RECURSOS DE BIOINFORMÁTICA APLICADOS ÀS CIÊNCIAS ÔMICAS COMO GENÔMICA, TRANSCRIPTÔMICA, PROTEÔMICA, INTERATÔMICA E METABOLÔMICA

BIOINFORMATIC RESOURCES APPLIED ON THE OMIC SCIENCES AS GENOMIC, TRANSCRIPTOMIC, PROTEOMIC, INTERATOMIC AND METABOLOMIC

Foued Salmen ESPINDOLA[1]; Luciana Karen CALÁBRIA[1,2]; Alexandre Azenha Alves de REZENDE[1,2]; Boscolli Barbosa PEREIRA[1,2]; Flávia Assumpção SANTANA[1,2]; Isabel Marques Rodrigues AMARAL[1,2]; Janaina LOBATO[1,2]; Juliana Luzia FRANÇA[1,2]; Justino Luiz MARIO[1,2]; Leonardo Bruno FIGUEIREDO[1,2]; Luana Pereira dos SANTOS-LOPES[1,2]; Neire Moura de GOUVEIA[1,2]; Rafael NASCIMENTO[1,2]; Renata Roland TEIXEIRA[1,2]; Taís Alves dos REIS[3]; Thaise Gonçalves de ARAÚJO[1,2]

1. Instituto de Genética e Bioquímica - INGEB, Universidade Federal de Uberlândia - UFU, Campus Umuarama, Uberlândia, MG, Brasil. foued@ufu.br . 2. Programa de Pós-Graduação em Genética e Bioquímica INGEB - UFU; 3. Programa de Pós-Graduação em Odontologia, Faculdade de Odontologia – Universidade Federal de Uberlândia - UFU, Uberlândia, MG, Brasil.

RESUMO: As ciências ômicas tratam da análise global dos sistemas biológicos, integrando diferentes áreas do conhecimento, como a bioquímica, genética, fisiologia e computação, com o objetivo de isolar e caracterizar genes, proteínas e metabólitos, assim como estudar as interações entre eles, com base em técnicas experimentais, *softwares* e bancos de dados. A bioinformática por sua vez, propõe novas formas de ciência baseada na experimentação *in silico*, sendo muito dinâmica na sua atualização e fornecendo a base para geração de novos dados e conhecimentos que podem ser aplicados na pesquisa básica e na aplicada com o desenvolvimento de novos produtos e soluções. Este processo está intimamente relacionado à inovação tecnológica, que é conseguida unindo-se a biotecnologia e a bioinformática. Contudo, o objetivo desta revisão é apresentar uma pequena abordagem dos recursos de bioinformática aplicados às ciências ômicas, como genômica, transcriptômica, proteômica, interatômica, metabolômica, farmacogenômica, dentre outras.

PALAVRAS-CHAVE: Ômica. Bioinformática. Biotecnologia. Bancos de dados.

INTRODUÇÃO

Os dados gerados pelo sequenciamento dos genomas de diferentes organismos transformaram a biologia. A integração de várias áreas do conhecimento permitiu avançar os estudos em relação à genômica, os processos de transcrição das informações contidas nos genes, a transcriptômica, bem como a compreensão do conjunto dos produtos destes genes pela proteômica. No início desta década com o advento do genoma humano também se iniciava as discussões e as ações para uma nova era da biologia, a "era pós-genômica". Neste contexto, promoveu-se o desenvolvimento e o aperfeiçoamento das técnicas que permitiram os avanços destas novas ciências ômicas (Figura 1), como a transcriptômica, proteômica e metabolômica, com o objetivo de isolar e caracterizar o RNA, as proteínas e os metabólitos, respectivamente; sendo possível devido também ao desenvolvimento da bioinformática.

O termo "ômicos" refere-se à análise global dos sistemas biológicos. Além das citadas anteriormente, uma variedade de subdisciplinas ômicas têm surgido, cada uma com seu próprio conjunto de instrumentos, técnicas, *softwares* e base de dados. Entre as tecnologias ômicas que impulsionam estas novas áreas de investigação, mencionam-se as tecnologias de DNA e *microarrays*, a espectrometria de massas e uma série de outras tecnologias e instrumentação que permitiram uma alta capacidade de análise (WINGENDER et al., 2007).

O domínio da bioinformática cresceu em paralelo e com a internet, em que a rápida análise de dados e a troca de informações sobre os códigos biológicos e computacionais estão em convivência harmônica, por meio de suas múltiplas ramificações, gerenciando e integrando bancos de dados aplicáveis, e construindo sistemas *in silico* para simulação de formas naturais e modificadas de produtos específicos.

Todos os projetos de sequenciamento genômico realizados e em andamento, tanto de procariotos como de eucariotos, continuam a nos lembrar que o nosso conhecimento sobre o funcionamento de um organismo ou célula, a nível molecular, é realmente muito limitado. Desta forma,

Received: 24/04/09
Accepted: 06/10/09

o aumento substancial de sequências e de informações produzidas pelo rápido avanço das ciências ômicas está ajudando a prover novos caminhos da exploração de textos pela bioinformática (YANDELL; MAJOROS, 2002).

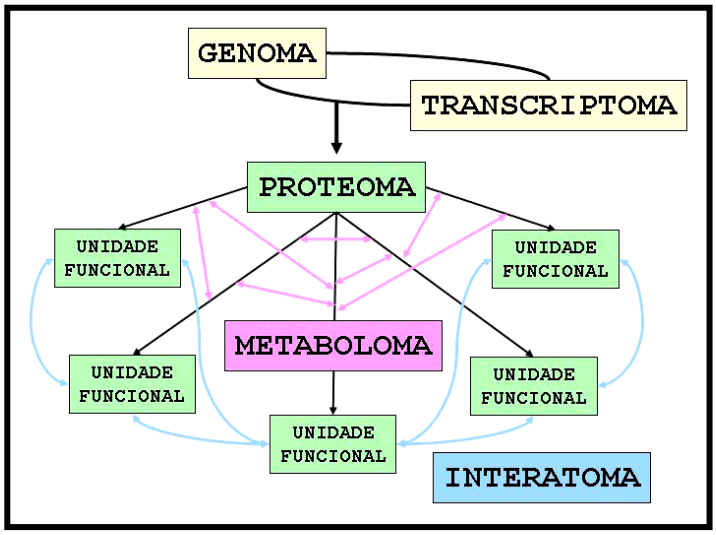

Figura 1. Representação esquemática das principais ciências ômicas

Tantos são os temas a serem discutidos sobre os vários aspectos da bioinformática, todos eles na fronteira do conhecimento. Por isso, apresentamos aqui uma pequena abordagem baseada nos seminários desenvolvidos no Programa de Pós-Graduação em Genética e Bioquímica da Universidade Federal de Uberlândia na disciplina de Bioinformática. O quadro 1 apresenta páginas da *web* com recursos de bioinformática.

Quadro 1: Lista de páginas da Web com recursos de bioinformática e que são abordadas neste artigo.

