As ômicas:
Integrando a bioinformação

O papel da bioinformática em expansão

Dr. Eliseu Binneck
Consultor/Pesquisador na área de Bioinformática
Embrapa Soja, Londrina – PR.
binneck@cnpso.embrapa.br

Imagens cedidas pelo autor

Como resultado dos crescentes investimentos na área da genômica nos últimos anos, a lista de sequências de genomas completos vem crescendo a uma velocidade cada vez maior e contribuindo com a disposição de um volume de dados para acesso público sem precedentes na história. Hoje (maio de 2004) são 190 genomas completos publicados, dos quais, 145 de procariotos, 18 de archaea e 27 de eucariotos. Além disso, existem 900 genomas sendo seqüenciados; 460 de procariotos, 26 de archaea e 414 eucariotos (http://www.genomesonline.org). A Figura 1 apresenta a evolução na obtenção de sequências genômicas completas de organismos de vida livre.

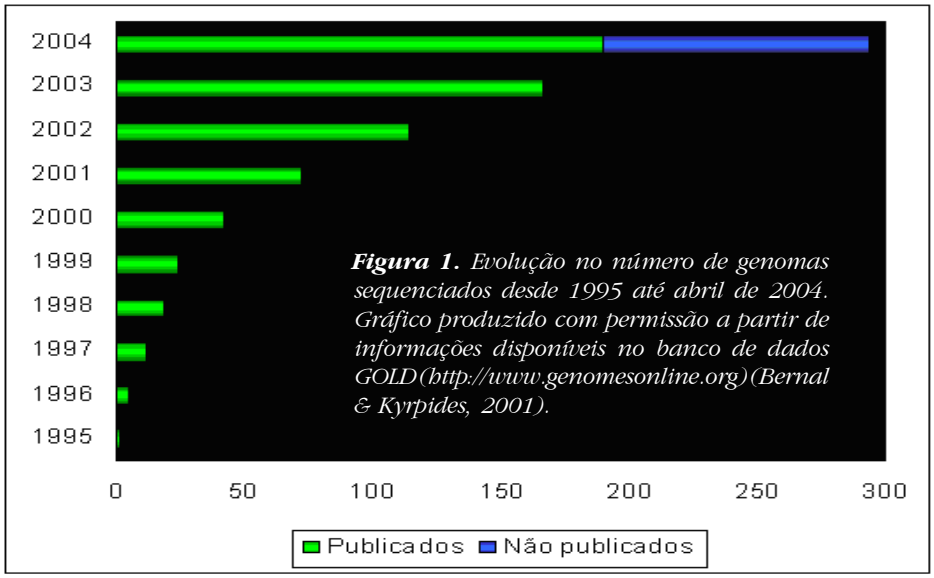

Figura 1. *Evolução no número de genomas sequenciados desde 1995 até abril de 2004. Gráfico produzido com permissão a partir de informações disponíveis no banco de dados GOLD (http://www.genomesonline.org) (Bernal & Kyrpides, 2001).*

Um forte componente que tem auxiliado tremendamente essa evolução da informação genômica são as ferramentas de bioinformática. Atualmente os dados de sequências podem ser explorados com o uso de poderosas ferramentas de busca, acessando fontes de informação eletrônica associada e integrada de um modo inconcebível há menos de uma década, quando, em 1995, foi seqüenciado o primeiro genoma de um organismo de vida livre, *Haemophilus influenzae* (Fleischmann et al, 1995). Muitas dessas ferramentas, como Ensembl Genome Browser (http://www.ensembl.org/) (Stalker et al, 2004), KEGG (http://www.genome.ad.jp/kegg/kegg2.html) (Kanehisa et al, 2004), GeneQuiz (http://www.sander.ebi.ac.uk/gqsrv/submit/) (Hoersch et al, 2000) e MIPS (http://www.mips.biochem.mpg.de/) (Mewes et al, 2004), são de domínio público e possibilitam a obtenção de informações organizadas, além de integrarem ferramentas poderosas, possibilitando, por exemplo, a análise comparativa entre dados de diferentes genomas.

