freqüentemente presente em artigos de genoma. Que informação relevante se poderia tirar desse monte de dados? Usando 'peneiras' específicas, os cientistas podem gerar conhecimento mais aprofundado sobre aspectos de seu interesse. A construção de bancos de dados de seqüências de genes que tenham uma ou outra função específica ou de estruturas tridimensionais de proteínas, por exemplo, também está incluída no âmbito da bioinformática 'peneira'. Todo ano, a primeira edição da revista britânica *Nucleic Acids Research* traz um resumo dos bancos de dados mais utilizados na área.

Entretanto, é nos trabalhos de bioinformática 'lupa' que a ciência aparece com maior clareza na genômica. Vale ressaltar que os estudos de genoma e bioinformática citados até agora são indispensáveis para o aumento do conhecimento científico sobre os organismos e sua constituição molecular. Nos estudos do tipo 'lupa', porém, o método científico é rigorosamente aplicado. Aqui, empregando as mais variadas ferramentas computacionais, o processo investigativo científico é retomado: observam-se os dados, criam-se hipóteses e realizam-se experimentos *in silico* (dentro do computador) para comprová-las ou refutá-las através de algoritmos (processos de cálculo que permitem solucionar problemas) bioinformáticos.

É interessante verificar que estudos 'lupa' não são necessariamente publicados em revistas especializadas em bioinformática. Isso acontece porque os algoritmos usados nesses estudos são vistos apenas como a metodologia de um trabalho que tenta buscar um resultado biológico mais específico. A bioinformática não é o centro do trabalho, como ocorre nas abordagens 'tijolo' e 'peneira'. Nos trabalhos 'lupa', a hipótese e os resultados são mais importantes que as ferramentas bioinformáticas usadas como meio investigativo. Assim, tais estudos são freqüentemente publicados nas revistas relacionadas com o organismo em que se está estudando o fenômeno ou em revistas específicas de genética, biologia molecular ou bioquímica.

Exemplos de estudos de bioinformática 'lupa' são aqueles nos quais alguma característica biológica de um organismo é explicada a partir da observação de suas seqüências gênicas ou protéicas e da comparação com seqüências similares em organismos proximamente relacionados. Esses estudos de genômica comparativa permitem associar aspectos da biologia dos organismos comparados à presença ou à ausência de determinado gene, grupo de genes ou processos metabólicos.

Assim, a bioinformática traz uma abordagem científica aos dados gerados em projetos genoma, como já fazem outras ciências bem estabelecidas, como a biologia molecular, a genética e a bioquímica. Vale registrar, no Brasil, a iniciativa pioneira da Coordenação para o Aperfeiçoamento de Pessoal de Nível Superior (Capes) de induzir a criação de cursos de doutorado na área de bioinformática, o que já aconteceu em duas universidades (a de São Paulo e a Federal de Minas Gerais).

Os estudos de genomas, como vimos, são importantes para produzir um grande volume de informações sobre a anatomia molecular de uma espécie. Tais informações podem ser usadas como pontos de partida para a produção de novos conhecimentos científicos através de diferentes modelos experimentais, seja *in vitro*, *in vivo* ou *in silico*. Essa última abordagem é representada por metodologias baseadas na criação de algoritmos dentro dessa nova e importante ciência do século 21, a bioinformática. ■