NOME	ENDEREÇO
PFAM	http://pfam.jouy.inra.fr/
SANGER	http://www.sanger.ac.uk/Software/Pfam/
Blast	http://www.ncbi.nlm.nih.gov/BLAST
CaM Target	http:// calcium.uhnres.utoronto.ca/ctdb/ctdb/home.html
CAP3	http://genome.cs.mtu.edu/cap/cap3.htm
CAS	http://www.cas.org/
Clustal	http://www.clustal.org/
Cytoscape	http://www.cytoscape.org/
Drug DataBase	http://chrom.tutms.tut.ac.jp/JINNO/DRUGDATA/00database.html
Easy Align	http://www.scriptspot.com/3ds-max/easyalign
Entrez Protein	http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein
GenBank	http://www.ncbi.nlm.nih.gov/Genbank
GENE 3D	http://gene3d.biochem.ucl.ac.uk/Gene3D/
Gene Ontology	http://www.geneontology.org/
Google	http://www.google.com
GoogleScholar	http://scholar.google.com
HiMAP	http://www.himap.org/
HoGenom	http://ralyx.inria.fr/2007/Raweb/helix/uid41.html
INSDC	http://insdc.org
Interpare	http://interpare.net/
InterPro	http://www.ebi.ac.uk/interpro/
KEGG	http://www.genome.jp/kegg/
NCBI	http://www.ncbi.nlm.nih.gov
Osprey	http://biodata.mshri.on.ca/osprey/servlet/Index

PANTHER	http://www.pantherdb.org/
Pfam	http://www.sanger.ac.uk/Software/Pfam/
Phrap	http://www.phrap.org
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
PIRSF	http://pir.georgetown.edu/iproclass/
PRF	http://www.prf.or.jp/en/index.shtml
PRINTS	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/
ProDom	http://prodom.prabi.fr/prodom/current/html/home.php
PROSITE	http://ca.expasy.org/prosite/
PubChem	http://pubchem.ncbi.nlm.nih.gov/
PubMed	http://www.pubmed.com
RNAMOTIF	http://www.scripps.edu/mb/case/casegr-sh-3.5.html
SBBiotec	http://www.sbbiotec.org.br/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
SetupX	http://fiehnlab.ucdavis.edu
SMART	http://smart.embl-heidelberg.de/
String	http://string.embl.de/
SUPERFAMILY	http://supfam.cs.bris.ac.uk/SUPERFAMILY/
SwissProt	http://www.expasy.ch/spro/
TIGRFAMs	http://www.tigr.org/TIGRFAMs/index.shtml
Tree View	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
UniProt	http://www.ebi.ac.uk/uniprot/
UniProtKB	http://www.ebi.ac.uk/trembl/

RECURSOS DE BIOINFORMÁTICA

Mineração de textos e dados (*text mining* e *data mining)*

O PubMed contém mais de 15 milhões de artigos científicos, e em 2003 aproximadamente 560.000 artigos foram adicionados ao Medline, sendo que o número de artigos adicionados por ano aumentou mais de 20.000/ano entre 2000 e 2003. Além de dados publicados no formato de artigos, bancos *on-line* são criados sobre muitos aspectos da biologia, contendo informações sobre interações moleculares, vias metabólicas, caracterização de genes, compostos químicos, estrutura protéica, doenças e organismos. Esta produção de dados científicos em larga escala impossibilita a extração ou análise destes dados por especialistas através de métodos manuais tradicionais, e para solucionar este problema, na década passada, métodos de *text mining* foram desenvolvidos.

O *text mining* é uma ferramenta de obtenção de dados não-estruturados, escritos em linguagem natural, extraídos a partir de um banco de dados estruturado, com o auxílio de algoritmos para análise de textos não-estruturados. O processo de *text mining* envolve três subáreas: recuperação, extração, questões e respostas da informação, permitindo a identificação de identidades biológicas e suas interações, facilitando a análise de dados.

A primeira sub-área (*information retrieval*), e mais comum em biologia molecular, consiste na extração de documentos a partir de uma grande coleção. Neste caso, há dois tipos de estratégia de busca: uma é a partir da combinação de palavras-chave e busca usando documentos como referência, para selecionar outros documentos semelhantes. Como muitas palavras são encontradas com grande freqüência e levam ao encontro de documentos com pouca informação, elas são excluídas durante o processo de busca.

Uma ferramenta de busca de dados amplamente usada é o sistema de *Information Retrieval* do EntrezPubMed fornecido pelo NCBI (*NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION*). O popular Google foi recentemente incorporado como ferramenta de busca específica para literatura acadêmica com o GoogleScholar, que recupera artigos científicos, livros e reportagens. Na figura 2 estão esquematizados outros bancos de dados separados de acordo com a aplicação biológica.

Por outro lado, o *data mining* é uma das novidades da Ciência da Computação, que utiliza vastos repertórios para tentar descobrir se há algum conhecimento escondido neles. A definição mais importante de *data mining* foi elaborada em 1996 por Fayyad et al. (1996): "... *Data mining* é um processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis". Este processo vale-se de diversos algoritmos desenvolvidos recentemente que processam os dados e encontram padrões

válidos e novos. Embora os algoritmos atuais sejam capazes de descobrir esses padrões, os analistas humanos são os principais responsáveis por essa determinação.

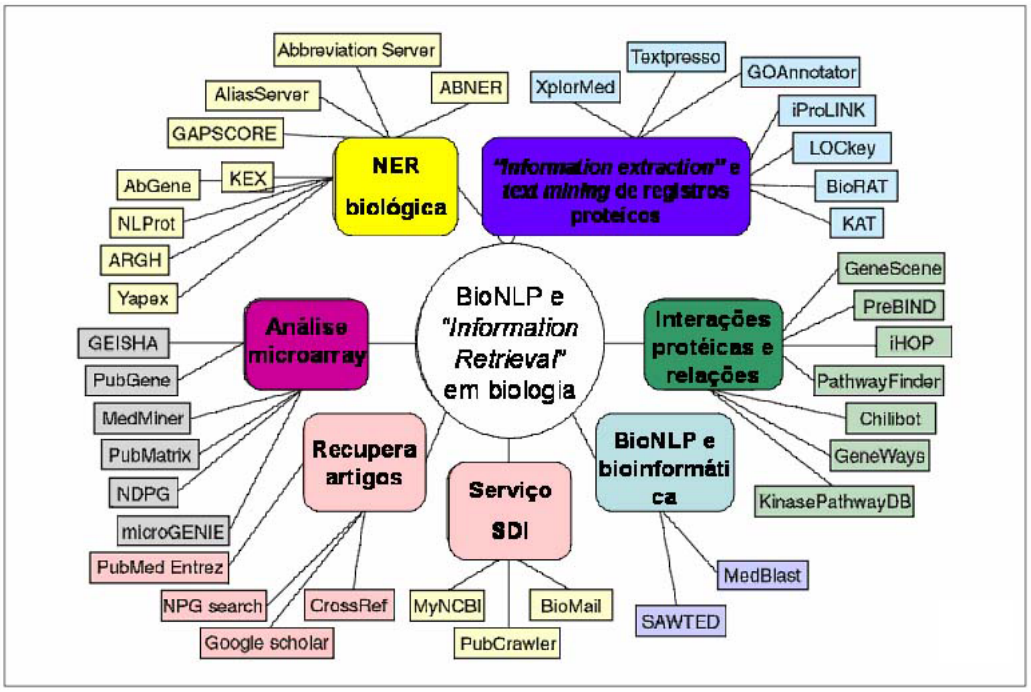

Figura 2. Uma visão do processamento da linguagem natural biológica (BioNLP) e aplicações de *text mining* na biologia. O tópico central é envolvido por sete círculos com suas aplicações correspondentes dadas por outras caixas posteriores: NER (nome da identidade reconhecida), SDI (informação seletiva disseminada), *Information extration* (obtenção de informações), Análise por *microarrays*, Recuperação de artigos, Interações protéicas e relações e BioNLP. Modificado de: Krallinger e Valencia (2005).

