Bioinformática como suporte às biociências

Lurdes Jorge

Escola Superior Agrária de Bragança

Nas últimas décadas os progressos verificados a nível da biologia molecular e das tecnologias do DNA (nomeadamente a utilização da PCR-"Polymerase Chain Reaction" e dos processos de sequenciação automática, de ESTs-"Expressed Sequence Tags" e de SNPs-"Single Nucleotide Polymorphisms") levaram ao crescimento exponencial da informação biológica produzida pela comunidade científica. Para tal contribuiu muito o aparecimento de inúmeros projectos de sequenciação de genomas, sobretudo o ''Human Genom Project'', com início em 1990.

No ano de 1996 existiam cerca de 280 000 ESTs nas bases de dados (1). Esta avalanche de informação teve de ser armazenada e organizada. No ano de 1998 as três grandes bases de dados de sequências nucleotídicas existentes (no Japão, Reino Unido e Estados Unidos da América, respectivamente "DNA Database of Japan": DDBJ, "European Molecular Biology Laboratory": EMBL, e GenBank) estabelecem um protocolo de colaboração, criando o "International Nucleotide Sequence Database" (INSD). Em cada uma delas se podem anotar novas sequências nucleotídicas decorrentes de trabalhos de investigação, ou corrigir as pré-existentes. Todas as correcções e novas entradas são partilhadas diariamente entre as três bases de dados, pelo que em todas elas a informação disponível é a mesma, embora com algumas diferenças de formato. O acesso aos dados é livre.

No entanto, para que todos estes dados pudessem ter utilidade para a comunidade científica, teriam de ser analisados e estruturados. Isto só se tornou possível com o desenvolvimento concomitante de um "sector" da informática computacional que teve de se adaptar especificamente à análise de sequências biológicas de nucleótidos e de aminoácidos. As bases de dados exigiam sistemas computacionais com elevada capacidade de processamento e armazenagem. Foram criadas ferramentas específicas para análise e interpretação de dados biológicos. Desenvolveram-se novos algoritmos e programas estatísticos. Surgiu o conceito de Bioinformática, ao qual se estabeleceu analogia com a tentativa de nadar num mar de dados (2).

Actualmente, para além da localização de genes numa sequência de DNA e da sua tradução para proteína, nalgumas bases de dados podemos obter de forma rápida uma listagem de outras sequências polipéptidicas semelhantes à nossa, saber a que

organismos pertencem, que funções têm, saber o grau de semelhança entre ambas (e o nível de confiança que podemos ter nesses dados), efectuar entre elas alinhamentos globais ou locais, ver relações filogenéticas, localizar motivos e domínios, predizer a estrutura tridimensional por comparação com outras proteínas existentes nas bases de dados e previamente cristalografadas, etc. Estas bases de dados permitem hiperligações a outras relacionadas com o mesmo tema de pesquisa, e acesso a ferramentas bioinformáticas diversas, permitindo a integração da informação. São sobretudo bases de dados de proteínas. Em 2002 formou-se a UniProt, consórcio entre a Swiss-Prot/TrEMBL e a PIR, pré-existentes. É uma base de dados com muita informação adicional, funções de proteínas, referências cruzadas, e que além da componente bioinformática tem curadores a avaliar a informação disponibilizada. Todas as sequências nucleotídicas existentes nas bases de dados do INSD estão traduzidas e constam nas entradas da Uniprot.

Actualmente há cada vez mais bases de dados específicas que recolhem dados sobre determinados temas, os estruturam e comentam (p.e. a "Tumor Gene Database", a "HIV Structural Database", "The Restriction Enzime Database" ou a "Genome Database of Naturally Occurring Plasmids"). As bases de dados específicas de um dado organismo também aumentaram, devido ao grande número de projectos de sequenciação de genomas existente. São alguns de muitos exemplos a SGD (*Saccharomyces* Genome Database), a BDGP (Berkeley *Drosophila* Genome Project) e a TAIR (The *Arabidopsis* Information Resource).

Os projectos de sequenciação do genoma humano, do genoma de organismos modelo e de organismos patogénicos surgiram com objectivos muito concretos:

- A nível da **Medicina** pretendia-se conseguir efectuar um diagnóstico mais correcto de doenças (quer das influenciadas por genes quer das causadas por agentes patogénicos), detectar predisposição genética para doenças, criar novos fármacos baseados em informações moleculares, adequar substâncias activas de medicamentos e dosagens a cada paciente com base em perfis genéticos, efectuar terapia génica. Pretendia-se também avaliar os riscos da exposição a radiações ou a agentes mutagénicos químicos nos locais de trabalho ou de residência das populações.
- As **Ciências do Ambiente** vêem na caracterização de genomas de microrganismos específicos expectativas de utilizar alguns deles como descontaminantes ambientais eficazes e seguros, como biosensores (detectores de poluição ambiental), ou como novas fontes de obtenção de energia (produção de biodiesel).

- Ciências como a **Bioarquelogia**, a **Antropologia** e as que estudam a evolução e as migrações humanas esperam obter novos conhecimentos por análise de linhagens de origem materna ou de mutações ocorridas no cromossoma Y.
- A **Ciência Agronómica** pretende a obtenção de plantas mais resistentes a pragas e doenças, mais adaptadas a situações ambientais extremas (seca, salinidade,...), mais produtivas e/ou com um maior valor alimentar.
- A **Zootecnia** utiliza os novos conhecimentos para a obtenção de animais mais produtivos e sãos.
- A **Ciência Forense**, pela análise de DNA presente em locais de crime pretende identificar suspeitos potenciais, e também ilibar suspeitos indevidamente acusados. Por análise de DNA é também possível a identificação de vítimas de catástrofes ou de crimes, ou o estabelecimento da paternidade. As análises de DNA permitem ainda detectar a contaminação e a origem ou autenticidade de alimentos, efectuar determinações de "pedigree" em animais ou avaliar a aceitação de órgãos em programas de transplantação.

No final do ano de 2005 estavam referenciados 243 genomas completos (3), dos quais apenas 24 eram eucariotas. Em Abril de 2007 estão referenciados 832 genomas completos, entre os quais 42 de eucariotas, 32 de *Archaea*, 204 de vírus, e a esmagadora maioria de bactérias – 554 (66,6%) (4). A informação contida nestes genomas só terá alguma utilidade para a comunidade científica se estiver facilmente disponível e organizada. A sequenciação de genomas envolve não só uma completa gama de genes, mas também a sua localização nos cromossomas, relações de similaridade intra-específicas e inter-específicas entre genes, e a classificação funcional e filogenética de proteínas. Para além de um maior volume de dados, a Bioinformática teve de (co)evoluir e adaptar-se a dados e a objectivos de análise diferentes, criando ferramentas diferentes.

Referências

1. http://www.ornl.gov/sci/techresources/Human_Genome/project

2. Ross, D., 2001. Computational biology. Bioinformatics-trying to swim in a sea of data. *Science*, **291**(5507):1260-1.

3. http://cgg.ebi.ac.uk/services/cogent/stats.html

4. http://igweb.integratedgenomics.com/GOLD/