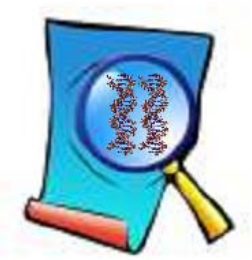
Este documento é a versão produzida pelos autores. A versão definitiva do documento está publicada na edição nº 3 de Outubro de 2003 na revista centroatlantico.pt (http://www.centroatl.pt/revista/).

Bioinformática - Exploração da Informação

Francisco M. Couto (http://www.di.fc.ul.pt/~fjmc)
Mário J. Silva (http://xldb.fc.ul.pt/mjs)



A biotecnologia tem como objectivo a produção e transformação industrial de materiais de natureza biológica. Os exemplos mais conhecidos de aplicações deste tipo de tecnologia são: o conhecimento do genoma de cada indivíduo para previsão de doenças e eventual tratamento pelos medicamentos mais apropriados; a manipulação genética de sementes permitindo obter plantas de maior rendimento; a substituição de materiais poluentes como os plásticos, combustíveis e antibióticos por materiais de origem biológica com um nível de poluição muito inferior.

A bioinformática é por sua vez uma disciplina científica recente, cujo principal objectivo é a produção de conhecimento de interesse para a biotecnologia. Estuda técnicas inovadoras de manipulação, gestão, e análise de grandes quantidades de informação biológica, permitindo aos cientistas extrair conhecimento a partir dessa informação. As fronteiras que limitam a variedade das aplicações da bioinformática são difíceis de identificar, pois esta integra conhecimentos de diversas áreas da ciência, como a biologia, a bioquímica, a estatística, a matemática e, naturalmente, a informática. O factor comum de todas as suas aplicações é o uso de sistemas computacionais no tratamento de informação biológica para a obtenção eficaz de importantes resultados científicos.

Nome	Principais Características	Endereço na Web
Oracle	*Muito utilizado pela indústria;* *Grande capacidade de dados;* *Sistema comercial*	http://www.oracle.com/
PostgreSQL	*Tipo de dados flexíveis;* *Vasto conjunto de funcionalidades;* *"Open Source"*	http://www.postgresql.org/
MySQL	*Facilidade na instalação e no uso;* *Radipez na execução das operações* *"Open Source"*	http://www.mysql.com/

Tabela 1. Principais sistemas de gestão de bases de dados (SGDBs) utilizados em Bioinformática.

O grande interesse pela bioinformática nos últimos anos deve-se sobretudo à explosão da informação disponível proveniente dos esforços de sequenciação dos genomas de diferentes

organismos. Esta informação permitiu o estudo de processos biológicos relacionados com o genoma, o que gerou ainda mais informação. Para a gerir têm sido criadas diversas bases de dados de grande dimensão (Tabela 1). Por exemplo, a base de dados GenBank (http://www.ncbi.nih.gov/GenBank/) disponibilizava através da Internet em Julho de 2003 cerca de 20GB só em sequências, resultante de um crescimento exponencial desde a sua criação. Este valor não conta com a informação descritiva de cada sequência, que é ainda de maior dimensão e de enorme importância.

A gestão destas bases de dados afigurou-se desde cedo como um processo complexo. A ausência de recursos para caracterização das entidades armazenadas foi infelizmente acompanhada pela utilização de métodos simplistas de anotação, causas da maioria das incongruências encontradas presentemente nas bases de dados. A integração de diversas fontes de informação é uma forma viável de completar e corrigir o conhecimento sobre as entidades biológicas, mas o objectivo, a estrutura, a nomenclatura e o tipo de informação variam nas diferentes bases de dados, tornando assim pouco viável a sua integração. Contudo, todo esse conhecimento biológico está presente na literatura, pois esta tem sido o meio preferido para divulgação do conhecimento científico. A maioria das bases de dados tem equipas de peritos que procuram informação relevante para a sua base de dados através da leitura de artigos científicos, cuja falta de estrutura dificulta o tratamento automático. Estes factos motivam o desenvolvimento de ferramentas automáticas que possam extrair parte desta informação, ou que pelo menos permitam uma melhor orientação no trabalho destas equipas.

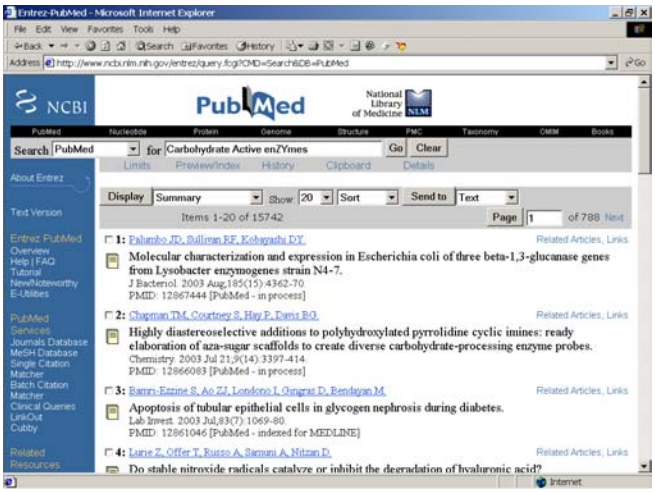


Figura 1. PubMed (http://www.ncbi.nih.gov/PubMed/) é um serviço da "National Library of Medicine" que actualmente disponibiliza o acesso a mais de 12 milhões de citações armazenadas no repositório de literatura biológica MEDLINE. Cada citação é composta pelo título, nome dos autores, sumário e outros dados que descrevem o artigo citado. Estes dados são disponibilizados também em formato XML, o que facilita a exploração desta informação de uma forma automática.