A evolução da informática conta um pouco sobre o surgimento do processo de *data mining*. Nos anos 60, os computadores tinham capacidade precária não dispondo de mecanismos eficientes para armazenamento de grandes volumes de dados. No início da década de 70 até 80, um grande avanço marcou os meios físicos de armazenamento de dados, o desenvolvimento de *softwares* para o gerenciamento de dados, denominados de Sistemas Gerenciadores de Bancos de Dados, e logo após o surgimento do modelo relacional, permitindo rápida recuperação de dados dirigidos. Nos anos 90, surgem os Bancos de Dados Multidimensionais ou *Data Warehouses* que propiciam o processo analítico *on-line* (OLAP). A diferença entre o OLAP e o *data mining* é que no primeiro o analista gera as hipóteses que podem ser validadas ou negadas, e no segundo o próprio sistema gera as hipóteses.

O processo geral de descoberta de conhecimento em banco de dados é composto por diversas etapas. As principais tarefas são associação, agrupamento e descoberta de regras de classificação. A tarefa de classificação pode ser realizada por algoritmos convencionais ou por métodos de inteligência artificial como, por exemplo, as redes neurais, algoritmos evolucionários, dentre outros. Além disso, trabalhos têm demonstrado a importância desta ferramenta para estudos científicos. Baseando-se na literatura e análises *in silico*, pesquisadores utilizaram o *data mining* e fizeram uma seleção de 189 candidatos à vacina contra o *M. tuberculosis*. Este repertório foi ranqueado para gerar uma lista com os 45 melhores antígenos, selecionando genes que abrangem todos os estágios da infecção, sendo incorporados no rBCG ou vacinas baseadas em subunidades (ZVI et al., 2008).

Sequenciamento do DNA e a Genômica

Os mecanismos envolvidos na expressão e interação dos genes, assim como a compreensão das redes funcionais estabelecidas pelas proteínas, fazem com que, no cenário científico atual, a genômica e a proteômica estejam cada vez mais em evidência.

Quatro são os principais bancos de dados utilizados para as diferentes análises de nucleotídeos. Um deles é o INSDC

(*INTERNATIONAL NUCLEOTIDE SEQUENCE DATABASE*) que disponibiliza um repertório de sequências e é resultado da associação de três bancos de dados parceiros, o DDBJ (*DATA BANK OF JAPAN*), o EMBL (*EMBL NUCLEOTIDE SEQUENCE DATABASE*) e o GenBank. Os registros da associação EMBL/GenBank/DDBJ incluem genes individuais, genomas completos, RNAs, anotações, sequências expressas, cDNAs e sequências sintéticas. Devido à sua designação como sendo um provedor de dados primários, o banco EMBL/DDBJ/GenBank é a fonte inicial de muitos bancos de dados em biologia molecular (TATENO et al., 2005; KANZ et al., 2005 e BENSON et al., 2005). Um exemplo de banco de dados de sequências genômicas secundárias de nucleotídeos é o Ensembl, uma fonte compreensível de anotações estáveis, em que genes são anotados por evidências derivadas de proteínas conhecidas, cDNAs e sequências expressas. Novos genes são determinados pelo sistema de construção de genes, incorporando uma variedade de métodos, incluindo homologia e predição pela aplicação do HMM (*HIDDEN MARKOV MODEL*) (HUBBARD et al., 2005).

O RefSeq (*REFERENCE SEQUENCE*) é um banco que disponibiliza sequências compreensíveis, integradas e não-redundantes, incluindo DNA genômico, transcritos e proteínas para diversos organismos (PRUITT et al., 2005). Por outro lado, o Genome Review representa uma versão da sequência original de um cromossomo ou plasmídeo, com informações importadas de fontes que incluem o UniProt (*UNIVERSAL PROTEIN RESOURCE*), Gene Ontology (GO), projeto GOA (*GO ANNOTATION*), InterPro e HoGenom, além de serem disponibilizadas referências cruzadas com 18 bancos de dados (KERSEY et al., 2005).

Milhares de sequências são obtidas através de técnicas, como o *shotgun* que foi usada, entre outros, no genoma de *Apis mellifera* (THE HONEYBEE GENOME SEQUENCING CONSORTIUM, 2006). As sequências obtidas são analisadas e reunidas por *software*, como CAP3. Após a organização das sequências é verificada a existência de redundância, a identificação das regiões codificadoras e, a seguir, identificação de funções.

Os níveis de redundância são comparados usando o *software* Phrap, que foi usado no genoma da cana de açúcar (VETTORE et al., 2003). Além disso, para a identificação do gene, em seres eucariotos, dois *softwares* foram muito utilizados, o Glimmer (DELCHER et al., 1999) e o Genemark (BORODOVSKY; MCININCH, 1993). Por meio deles se analisa as janelas de leitura na sequência (ORFs), sendo cada uma alinhada e comparada com as de outras espécies conhecidas, depositadas em bancos de dados como o GenBank e o SwissProt.

A anotação funcional se dá através da comparação das sequências obtidas com as depositadas em bancos de dados como o GenBank e o Blast, sendo o último uma ferramenta mais amplamente utilizada para esse tipo de comparação (ALTSCHUL et al., 1990). A anotação é considerada completa quando o genoma está decodificado e minimamente anotado, com seus genes identificados e conferidos.

Em 2001, por junção da iniciativa pública (*INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM*, 2001) e privada (CELERA) (VENTER et al., 2001) realizou-se o projeto Genoma Humano. Os objetivos deste projeto foram identificar todos os genes estimados do DNA humano, determinar as sequências de bases, armazenar as informações em banco de dados e desenvolver ferramentas para a análise dos dados. Entretanto, de todos os genes que já foram seqüenciados, em média, apenas 50% codificam proteínas de função conhecida. Sobre o genoma humano em particular, o banco Genew como parte do HUGO (*HUMAN GENOME ORGANIZATION*), mantém um depósito de nomes e símbolos de genes, para se definir uma nomenclatura de dados submetidos por este Genoma (WAIN et al., 2002).

Outros projetos também foram desenvolvidos, como o Projeto EST (*EXPRESSED SEQUENCE TAG*), que ao invés de sequenciar todo o genoma de um organismo e depois tentar descobrir quais são seus genes, apenas os genes expressos pelo organismo são capturados e seqüenciados. Contudo, surgiu um importante problema computacional, o *clustering*. Ou seja, qual o modo de agrupar todos os ESTs que correspondem ao mesmo gene em um único grupo (*cluster*)? Assim, vários métodos foram propostos e a partir do agrupamento de tais dados, foram criados bancos de dados EST (dbEST), que agrupam a informação de milhões de ESTs.

Mesmo com o grande avanço provocado pela bioinformática, ainda persistem vários desafios, como por exemplo, o Gene Ontology que ainda não possui condições consistentes para diferenciar todos os processos biológicos, e os dados de *microarray* que ainda não possuem uma grande reprodução de dados, resultando em um baixo poder estatístico. Muitos métodos estatísticos padrões falham por causa de problemas com o tamanho das amostragens podendo levar a uma desatualização dos bancos de dados comumente utilizados.

Filogenia

Uma árvore filogenética é uma representação gráfica em forma de árvore, que retrata a história de parentesco entre as espécies, sendo que cada aresta da árvore representa uma mutação; cada vértice, um ponto na escala evolutiva; e cada folha uma espécie. Além disso, os indivíduos partem de um ancestral comum, a raiz (PROSDOCIMI et al., 2003).