Entretanto, em meio a esse clima de novidade e excitação, parece ter se estabelecido uma expectativa excessiva sobre a aplicação de dados de sequências genômicas em busca de inferências biológicas. Por outro lado, existe um crescente reconhecimento e entendimento de que tais metodologias baseadas na seqüência de DNA terão que ser complementadas pela análise direta dos produtos codificados pelos genes; os RNAs e as proteínas. Sabe-se que conhecer a seqüência de um genoma não garante que as proteínas codificadas por esse genoma possam ser imediatamente determinadas (por exemplo, por homologia com proteínas, já conhecidas, de outros organismos). Há uma estimativa, baseada em genomas recém-completos, que cerca de 30% do conteúdo gênico de um organismo seja de proteínas específicas deste (Rubin et al, 2000). É claro que esse número tende a diminuir à medida que mais e mais genomas vão sendo seqüenciados, mas mostra a dificuldade em proceder-se a uma anotação automatizada [confiável e completa] dos genomas.

As predições computacionais a partir de dados de seqüências são complicadas e nem sempre geram resultados confiáveis, principalmente no caso de genomas mais complexos como o genoma humano. Embora o término do Projeto Genoma Humano tenha sido comemorado em abril de 2003

(Collins et al, 2003; Pennisi, 2003a), o número exato de genes codificados pelo genoma é ainda desconhecido e podem ser necessários anos ainda até que tenhamos uma contagem confiável do número de genes no genoma humano.

A razão para tanta incerteza é que as predições são derivadas a partir de diferentes métodos computacionais e programas de predição gênica. Alguns programas detectam genes procurando por parâmetros diferentes que definem onde um gene começa e termina (predição "*ab initio*"). Outros programas procuram por genes pela comparação de segmentos de sequência com homologia com genes e proteínas conhecidos (predição comparativa). Enquanto a predição *ab initio* tende a sobrestimar o número de genes pela contagem de qualquer segmento que pareça um gene, o método de predição comparativa tende a subestimar este número, já que é limitado por reconhecer somente os genes similares aos já conhecidos. A definição de gene é problemática porque pequenos genes podem ser difíceis de detectar, um gene pode codificar para vários produtos protéicos, alguns genes codificam para RNA, dois genes podem se sobrepor, e há muitas outras complicações (Pennisi, 2003b). Sendo assim, métodos computacionais por si só não são suficientes para gerar o número real e o conhecimento de todos os genes de um genoma eucariótico complexo; pelo menos com as informações existentes atualmente. Até que se gere um conjunto de dados bastante informativo para as predições comparativas, essas precisarão ser verificadas por trabalho intensivo de laboratório antes de se chegar a um consenso real.As últimas estimativas a partir de programas de predição gênica sugerem que no genoma humano devem existir 24500 ou menos genes que codificam para proteínas (Pennisi, 2003c). A estimativa do Ensembl (versão 20.34c.1, de 08-02-2004) é de 23531 genes, incluindo 1744 pseudogenes (http://www.ensembl.org/Homo_sapiens/) (Stalker, 2004). Essa estimativa é muito menor do que aquelas das anotações iniciais, que contavam mais de 70.000 genes (Write et al, 2001). Considerando que os genes no genoma humano apresentam um tamanho médio de 3000 pares de bases, menos de 2% do genoma codificam para proteínas. Assim mesmo, atualmente é desconhecida a função de mais de 50% dos genes descobertos.

