Data: 07-12-2020

Título: O código da vida, decifrado

Pub:

Tipo: Internet

Secção: Nacional

Arlindo Oliveira

OPINIÃO

O código da vida, decifrado

Um problema que, durante mais de meio século, desafiou os melhores grupos de investigação do mundo, foi agora resolvido pelo programa AlphaFold.

7 de Dezembro de 2020, 7:30

Sabe-se, há muito tempo, que determinadas características dos progenitores passam para os descendentes. Porém, os mecanismos de transmissão destas características só foram compreendidos há relativamente pouco tempo. Mendel descobriu, e publicou em 1866, algumas das regras que controlam a transmissão destas características, mas o seu trabalho permaneceu ignorado durante décadas e só foi redescoberto no princípio do século XX. Quando, em 1859, Charles Darwin comunicou a sua descoberta (que partilhou com Alfred Russel Wallace) de que este mecanismo de herança de características estava na origem de todas as espécies que existem no planeta, revolucionou a nossa compreensão do mundo. Existia, afinal, uma resposta simples, óbvia e definitiva para a questão: o que somos, de onde vimos, como aparecemos neste planeta? Mas Darwin não sabia como era feita a transmissão de características nem conhecia o trabalho de Mendel. Foi preciso esperar mais um século até ficar definitivamente esclarecido qual o mecanismo biológico usado pela natureza para passar as características dos progenitores para os seus descendentes, criando a variação, mas também a continuidade, que tornam possível o processo evolutivo.

De facto, foi apenas em 1953, quando James Watson e Francis Crick publicaram, com base em trabalho de Rosalind Franklin, a estrutura em hélice da molécula do ADN, que percebemos, finalmente, o código que especifica as características de cada ser humano, animal, planta e organismo do planeta. Embora o ADN fosse conhecido há muito tempo, não se conhecia o papel que desempenhava nas células. Max Delbrück, um conhecido biólogo, chamou-lhe mesmo “essa estúpida molécula”, porque ocorria com bastante abundância nas células mas não parecia desempenhar nenhuma função relevante. Porém, o ADN guarda apenas um código, e não desempenha directamente nenhuma operação nas células. A molécula de ADN é apenas uma descrição, um esquema, codificado, de uma célula. Cada uma das nossas células, que em muitos casos têm dimensões de apenas milésimos de milímetro, contém no seu núcleo o genoma, moléculas de ADN que guardam uma longa sequência de três mil milhões de símbolos, cada um deles um A, T, C ou G. Se

estas sequências fossem impressas em papel, cada uma das nossas células precisaria de usar 3000 livros de 500 páginas cada para guardar o seu genoma.

Para transformar essas sequências numa célula é necessário decifrar o código, interpretando os genes que estão algures nestas sequências de milhões de caracteres e que são as receitas para as proteínas, as máquinas moleculares que constituem cada uma das nossas células. Estas receitas são escritas de acordo com o código genético: a cada grupo de três letras no ADN corresponde uma molécula em particular, um aminoácido, escolhido de um conjunto de 20. Uma determinada sequência, possivelmente muito longa, destes 20 aminoácidos, corresponde a uma determinada proteína. Conhecemos os nomes de algumas destas proteínas, como hemoglobina, insulina ou colagénio, mas de facto existem milhares de proteínas diferentes, que executam diferentes funções nos organismos. A função de cada proteína é definida pela sua estrutura terciária (ou, simplesmente, estrutura), a designação que é dada à forma tridimensional que a proteína assume quando os diferentes aminoácidos que a compõem se organizam no espaço, atraídos e repelidos uns pelos outros, de acordo com as suas características físicas.

Em princípio, a análise da sequência de aminoácidos de uma proteína permite determinar univocamente a sua estrutura, que é o primeiro passo necessário para determinar a sua função. De facto, dentro de uma célula, uma proteína assume a sua conformação final numa fracção de segundo, após ter sido construída a partir da receita descrita no gene. Na prática, determinar, usando um computador, a forma que uma proteína assume, a partir da sequência de aminoácidos que a constitui, é um problema muito difícil, que era até há pouco tempo impossível de resolver eficazmente. Para muitas proteínas, a única forma de determinar a sua estrutura consistia em levar a cabo processos laboratoriais complexos, laboriosos e demorados. Para muitas proteínas, não é sequer possível determinar a sua estrutura experimentalmente, porque não se prestam aos métodos de análise existentes.

Físicos, biólogos e outros cientistas tentaram, durante décadas, escrever algoritmos, traduzidos em programas de computador, que determinassem a estrutura de proteínas, uma tarefa muito complexa porque a proteína pode, em princípio, assumir um número astronómico de configurações diferentes. Sabe-se que o problema é muito, muito, difícil, porque pertence a uma classe de problemas que ficam rapidamente mais complicados quando a dimensão (neste caso, o comprimento da cadeia de aminoácidos) aumenta. Em 1994, teve lugar a primeira edição da competição CASP, que desde então se tem repetido de dois em dois anos. Na CASP, equipas de investigação de todo o mundo competem para ver quem consegue determinar com maior precisão, por via computacional, a estrutura de determinadas proteínas a partir da sua sequência de aminoácidos. Como se trata de um problema de grande importância científica e prática, em medicina e na indústria farmacêutica, milhares de investigadores e equipas têm competido ao longo destes anos, obtendo resultados que têm melhorado de forma progressiva e gradual ao longo do tempo. Já foram usadas as abordagens mais diversas, que vão desde a modelação directa das forças entre átomos até métodos baseados em regras empíricas, mas até este ano, os resultados estavam longe de ser brilhantes. Mesmo os melhores programas erravam muito, especialmente quando se tratava de determinar a estrutura de proteínas que eram muito diferentes de quaisquer outras proteínas com estrutura conhecida.

Este ano, a empresa DeepMind, detida pela Google, que já tinha ficado famosa em 2016 com a criação de um programa, o AlphaGo, que bateu o campeão do mundo no difícil jogo do Go, venceu esta competição, com o programa AlphaFold. Este programa não só ficou em primeiro lugar como conseguiu obter resultados excepcionalmente bons, atingindo uma precisão superior a 92%, um valor que deve ser comparado com o valor de 40% do melhor programa de 2016. O resultado do AlphaFold é tão bom que permite a sua utilização prática na determinação da estrutura de quase todas as proteínas. Um problema que, durante mais de meio século, desafiou os melhores grupos de investigação do mundo, foi agora resolvido por este programa, que usa técnicas de aprendizagem automática, baseadas em redes neuronais, para aprender a determinar a forma como uma proteína se dobra sobre si mesma. Concretiza-se, assim, um sonho antigo e abre-se mais uma importantíssima área de aplicação para a Inteligência Artificial.

Muitas vezes, as discussões sobre a Inteligência Artificial focam-se nos riscos e nos problemas que a tecnologia levanta, ao criar visões distorcidas da realidade, invasões de privacidade ou desemprego. Mas é importante lembrar que a razão que levou ao desenvolvimento desta tecnologia é a esperança de que ela será útil e benéfica para a humanidade, ao permitir-nos fazer novas descobertas científicas, criar novos modelos de negócio e descobrir novas soluções para problemas que, de outra forma, não conseguiríamos resolver.

Professor do IST e director do INESC