A análise das relações filogenéticas pode ser dividida em duas etapas principais. A primeira consiste na construção das listas de caracteres para a análise, ou seja, o estabelecimento das homologias, sua codificação e ordenação, resultando na matriz de dados. Já na segunda etapa, se escolhe a árvore ótima por meio da análise numérica dessa matriz. A construção da matriz de dados é um passo muito importante na obtenção de filogenias, pois a qualidade da árvore filogenética final depende diretamente de sua construção. Com esse propósito, alguns programas estão disponíveis gratuitamente para auxiliar nas análises filogenéticas. O PHYLIP (*THE PHYLOGENY INFERENCE PACKAGE*) é um pacote para a análise filogenética criado por Joseph Felsenstein, na Universidade de Washington. Esse pacote pode resolver a maioria das análises filogenéticas existentes na literatura atual, incluindo métodos como o da parcimônia, o da matriz de distância e o *likelihood*. Como também, pode trabalhar com os seguintes tipos de entrada de dados: sequências moleculares, freqüência de gene, matriz de distância e características discretas.

Outro programa de acesso livre é o Easy Align que foi desenvolvido em plataforma Windows, e possibilita o alinhamento de sequências de DNA e proteínas pelo critério de máxima parcimônia. Por outro lado, o Clustal X (THOMPSON et al., 1997) é o algoritmo mais utilizado em análises filogenéticas de dados moleculares, e para finalizar a análise, a construção da árvore é feita pelo *software* Tree View, muito utilizado nesta área.

RNAs e a transcriptômica

Atualmente, estudos moleculares requerem a interação entre análises genômicas, celulares e dados de bioinformática, a qual apresenta, gradativamente, um papel essencial na geração de resultados aliados a alta tecnologia. A evolução da bioinformática, iniciada com análises de sequenciamento, tem oferecido avanços nas ciências "ômicas", principalmente nas anotações dos transcriptomas, permitindo a interrelação entre o genoma funcional e a informação codificada.

Contudo, o sequenciamento do genoma humano demonstrou que aproximadamente 98% de todos os produtos transcritos em humanos correspondem a RNAs não-codificantes (ncRNA) (MATTICK; GARDEN, 2001). Essa descoberta levantou questões referentes à diferenciação e desenvolvimento, tanto de espécies quanto de mecanismos moleculares individuais (MATTICK, 2001). Embora o repertório de isoformas protéicas expressas nos organismos complexos seja significativamente incrementado pelo processamento pós-transcricional alternativo (*splicing* alternativo), a combinação desses polipeptídeos com os sinais ambientais fornecem informações insuficientes sobre os processos biológicos, de maneira que a maioria dos sistemas regulatórios é controlada por moléculas de RNA (MATTICK, 2003).

A predição da estrutura do RNA é normalmente baseada nas características termodinâmicas do *folding* da molécula ou na conservação filogenética das regiões de pareamento de bases e, nesse sentido, existem duas estratégias básicas na predição de ncRNAs. A primeira é baseada na homologia genômica que existe ao longo da evolução e um dos métodos computacionais é o Rfam, capaz de alinhar mais de 500 famílias de ncRNAs e determinar a estrutura predominante que resulta desse alinhamento. Outros *softwares* desenvolvidos para detectar ncRNAs é o $_{T}RNA_{SCAN}$-SE, especializado na busca por tRNAs com elevada sensibilidade e baixa proporção de falso–positivos; e os PROMIR, MIR-ABELA, MIR-SCAN, MIRSEEKER e HARVESTER, dedicados na busca por miRNAs em humanos, mamíferos, vertebrados, *D. melanogaster* e plantas, respectivamente. Esses programas utilizam tanto métodos heurísticos quanto modelos probabilísticos para capturar as sequências e características estruturais dos miRNAs (MEYER, 2007).

Se a função do ncRNA depende de elementos estruturais bem definidos e há uma sequência ortóloga de um organismo correlato, pode-se usar dois outros programas: o INFERNAL ou CMFINDER (YAO et al., 2006). Caso exista apenas um único ncRNA cuja estrutura funcional é determinada, mas nenhuma sequência equivalente é conhecida em outro genoma relacionado evolutivamente, pode-se buscar uma sequência alvo para regiões que são similares ao RNA conhecido. Nesse caso, encontra-se disponível o programa RSEARCH (KLEIN; EDDY, 2003). Como o método compara apenas a sequência alvo a uma sequência única de RNA e não a um conjunto de sequências relacionadas filogeneticamente, sua sensibilidade é tipicamente menor. Portanto, como método alternativo, encontra-se o programa RNAMOTIF, o qual permite uma definição manual de motivos,

capturando a sequência e as estruturas secundárias e terciárias (MACKE et al., 2001).

Alguns ncRNAs não dependem de uma estrutura bem definida, permitindo-se optar por uma busca baseada essencialmente na similaridade de sequência conservada ao longo da evolução, sendo suficiente para identificá-los no genoma. Nesse sentido, existem programas especializados na busca por apenas homologia de sequência (perfil-HMMs) e aqueles disponíveis para realizar o alinhamento e *folding* ao mesmo tempo. Dentre esses se encontram: o FOLDALIGN, capaz de detectar estruturas locais ao invés de identificar estruturas globais com vários *loops* (GORODKIN et al., 2001), e o DYNALIGN, que reduz a complexidade computacional limitando o espaço de busca e o tamanho dos *loops* presentes nas estruturas internas do RNA (MATHEWS; TURNER, 2002). As desvantagens desses dois últimos residem na sua incapacidade de explicitar regiões que não adquirem uma determinada estrutura e na ineficiência do alinhamento baseado na estrutura da molécula.

Em casos excepcionais, em que as sequências correspondem a determinados transcritos e sabe-se que a estrutura global desempenha um papel essencial na sua função, pode-se predizer essa estrutura para cada sequência individual através de *softwares* específicos. Nesse sentido, as duas ferramentas mais conhecidas são MFOLD e RNAFOLD, que predizem a estrutura mais estável a partir de uma sequência pré-estabelecida (ZUKER, 1989; SCHUSTER et al., 1994).

A segunda estratégia inclui a predição *ab-initio* de ncRNAs, o que constitui o maior desfio na busca dessas moléculas não-codificantes. Também existem algoritmos computacionais eficientes que objetivam predizer estruturas estáveis de RNA em grande escala genômica, como o RNAPLFOLD. Contudo, o potencial desse programa na detecção de ncRNAs ainda não foi sistematicamente investigado (MEYER, 2007).

Uma das ferramentas mais utilizadas na análise transcriptômica é a tecnologia de *microarrays* (Figura 3) que constitui uma das principais ferramentas para estudos de expressão gênica (SCHENA et al., 1996), sendo muito aproveitada na avaliação de aspectos da biologia de sistemas e o estudo dos perfis de interação entre diversas biomoléculas (KITANO, 2002).