Observando a inesperada equidade relativa no número de genes de organismos bastante diferentes em termos de complexidade (Quadro 1), sugere-se que o fator que determina a complexidade de um organismo não está no número de genes, mas em como as partes gênicas são usadas para construir diferentes produtos em um processo chamado *splicing* alternativo. Outra razão para essa maior complexidade são as milhares de modificações químicas pós traducionais que ocorrem nas proteínas e o repertório de mecanismos de regulação que controlam esses processos (Genomics and Its Impact on Science and Society: The Human Genome Project and Beyond, 2003). A versão 34.00 do banco de dados RESID (http://pir.georgetown.edu/pirwww/dbinfo/resid.html) (Garavelli, 2003) apresenta 339 modificações pós ou co-traducionais conhecidas em proteínas, modificações essas que não podem ser evidenciadas diretamente a partir da seqüência gênica.

Informação versus ação

Em qualquer sistema biológico, se um trabalho é realizado, quase sempre a molécula responsável por essa ação é uma proteína. A vida depende de milhares de proteínas diferentes, cujas estruturas são ajustadas para que moléculas individuais de proteínas combinem, numa precisão impressionante, com outras moléculas. Reações químicas na célula dependem da combinação de enzimas com substratos e essas são geralmente controladas por outras moléculas combinando com sítios específicos da proteína. Estruturas como os músculos dependem da interação proteína-proteína, o controle da expressão gênica depende da combinação proteína-DNA, o controle hormonal depende da interação do hormônio com receptores protéicos, o transporte através da membrana envolve interações soluto-proteína, proteções imunes requerem a interação antígeno-anticorpo, atividades neuronais requerem a interação substância transmissora-proteína. Estes são apenas alguns exemplos do universo quase infindável de interações específicas em que as proteínas são envolvidas. Todas essas interações dependem do reconhecimento exato de estruturas específicas nas moléculas das proteínas envolvidas (Goodsell, 1991). Neste contexto, bancos de dados como o LIGAND (http://www.genome.ad.jp/ligand/) (Goto, 2002) possibilitam visualizar cada uma entre o universo de reações químicas conhecidas envolvendo a interação de enzimas com metabólitos e outros compostos. Interações proteína-proteína, proteína-DNA e proteína-RNA podem ser encontradas em bancos de dados como BIND – *Biomolecular Interaction Network database* (http://www.bind.ca/) (Bader et al, 2003), DIP – *Database of Interacting Proteins* (http://dip.doe-

Quadro 1 – Tamanho do genoma e número estimado de genes de diferentes organismos.

Organismo	Tamanho do Genoma (pares de bases)	N° Estimado de Genes
Homem (*Homo sapiens*)	3 bilhões	30.000
Rato (*M. musculus*)	2,6 bilhões	30.000
Mostarda (*A. thaliana*)	100 milhões	25.000
Roundworm (*C. elegans*)	97 milhões	19.000
Mosca das frutas (*D. melanogaster*)	137 milhões	13.000
Levedura (*S. cerevisiae*)	12,1 milhões	6.000
Bactéria (*E. coli*)	4,6 milhões	3.200
Virus da AIDS (HIV)	9700	9

mbi.ucla.edu/) (Salwinski et al, 2004) e MINT – *Molecular INTeractions* (http://cbm.bio.uniroma2.it/mint/) (Zanzoni, et al, 2002). Informações sobre interação antígeno-anticorpo são disponíveis no IMGT – International Immunogenetics Database (imgt.cines.fr) (Lefranc, 2004).

