Bioinformática e Biodiversidade

por João Carlos Setubal

Instituto de Química da Universidade de São Paulo

setubal@iq.usp.br

Resumo de palestra apresentada em mesa redonda da 71ª reunião da SBPC, Campo Grande, MS, julho de 2019

Biodiversidade é um conceito bem conhecido do público em geral; bioinformática nem tanto. O que um teria a ver com o outro? Este texto procura responder a esta pergunta.

Bioinformática pode ser entendida como o desenvolvimento e uso de métodos e técnicas computacionais para o estudo de problemas da biologia em geral, especialmente da biologia molecular. Há uma grande variedade de tais métodos e técnicas, incluindo por exemplo gerenciamento de dados e algoritmos analíticos. Colocado deste forma, a relação da bioinformática com pesquisa em biodiversidade poderia ser considerada semelhante à relação da bioinformática com qualquer outro ramo da biologia. Mas o que possivelmente destaca a bioinformática quando falamos de biodiversidade é a importância de métodos computacionais para o tema específico da genômica.

A genômica vem passando por uma revolução desde o início dos anos 1990, quando surgiram os primeiros sequenciadores de DNA automáticos comerciais. Desde essa época, os sequenciadores de DNA vem tendo um aperfeiçoamento tecnológico vertiginoso, com um decréscimo exponencial no custo e um crescimento exponencial na capacidade de geração de dados. Em consequência, tem havido uma geração superexponencial de dados genômicos, o que tem causado enorme impacto nas mais variadas áreas das ciências da vida, mas especialmente em qualquer área para a qual o conhecimento e o levantamento da biodiversidade é importante.

A bioinformática já era importante mesmo antes desta revolução, visto que a simples tarefa de comparar duas sequências de DNA ou de proteínas é muito melhor executada por um programa do que manualmente. Com o crescimento exponencial dos dados genômicos, a importância da bioinformática tem crescido ainda mais. Cabe também mencionar que a bioinformática é não apenas essencial para análise de dados genômicos, como também de dados que são derivados da genômica, especialmente dados de expressão gênica, que habitualmente recebe o nome de transcritômica.

A relação entre a genômica e biodiversidade pode ser ilustrada por uma recente iniciativa internacional chamada *Earth Biogenome Project* (EBP) [1]. Trata-se de um projeto em consórcio cujo objetivo é sequenciar o genoma de representantes de todas as espécies de eucariotos que se conhece, num horizonte de 10 anos. Na verdade o EBP pode ser melhor descrito como um projeto guarda-chuva, abrigando diversas outras iniciativas de menor escopo, como projetos de sequenciamento de genomas de vertebrados e projetos relacionados a plantas. A justificativa principal para esses projetos é ajudar a compreensão, a nível molecular, da biodiversidade do planeta. A primeira frase do artigo que lançou o

EBP [1] é: “Increasing our understanding of Earth’s biodiversity and responsibly stewarding its resources are among the most crucial scientific and social challenges of the new millennium.” Dado que vastas quantidades de dados genômicos serão gerados pelo EBP, fica evidente a importância da bioinformática para essa iniciativa.

Quando se fala em biodiversidade, a atenção do público em geral é cativada pelos grandes animais, como elefantes, leões, ursos polares e baleias azuis. Entretanto é importante notar que existe uma outra biodiversidade, invisível, que é a biodiversidade dos microorganismos. Estes podem ser encontrados em grande abundância por todo o planeta, e desempenham funções absolutamente fundamentais nos mais variados processos biológicos.

Inevitavelmente, a revolução genômica também alcançou os genomas de microorganismos. Historicamente, na verdade, foi com eles que a revolução se iniciou, com o sequenciamento do genoma da bactéria *Haemophilus influenzae* em 1995 [2] representando um grande marco na história da genômica (e da bioinformática). Hoje em dia já existem centenas de milhares de genomas de procariotos disponíveis no GenBank, e num futuro próximo esse número deverá passar de um milhão.

Nos últimos 10 anos ocorreu a popularização de uma técnica conhecida como metagenômica [3], que permite a extração do DNA de amostras de ambientes de interesse onde vivem comunidades microbianas. Essa técnica permite que conheçamos genomas de bactérias, arquéias e virus que vivem em tais ambientes e que de outra forma não poderíamos obter. Dados metagenômicos portanto estão permitindo o acesso à biodiversidade de microorganismos numa escala sem precedentes. Uma iniciativa importante nessa área é o projeto do Microbioma Humano, que visa fazer um levantamento detalhado das comunidades microbiomas que habitam as diversas parte de nosso corpo [].

O processamento de dados metagenômicos depende de técnicas sofisticadas de bioinformática, em particular para que sejamos capazes de recuperar os genomas neles contidos e identificar os organismos aos quais pertencem (a vasta maioria dos quais são novos para a ciência). Tal como anteriormente aconteceu com a genômica de isolados, a metagenômica tem permitido (e exigido) o desenvolvimento de diversos novos métodos computacionais, ou seja, tem propiciado grandes avanços na bioinformática.

Um dos mais recentes resultados das explorações do mundo dos microorganismos tem sido a descoberta da diversidade de virus, em particular de uma classe de virus conhecida como bacteriófagos, ou simplesmente fagos. Embora fagos sejam conhecidos desde o início do século 20, por muitas décadas se pensava que genomas de fagos ocupavam a menor posição na escala genômica, com tamanho variando de poucos milhares de pares de bases (kbp) a dezenas de kbp. No entanto, hoje sabemos que há fagos com centenas de kbp, sendo maiores do que os menores genomas de bactérias [5]. O recordista no momento é um fago de 716 kpb [6], sendo que o genoma da bactéria *Mycoplasma genitalium* tem apenas 580 kbp.

A menção aos fagos me permite finalizar com um exemplo concreto de uma contribuição da bioinformática ao estudo da biodiversidade: trata-se do programa MARVEL [7], desenvolvido para identificar genomas de fagos em dados metagenômicos. Com o uso dessa ferramenta e de outras, é de se esperar que muitas outras descobertas ainda venham a ser feitas no campo da genômica de

microorganismos, contribuindo desta forma para aumentar nosso conhecimento sobre essa biodiversidade invisível.

Referências

[1] H. A. Lewin et al. Earth BioGenome Project: Sequencing life for the future of lifes. *Proceedings of the National Academy of Sciences of the USA* 115 (17):4325-4333, 2018.

[2] R. D. Fleischmann et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496-512, 1995.

[3] S. Nayfach and K. S. Pollard. Toward Accurate and Quantitative Comparative Metagenomics. *Cell* 166:1103-1116, 2016.

[4] E. Pasolli et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 176(3):649-662, 2019.

[5] R. A. Edwards. Prodigious Prevotella phages. *Nature Microbiology*, 4:550-551, 2019.

[6] B. Al-Shayeb et al. Clades of huge phage from across Earth's ecosystems. bioRxiv, http://dx.doi.org/10.1101/572362, 2019.

[7] D. Amgarten, L.P.P. Braga, A.M. da Silva, J.C. Setubal. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Frontiers in Genetics*, section Bioinformatics and Computational Biology 304, 2018.