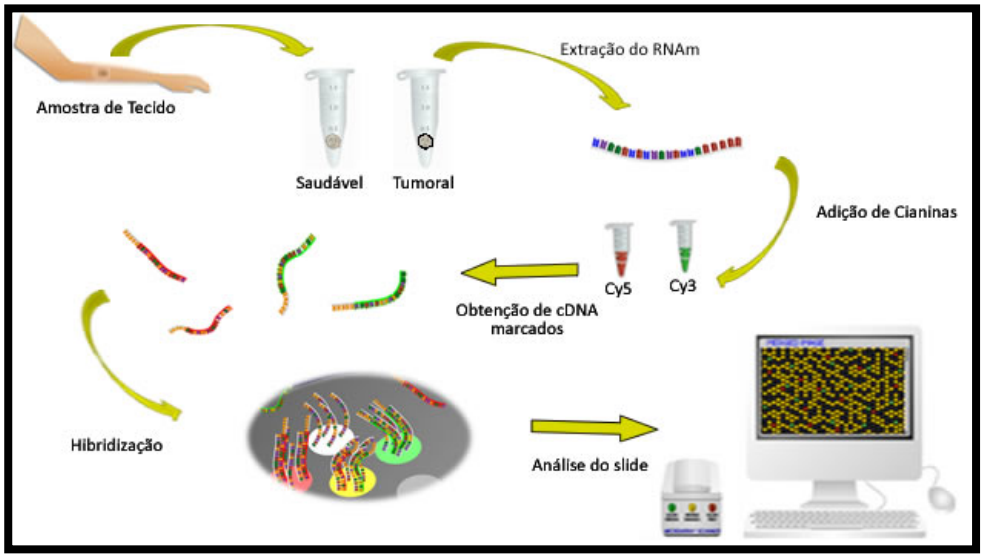

Figura 3: Experimento de *microarray.*

O primeiro *microarray* surgiu em meados da década de 1990 e possuía 45 sondas de cDNA (SCHENA et al., 1995). Com os aprimoramentos tecnológicos, no ano seguinte à sua publicação, pesquisadores apresentaram trabalhos com cerca de 1000 sondas de *arrays* (SCHENA et al., 1996; SHALON et al., 1996), sendo que atualmente é comum encontrar trabalhos que utilizem dezenas de milhares de sondas. A AFFYMETRIX foi a empresa pioneira em *microarrays*, trabalhando com a metodologia de apenas um canal (uma cor). Na tecnologia de *microarrays* com lâminas de vidro, várias sequências de DNA conhecidas (sondas), são impressas em uma mesma lâmina. Nos *arrays* de duas cores, os mRNAs são extraídos de células pertencentes às duas condições distintas e por meio de transcrição reversa, utilizando oligonucleotídeos marcados, o cDNA é obtido. Os oligonucleotídeos são marcados com corantes fluorescentes (cianinas), sendo o corante *Cy3* verde, e o *Cy5*, vermelho como

mostra o esquema 1. Após toda a experimentação biológica, as marcações são interpretadas por um *software* específico e os dados são analisados por ferramentas estatísticas.

Proteômica e estrutura de proteínas

"Como", "onde", "quando" e "por que" são produzidas centenas de milhares de proteínas individuais em um organismo vivo? Como elas interagem entre si e com outras moléculas para construir uma célula? Como elas funcionam e conduzem o desenvolvimento e crescimento programado e interagem com os ambientes biótico e abiótico? Responder todas essas questões é o objetivo da proteômica, que como uma metodologia, deve ser considerada parte de uma análise integrativa e multidisciplinar em diferentes níveis, estendendo desde os genes até o fenótipo expresso nas proteínas. Estas análises devem envolver as tecnologias "ômicas" (genômica, transcriptômica, proteômica e metabolômica) bem como as técnicas de bioquímica clássica e biologia celular.

No estudo completo das proteínas, integrando estrutura e função, os pesquisadores utilizam bancos de dados diversos que possam atender os diferentes ramos da proteômica. Um dos mais usados é o banco de dados Entrez Protein, um depósito de sequências disponibilizado pelo NCBI e compilado através de uma variedade de fontes. O banco contêm as sequências de proteínas submetidas aos bancos PIR (*PROTEIN INFORMATION RESOURCE*) (WU et al., 2003), UniProtKB/Swiss-Prot, PRF (*PROTEIN RESEARCH FOUNDATION*) e PDB.

Outro, também muito utilizado é o UniProt, um catálogo de dados de sequências e funções de proteínas, mantido pelo consórcio UniProt. O consórcio é uma colaboração entre o SIB (*SWISS INSTITUTE OF BIOINFORMATICS*), o EBI (*EUROPEAN BIOINFORMATICS INSTITUTE*) e o PIR. O banco UniProt é compreendido por três componentes, o acurado UniProtKB (*UNIPROT KNOWLEDGEBASE*), que continuou o trabalho do UniProtKB/Swiss-Prot; o UniProtKB/TrEMBL (BOECKMANN et al., 2003) e o PIR. O UniProtKB/Swiss-Prot é um banco anotado manualmente com informações extraídas da literatura e análises computacionais, contendo níveis mínimos de redundância e alto nível de integração com outros bancos de dados (BAIROCh et al., 2005).

Na análise de dados obtidos utilizando a eletroforese bidimensional, o banco de dados SWISS-2DPAGE (HOOGLAND et al., 2004) é o mais útil, pois armazena resultados experimentais que utilizam esta metodologia e acrescenta uma variedade de referências cruzadas com outros bancos de dados semelhantes, além do UniProtKB/Swiss-Prot. No entanto, se o objetivo é descrever a função molecular, o contexto biológico e a localização celular do produto gênico, o Gene Ontology é o mais indicado (CAMON et al., 2004).

O grande desafio enfrentado por estudiosos e bioinformatas é descobrir qual a estrutura tridimensional adotada pelas proteínas a partir da estrutura primária. No entanto, as ferramentas *in silico* disponíveis atualmente ainda não são totalmente confiáveis. Os métodos experimentais utilizados para obtenção da estrutura tridimensional são cristalografia por difração de raio-X e ressonância magnética nuclear. Entretanto, esses métodos podem ser onerosos e de difícil execução, além de apresentarem limitações técnicas. Estas e outras dificuldades fazem com que a quantidade de estruturas de proteínas decifradas ainda compõe uma pequena fração do total de proteínas existentes (PROSDOCIMI et al., 2003*)*.

Um método alternativo e não-experimental é a modelagem molecular, baseada em conhecimentos estereoquímicos dos aminoácidos. Uma das maneiras de se fazer a modelagem molecular é através da homologia entre sequências, em que uma delas já possui forma tridimensional definida. O primeiro passo é a pesquisa de proteínas homólogas em bancos de dados de estruturas de proteínas como o PDB (*PROTEIN DATABASE BANK*) (HULO et al., 2008), que é uma colaboração entre o RCSB (*RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS*), o MSD-EBI (*MACROMOLECULAR STRUCTURAL DATABASE*) e o PDBj (*PROTEIN DATA BANK OF JAPAN*) (BERMAN et al., 2000). A seguir, deve-se realizar o alinhamento das sequências de aminoácido das proteínas homólogas e a proteína-alvo, através do Clustal, por exemplo. A modelagem é realizada através de *softwares* como o Modeller, SWISS-MODEL, 3D-PSSM, dentre outros. Esses programas normalmente procuram encontrar a estrutura terciária que melhor se aproxime da disposição dos átomos das proteínas utilizadas como modelo, e ao mesmo tempo atenda às restrições físico-químicas (FORSLUND et al., 2008).

Outro tipo de modelagem é o Threading, que compara estrutura de uma proteína teste com a estrutura de outra proteína conhecida com uma pequena similaridade de sequência. Neste modelo é levada em consideração a distância entre os resíduos de aminoácidos, a estrutura secundária e as características físico-químicas (RATTEI et al., 2008).

Interatômica

O termo interatoma é menos bem definido do que genoma, proteoma ou transcriptoma, porque diferentes comunidades usam o termo "interação de proteína" para se referir de qualquer interação física às interações funcionais. No entanto, Sharan e Ideker (2006) caracterizaram interatoma como o estudo do conjunto de interações macromoleculares, físicas e genéticas, e uma das chaves da análise em larga escala é o alinhamento ou comparação global de duas ou mais redes (alinhamento múltiplo) para identificar regiões similares.

No estudo interatômico, uma forma de avaliar a qualidade das redes de interação proteína-proteína é comparando as interações sugeridas com a localização subcelular ou as classes funcionais da proteína, tais como o Gene Ontology (BADER; HOGUE, 2002). A suposição de tal análise é que os integrantes à interação devem pertencer à mesma categoria e a validez depende fortemente da escolha das classes. Além disso, a co-expressão dos genes correspondentes também é usada como um critério de avaliação (KEMMEREN et al., 2002).