Técnica	Aplicação
Clustering	*Agrupar as entidades biológicas de acordo com propriedades comuns.*
Classificação	*Atribuir uma determinada propriedade a um*

	conjunto de entidades biológicas.
Regressão	*Extrapolar uma tendência num conjunto de experiências biológicas.*
Combinação de Estimativas	*Melhorar a precisão das estimações através da combinação de diferentes técnicas.*

Tabela 2. Principais técnicas de prospecção de dados utilizadas em Bioinformática

Com a disponibilização da literatura biológica na Internet em formato electrónico (Figura 1), o estudo de métodos de extracção e prospecção de dados (ou "data mining") desta literatura constituiu-se recentemente como um tópico de investigação muito activo (Tabela 2). Estes métodos têm por objectivo identificar e estruturar informação relevante expressa nos textos de publicações científicas para posterior inserção em bases de dados. Estão em curso um grande número de projectos que têm como objectivo o desenvolvimento de sistemas de extracção automática de informação da literatura científica para catalogação em bases de dados de informação biológica. Todavia, o uso de diferentes nomenclaturas, a heterogeneidade da informação, e a subjectividade dos resultados têm sido obstáculos difíceis de transpor, ao contrário do que hoje se alcança noutros domínios, como na identificação automática de entidades reconhecidas em texto retirado de jornais noticiosos, onde é já possível alcançar níveis de qualidade equivalentes aos de um perito humano. As soluções mais utilizadas para aumentar a eficácia dos métodos de extracção de conhecimento a partir da literatura biológica baseiam-se na integração de informação específica a cada problema[i]. Este tipo de abordagem tem custos muito elevados, pois exige um grande esforço dos peritos para adaptar o método ao problema a resolver. Custos esses que muitas vezes não são compensados pelos resultados obtidos.

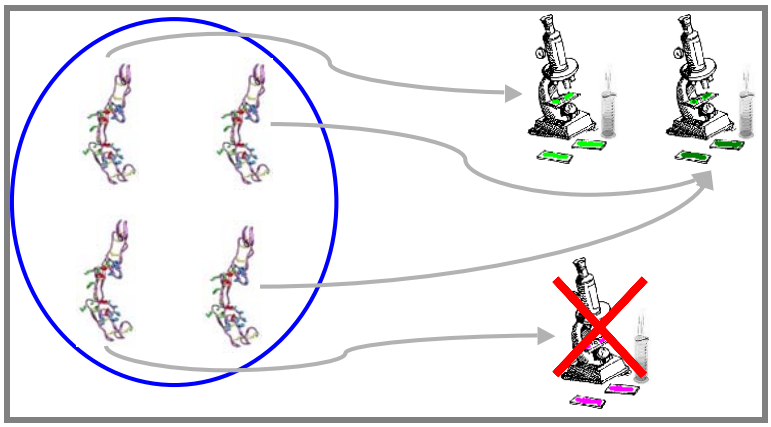


Figura 2. Um grupo de entidades biológicas com estrutura semelhante (dentro do círculo azul) tendem a ser anotadas com propriedades biológicas semelhantes. Desta forma as propriedades que não seguem esta regra tem uma forte probalidade de estarem erradas e por isso devem ser corrigidas, como é o caso da propriedade assinalada com uma cruz encarnada.

Com vista a atenuar este problema, o projecto ReBIL (Relacionamento de Informação Biológica através da Literatura) (http://xldb.fc.ul.pt/rebil/) propõe-se desenvolver métodos para melhorar a extracção de informação a partir da literatura biológica de uma forma totalmente automática, explorando a correlação biológica entre a estrutura e a função das entidades biológicas como forma de validar a informação extraída automaticamente (Figura 2). O projecto faz parte de um conjunto de iniciativas para dinamizar a investigação e

ensino da bioinformática em Portugal. A Universidade de Lisboa lançou recentemente os primeiros graus de Mestre e de Doutor em bioinformática, integrados num novo programa de pós-graduação criado numa iniciativa conjunta da Faculdade de Ciências da Universidade de Lisboa e do Instituto Gulbenkian de Ciência (http://bioinformatics.fc.ul.pt/). A integração dos formandos no tecido das infra-estruturas existentes, tanto nacionais como internacionais em que Portugal participa, poderá levar à progressiva criação de estruturas empresariais que levarão as aplicações desta nova ciência aos vários sectores da sociedade portuguesa.

---

[i] Na maioria dos casos estas soluções baseiam-se no desenvolvimento de regras gramaticais que quando aplicáveis a uma parte do texto permitem identificar informação relevante. Recentemente, tem sido também utilizado "Hidden Markov Models" para evitar a criação manual destas regras.