Outra forma de validar interações é associar as proteínas dentro de uma via metabólica. Para isso existe um banco de dados, ou também nomeada de enciclopédia metabólica, chamada KEGG. Desde 1995 têm sido desenvolvidos métodos bioinformáticos para descobrir comportamentos sistêmicos de informações bioquímicas e/ou genéticas. Os resultados são armazenados nesse banco de dados que possibilita pesquisa básica e aplicada das vias descobertas, como também interações com drogas (AOKI; KANEHISA, 2005).

As redes de interação proteína-proteína em larga escala têm sido avaliadas para diferentes modelos de organismos. Os primeiros estudos foram realizados por Sanchez et al. (1999) em *Drosophila melanogaster*, por Ito et al. (2001) e Ge et al. (2001) em *S. cerevisiae*, e em 2005 foi publicado por Rual et al. o interatoma humano. Além disso, interatomas específicos como o de desordens neurodegenerativas (LIMVIPHUVADH et al., 2007) e de esquizofrenia (CAMARGO et al., 2008) têm sido revelados.

Na busca por domínios e famílias gerais o banco de dados mais utilizado é o Pfam. Nele, cada família é manualmente refinada e representada por dois alinhamentos múltiplos de sequência, dois perfis HMMs e um arquivo de anotação. Outro recurso muito utilizado, mas baseado na similaridade da sequência ou estrutura é o SCOP (*STRUCTURAL CLASSIFICATION OF PROTEINS*). Se a intenção é a busca por domínios específicos existem bancos de dados disponíveis para os diferentes tipos como, por exemplo, para *motif* ligante de calmodulina, Calmodulin Target Database.

Por outro lado, o Interpro (MULDER et al., 2005) realiza buscas contra diferentes bancos de dados de domínios e famílias de proteínas, integrando os serviços oferecidos pelo Pfam, Uniprot, PROSITE, SMAR, PANTHER, PIRSF, SUPERFAMILY PRINTS, ProDom, GENE 3D e TIGRFAMs. Este banco de dados combina os diferentes métodos de reconhecimento de proteínas e na ausência da caracterização bioquímica, a predição de domínios pode ser um bom guia em direção à sua função (QUEVILLON et al., 2005).

O Interpare, da mesma forma que o Interpro, é um banco de dados para busca de domínios em conjunto com o PDB, SCOP, Uniprot e Swiss-Prot. Contudo, este banco de dados também utiliza um método computacional para identificar sítios de interação e moléculas ligantes, e classifica as proteínas pelos alvos de interação com drogas.

Na construção das redes interatômicas, os *softwares* mais utilizados são: **String**, **Cytoscape**, **Osprey** e HiMAP. A escolha depende do organismo estudado, como também do banco de dados no qual está depositada a sequência estudada, se NCBI ou Swiss-Prot. Além disso, é importante que as redes interatômicas sejam feitas por diferentes *softwares* e depois comparadas, para confiabilizar os dados finais. O objetivo final do interatoma é unir as informações do genoma, proteoma e metaboloma, gerando informações que auxiliam no entendimento de funções e ações direcionadas a fármacos e moléculas biologicamente ativas.

Metabolômica

O termo metabolômica foi criado e tem sido usado na última década para abranger o estudo do metabolismo sob perturbações ambientais e genéticas. No entanto, os primeiros trabalhos que envolviam técnicas relacionadas a metabólitos foram publicados há mais de 30 anos, a fim diagnósticos médicos (HORNING; HORNING, 1971).

Os resultados em metabolômica são geralmente ricos em dados, sendo necessário o uso de ferramentas estatísticas e de bioinformática para avaliação e sistematização dos dados, em que propriedades bioquímicas e relações celulares podem ser mapeadas em plataformas de *software* que podem reforçar a interpretabilidade dos dados como, por exemplo, o SetupX que organiza e armazena os resultados de várias pesquisas em metabolômica.

Em 2004, uma série de relatos destacaram a importância de se fornecer informações. Entre elas, a base de dados ArMet, que descreve a arquitetura geral para metabolômica (JENKINS et al., 2004) e MIAMet, que demonstra considerações sobre o mínimo de informações de um experimento em metabolômica (Bino et al., 2004). Estas considerações têm sido concretizadas apenas parcialmente em bases de dados disponíveis sobre metabolômica de plantas (KOPKA et al., 2005). Para uma série de compostos vegetais, várias empresas de agro-biotecnologia têm publicado dados dos metabólitos referentes ao valor nutricional das culturas. O mais abrangente é o CAS (*CHEMICAL ABSTRACTS*), que inclui informações sobre milhões de compostos, entre eles, metabólitos biogênicos. No entanto, este serviço vem com elevados encargos e não contém *links* para bases de dados genômicos.

Farmacogenômica

Um objetivo nos estágios iniciais do desenvolvimento de fármacos é a identificação de um ou mais compostos bioativos. Um composto bioativo é qualquer substância que apresenta a atividade biológica que se procura (BUCHWALD; BODOR, 1998). Qualquer composto com atividade farmacológica ou compostos similares normalmente possuem atividades parecidas, mas variam em sua potência e especificidade. Baseados em um composto bioativo, os cientistas investigam um grande número de moléculas parecidas de forma a otimizar as propriedades farmacológicas desejadas.

Para uma busca sistemática, seria muito importante o entendimento de como as variações nas características estruturais e físico-químicas da família de moléculas estão relacionadas com suas propriedades farmacológicas. O problema é que existem muitos descritores diferentes para caracterizar as moléculas. Eles incluem características estruturais, como a natureza e distribuição dos substituintes; características experimentais, como solubilidade em solventes aquosos e orgânicos, ou momentos dipolo; e características calculadas computacionalmente, como cargas parciais dos átomos. Estes fatores sejam eles de caráter eletrônico, hidrofóbico ou estérico, influenciam na interação do fármaco com a biofase, e na sua distribuição nos compartimentos que compõem o sistema biológico.

Assim, dois fármacos com estruturas químicas semelhantes, diferenciando-se apenas por um átomo ou posição que este ocupa na molécula, podem apresentar diferenças quanto às suas propriedades físico-químicas e, conseqüentemente, quanto à atividade biológica, tanto do ponto de vista quantitativo como qualitativo (ESTRADA, 2008). Os bancos de dados mais utilizados na análise da interação de fármacos ou compostos ativos e outra molécula biologicamente ativa são: KEGG, Drug DataBase e PubChem.

A farmacogenômica surgiu em 1995, da união da farmacogenética com a genômica e a biotecnologia (NEBERT; VESELL, 2004), sendo definida como o estudo da expressão de genes individuais relevantes na susceptibilidade a doenças, bem como resposta a fármacos em níveis celular, tecidual, individual ou populacional (PIRAZZOLI; RECCHIA, 2004). Como muitos outros ramos das ciências biomédicas, foi impulsionada pelos avanços da genômica, que conduziram às expectativas de que a segurança e a eficácia dos medicamentos seriam melhoradas pela personalização da terapêutica, com base nos dados genéticos (FONTANA et al., 2006).

Para o seu estudo, a farmacogenômica utiliza técnicas genômicas, como o sequenciamento de DNA, mapeamento genético e a bioinformática para facilitar as pesquisas na identificação das bases genéticas da variação inter-individual e inter-racial na eficácia, metabolismo e transporte com fármacos (MANCINELLI et al., 2000). A genômica combinada com as ferramentas da bioinformática permite dissecar as bases genéticas das doenças multifatoriais e têm mostrado pontos mais convenientes para melhor ação medicamentosa, aumentando o número de opções moleculares para o tratamento de doenças (DREWS, 2000).

Biotecnologia

A biotecnologia é o uso de conhecimentos sobre os processos biológicos e sobre as propriedades dos seres vivos, com o fim de resolver problemas e criar produtos de utilidade (ANTUNES et al., 2006). Esse processo surgiu da necessidade de se suprir as transformações globais que ocorreram na ciência e no mercado. Desta forma, a biotecnologia está intimamente relacionada à inovação tecnológica, uma vez que propõe o desenvolvimento de novas tecnologias e produtos, aplicando as informações desenvolvidas na pesquisa. Neste contexto, dentro da biotecnologia estão incluídas as pesquisas sobre transgênicos, genômica, proteômica, terapia gênica, entre outras, sendo que para todas essas áreas a bioinformática vem se tornando uma das ferramentas mais utilizadas.

A bioinformática consiste na análise em bancos de dados e utilizando *softwares* visam dar novos rumos à pesquisa, analisando dados e

simulando experimentos. Essa tecnologia propõe novas formas de ciência baseada na experimentação *in silico*, onde podemos prever estruturas de proteínas e moléculas, realizar testes de interação, inibição ou excitação de moléculas, criar inibidores, moléculas de interferência, entre outras atividades. Porém, é fundamental que sejam desenvolvidas pesquisas para alimentar esses bancos de dados, assim como organizá-los em uma linguagem universal de forma a facilitar o *text mining* e *data mining*. Desta forma, o desenvolvimento da bioinformática está relacionado à biotecnologia a partir do momento que geramos novos dados e conhecimentos que podem ser aplicados para o desenvolvimento de novos produtos e soluções.

Atualmente no Brasil existem 39 empresas e entidades cadastradas no site da SOCIEDADE BRASILEIRA DE BIOTECNOLOGIA que atuam na área de Biotecnologia. Além disso, temos 53 grupos de pesquisa biotecnológica na área de biologia animal, 16 em biologia humana e 50 em biologia vegetal. Em relação ao ensino, são 18 cursos de graduação e 12 de pós-graduação com ênfase ou em biotecnologia em Universidades federais, estaduais e faculdades particulares (http//:www.sbb.br, acessado em 24/04/2008).

Para o desenvolvimento da Biotecnologia e, consequentemente, de todas as tecnologias no Brasil é necessário que o governo, a universidade e as empresas percebam esse processo como um sistema multisetorial tecnológico de inovação que abrange diversos setores econômicos (ANTUNES et al., 2006).

Um reflexo disso está na análise dos números de patentes no Brasil que vêm crescendo nos últimos anos, mostrando um retrato dos avanços tecnológicos e do domínio de tecnologias que os centros de pesquisa vêm alcançando. De 2005 até março de 2007, foram realizados 550 depósitos de patentes no Brasil, sendo que destes os principais depositantes são empresas norte-americanas e européias, e apenas 4 entidades brasileiras apresentam um desempenho considerável na área do meio ambiente. Talvez o grande problema não seja o baixo avanço tecnológico, mas a falta de agilidade dos julgamentos dos processos de patente no INPI.

Desta forma, é imprescindível conhecer as tecnologias mais avançadas e capacitar profissionais para o domínio da bioinformática, uma vez que existe uma tendência da evolução da economia global baseada na biotecnologia. As decisões sobre a participação nesse mercado dependem das ações que estão sendo desenvolvidas no presente, sendo que a interação universidade, empresa e governo é a base para garantir ao Brasil essa gestão.

AGRADECIMENTOS

Este trabalho é resultado do aprendizado obtido na disciplina Bioinformática oferecida pelo Prof. Dr. Foued Salmen Espindola e realizada de março a maio de 2008 no Curso de Pós-graduação em Genética e Bioquímica, da Universidade Federal de Uberlândia, Uberlândia/MG, e desta forma agradecemos às Instituições e Agências de fomento que apóiam cada um de nós, como UFU, CAPES, CNPq e FAPEMIG.

ABSTRACT: The omic sciences had a wide point of view of the biological systems, integrating different knowledgement areas, as biochemistry, genetics and physiology, with the aim of isolation and characterization of genes, proteins and metabolites as well study their interactions, based on experimental techniques, softwares and data banks. Bioinformatics proposes a new science, which is based on in silico experimentation, being very dynamic in its update and also can provides the basis for generation of new data and knowledge that can be applied in basic research and applied to the development of new products and solutions. This process is closely related to technological innovation, which is achieved joining biotechnology and bioinformatics. However, the objective of this review is to present a small approach of bioinformatics resources applied to the omics science, like genomics, transcriptomics, proteomics, interatomics, metabolomics, pharmacogenomics, among others.

KEYWORDS: Omics. Bioinformatics. Biotechnology. Data base.

REFERÊNCIAS

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of Molecular Biology**, San Diego, v. 215, p. 403-410, 1990.

ANTUNES, A.; PEREIRA JR, N.; EBOLE, M. F. **Gestão em biotecnologia**, 1. ed., Rio de Janeiro: E-papers, 2006. 324p.

AOKI, K. F.; KANEHISA, M. Using the KEGG database resource. **Current Protocols in Bioinformatics**, Somerset, v. 1, p. 1-12, 2005.

BADER, G. D.; HOGUE, C. W. V. Analyzing yeast protein–protein interaction data obtained from different sources. **Nature Biotechnology**, New York, v. 20, p. 991–997, 2002.

BAIROCH, A. et al. The universal protein resource (UniProt). **Nucleic Acids Research**, Oxford, v. 33, p. 154–159, 2005.

BENSON, D. A. et al. GenBank. **Nucleic Acids Research**, Oxford, v. 33, p. 34–38, 2005.

BERMAN, H. M. et al. The protein data bank. **Nucleic Acids Research**, Oxford, v. 28, p. 235–242, 2000.

BINO, R. J. et al. Potential of metabolomics as a functional genomics tool. **Trends in Plant Science**, London, v. 9, p. 418–425, 2004.

BOECKMANN, B. et al. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. **Nucleic Acids Research**, Oxford, v. 31, p. 365–370, 2003.

BORODOVSKY, M.; MCININCH, J. GeneMark: parallel gene recognition for both DNA strands. **Computers and Chemistry**, London, v. 17, p. 123-133, 1993.

BUCHWALD, P.; BODOR, N. Proteins: structure and function. **Genetics**, Bethesda, v. 30, p. 86-88, 1998.

CAMARGO, L. M.; WANG, Q.; BRANDON, N. J. What can we learn from the disrupted in schizophrenia 1 interactome: lessons for target identification and disease biology? **Novartis Foundation Symposium**, London, v. 289, p. 208-216, 2008.

CAMON, E. et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. **Nucleic Acids Research**, Oxford, v. 32, p. 262–266, 2004.

DELCHER, A. L. et al. Improved microbial gene identification with GLIMMER. **Nucleic Acids Research**, Oxford, v. 27, p. 4636-4641, 1999.

DREWS, J. Drug discovery: a historical perspective. **Science**, Washington, v. 287, p. 1960-1964, 2000.

ESTRADA, E. Quantum-chemical foundations of the topological substructural molecular design. **The Journal of Physical Chemistry A**, Washington, v. 10, p. 1021-1027, 2008.

FAYYAD, U. M. Data Mining and knowledge discovery: making sense out of data. **IEEE Expert: Intelligent Systems and Their Applications**, Washington, v. 11, p. 20-25, 1996.

FONTANA, V. et al. O conceito de gene está em crise. A farmacogenética e a farmacogenômica também? **Revista Biotemas**, Florianópolis, v. 19, p. 87-96, 2006.

FORSLUND, K. et al. Domain tree-based analysis of protein architecture evolution. **Molecular Biology and Evolution**, Cary, v. 25, p. 254–264, 2008.

GE, H. et al. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. **Nature Genetics**, New York, v. 29, p. 482-486, 2001.

GORODKIN, J.; STRICKLIN, S. L.; STORMO, G. D. Discovering common stem-loop motifs in unaligned RNA sequences. **Nucleic Acids Research**, Oxford, v. 29, p. 2135-2144, 2001.

HOOGLAND, C. et al. SWISS-2DPAGE, ten years later. **Proteomics**, Weinheim, v. 4, p. 2352-2356, 2004.

HORNING, E. C.; HORNING, M. G. Human metabolic profiles obtained by GC and GC/MS. **Journal of Chromatographic Science**, Niles, v. 9, p. 129–140, 1971.

HUBBARD, T. et al. Ensembl 2005. **Nucleic Acids Research**, Oxford, v. 33; p.447–453, 2005.

HULO, N. et al. The 20 years of PROSITE. **Nucleic Acids Research**, Oxford, v. 36, p. 245-249, 2008.

ITO, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. **Proceedings of the National Academy of Sciences,** Washington, v. 98, p. 4569-4574, 2001.

JENKINS, H. et al. A proposed framework for the description of plant metabolomics experiments and their results. **Nature Biotechnology**, New York, v. 22, p. 1601–1605, 2004.

KANZ, C. et al. The EMBL nucleotide sequence database. **Nucleic Acids Research**, Oxford, v. 33, p. 29–33, 2005.

KEMMEREN, P. et al. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. **Molecular Cell**, St. Louis, v. 9, p.1133–1143, 2002.

KERSEY, P. J. et al. Integr8 and genome reviews: integrated views of complete genomes and proteomes. **Nucleic Acids Research**, Oxford, v. 33, p.297–302, 2005.

KITANO, H. Systems biology: a brief overview. **Science**, Washington, v. 295, p. 1662-1664, 2002.

KLEIN, R. J.; EDDY, S. R. RSEARCH: Finding homologs of single structured RNA sequences. **BMC Bioinformatics**, London, v. 4, p. 44, 2003.

KOPKA, J. et al. GMD@CSB.DB: the Golm metabolome database. **Bioinformatics**, Oxford, v. 21, p.1635–1638, 2005.

KRALLINGER, M.; VALENCIA, A. Text-mining and information-retrieval services for molecular biology. **Genome Biology**, London, v. 6, p. 224, 2005.

LIMVIPHUVADH, V. et al. The commonality of protein interaction networks determined in neurodegenerative disorders (NDDs). **Bioinformatics**, Oxford, v. 23, p. 2129-2138, 2007.

MANCINELLI, L.; CRONIN, M.; SADÉE, W. Pharmacogenomics: the promise of personalized medicine. **American Association of Pharmaceutical Scientists**, Arlington, v. 2, p. E4, 2000.

MATHEWS, D. H.; TURNER, D. H. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. **Journal of Molecular Biology**, San Diego, v. 317, p. 191-203, 2002.

MATTICK, J. S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. **BioEssays**, Hoboken, v. 25, p. 930-939, 2003.

MATTICK, J. S. Non-coding RNAs: the architects of eukaryotic complexity. **EMBO Reports**, Heidelberg, v. 2, p. 986-991, 2001.

MATTICK, J. S.; GARDEN, M. J. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. **Molecular Biology and Evolution,** Oxford, v. 18, p.1611-1630, 2001.

MEYER, I. M. A practical guide to the art of RNA gene prediction. **Brief in Bioinformatics**, Oxford, v. 8, p. 396-414, 2007.

MULDER, N. J. et al. InterPro: progress and status in 2005. **Nucleic Acids Research**, Oxford, v. 33, p. 201–205, 2005.

NEBERT, D. W.; VESELL, E. S. Advances in pharmacogenomics and individualized drug therapy: exciting challenges that lie ahead. **European Journal Pharmacology**, Amsterdam, v. 500, p. 267-280, 2004.

PIRAZZOLI, A.; RECCHIA, G. Pharmacogenetics and pharmacogenomics: are they still promising? **Pharmacology Research**, Maryland Heights, v. 49, p. 357-361, 2004.

PROSDOCIMI, F. et al. Bioinformática: manual do usuário. **Biotecnologia Ciência e Desenvolvimento**, Brasília, v. 29, p. 12-25, 2003.

PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. **Nucleic Acids Research**, Oxford, v. 33, p. 501–504, 2005.

QUEVILLON, E. et al. InterProScan: protein domains identifier. **Nucleic Acids Research**, Oxford, v. 33, p. 116–120, 2005.

RATTEI, T. et al. SIMAP-- Structuring the network of protein similarities. **Nucleic Acids Research**, Oxford, v. 36, p. 289-292, 2008.

RUAL, J. F. Towards a proteome-scale map of the human protein-protein interaction network. **Nature**, London, v. 437, p. 1173-1178, 2005.

SANCHEZ, C. et al. Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. **Nucleic Acids Research**, Oxford, v. 27, p. 89-94, 1999.

SCHENA, M. et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science**, Washington, v. 270, p. 467-470, 1995.

SCHENA, M. et al. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. **Proceedings of the National Academy of Sciences,** Washington, v. 93, p. 10614-10619, 1996.

SCHUSTER, P. et al. From sequences to shapes and back: a case study in RNA secondary structures. **Proceedings of the National Academy of Sciences,** Washington, v. 255, p. 279-284, 1994.

SHALON, D.; SMITH, S. J.; BROWN, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. **Genome Research**, New York, v. 6, p. 639-645, 1996.

SHARAN, R.; IDEKER, T. Modeling cellular machinery through biological network comparison. **Nature Biotechnology**, New York, v. 24, p. 427-433, 2006.

TATENO, Y. et al. DDBJ in collaboration with mass-sequencing teams on annotation. **Nucleic Acids Research**, Oxford, v. 33, p. 25–28, 2005.

THE HONEYBEE GENOME SEQUENCING CONSORTIUM. Insights into social insects from the genome of the honeybee *Apis mellifera*. **Nature**, London, v. 443, p. 931–949, 2006.

THOMPSON, J. D. et al. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. **Nucleic Acids Research**, Oxford, v. 25, p. 4876-4682, 1997.

VENTER, J. C. et al. The sequence of the human genome. **Science**, Washington, v. 291, p. 1304-1351, 2001.

VETTORE, A. L. et al. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. **Genome Research**, New York, v. 13, p. 2725–2735, 2003.

WAIN, H. M. et al. Genew: The Human Gene Nomenclature Database, 2004 updates. **Nucleic Acids Research**, Oxford, v. 30, p. 169–171, 2002.

WINGENDER, E. et al. Integrative content-driven concepts for bioinformatics “beyond the cell”. **Journal of Biosciences**, Karnataka, v. 32, p. 169-180, 2007.

WU, C. H. et al. The protein information resource. **Nucleic Acids Research**, Oxford, v. 31, p. 345–347, 2003.

YANDELL, M. D.; MAJOROS, W. H. Genomics and natural language processing. **Nature Reviews Genetics**, London, v. 3, p. 601-610, 2002.

YAO, Z.; WEINBERG, Z.; RUZZO, W. L. CMfinder--a covariance model based RNA motif finding algorithm. **Bioinformatics**, Oxford, v. 22, p. 445-452, 2006.

ZUKER, M. Computer prediction of RNA structure. **Methods in Enzymology**, San Diego, v. 180, p. 262-288, 1989.

ZVI, A. et al. Whole genome identification of *Mycobacterium tuberculosis* vaccine candidates by comprehensive data mining and bioinformatic analyses. **BMC Medical Genomics**, London, v. 28, p. 1-18, 2008.