Cada vez mais se torna evidente que a aplicação de dados de seqüências de DNA, utilizando informações sobre a relação entre a seqüência de DNA do gene e a função protéica, não sustenta a atribuição infalível de função para as proteínas. Muitas evidências mostram a fragilidade das constatações feitas puramente a partir de seqüências genômicas, sugerindo que (i) embora a seqüência genômica possa ser usada para predizer "open reading frames" (ORFs), tais predições são ainda muito grosseiras e passíveis de erro, principalmente em eucariotos. (ii) O processamento do mRNA tem uma influência importante no produto final da expressão gênica; o proteoma. É o caso do *splicing* alternativo, em que, pela montagem de diferentes combinações de exons, um pré-mRNA dá origem a dois ou mais mRNAs diferentes, que codificam para produtos protéicos diferentes. Como resultado, as modificações advindas do processamento do mRNA permitem que seja produzida uma variedade de proteínas superior ao número de genes do genoma. (iii) Existe uma enorme diversidade de modificações pós-traducionais que uma proteína pode sofrer, influenciando a sua função, localização celular e atividade. A informação da seqüência de DNA ainda não dá um discernimento claro sobre modificações pós-traducionais a que cada produto protéico está sujeito, sendo difícil, se não impossível, estabelecer um número de proteínas produtos que cada gene codifica. (iv) Os mecanismos de controle da expressão gênica envolvem uma rede complexa e variável de interações moleculares, cujo entendimento é ainda bastante rudimentar. Esses mecanismos não são prontamente evidentes a partir do conhecimento da seqüência de DNA do genoma, havendo ainda grandes limitações em se utilizar a informação da seqüência de DNA com o intuito de conhecer o conteúdo e a dimanicidade das proteínas codificadas por um determinado genoma.

Fotografia versus filme

Certos grupos de proteínas interagem entre si para realizar determinados trabalhos celulares. Um exemplo bem típico são as proteínas organizadas em vias metabólicas como a glicólise, o ciclo de Krebs, e outras, em que os produtos gênicos chamados enzimas precisam trabalhar em harmonia. Outro exemplo bem conhecido é o caso das proteínas estruturais que devem estar juntas e organizadas precisamente para exercer a sua função, como exemplo, os componentes de uma unidade ribossomal, as histoproteínas que são essenciais para manter a estrutura da cromatina etc. Desse modo, em estudos de expressão gênica é habitual assumir que grupos de genes cujos modelos de expressão são similares entre si, sejam provavelmente funcionalmente relacionados.

Um problema com as técnicas de agrupamento de dados de expressão gênica (ESTs, SAGE, Microarrays), no entanto, é que elas são baseadas na suposição de que os genes que apresentam modelos de expressão similares são de fato relacionados funcionalmente, isto é, eles têm funções que são relacionadas. Essa interpretação geralmente leva a erros na tentativa de entender a relação real entre os genes [através dos seus produtos].

Existem razões para pôr em dúvida essa suposição: primeiro, ainda é muito inconsistente o conhecimento de quão discretamente trabalham os grupamentos funcionais de genes na maquinaria celular. Pode ser que produtos gênicos individuais tenham tantos papéis diferentes em diferentes circunstâncias, que vários deles participem de papéis essenciais em mais de uma função. Por exemplo, os processos de defesa contra estresses bióticos (originados do ataque de agentes patogênicos), ou estresses ambientais, podem ser extremamente complexos e envolverem diferentes mecanismos atuando em conjunto. Segundo, o termo "relacionados funcionalmente" é por si só mal especificado. Se o modelo de expressão de um gene é similar ao de um outro gene, isso pode significar vários tipos de relacionamento, desde "dois genes tendo produtos que interagem fisicamente", "um gene que codifica para um fator de transcrição para outro gene", "dois genes ambos com seqüências promotoras ligadas por repressores que são liberados quando um receptor nuclear é ativado, mesmo que os dois genes tenham funções muito distantes". É claro que existe um nível de abstração no qual todos os genes são funcionalmente relacionados no trabalho de manter a célula viva e produzindo todos os componentes necessários para o organismo como um todo. Mas abaixo desse nível de abstração existem muitos alternativos, pela sua natureza, favorecendo a definição de agrupamento. Portanto, é perfeitamente questionável a atribuição indistinta de que similaridade em expressão corresponde à similaridade em função.

Além disso, o que constitui realmente um modelo de expressão similar é ainda pouco preciso, ou pelo menos existem múltiplas definições alternativas. Por exemplo, similaridade poderia significar ter um modelo de mudança similar ao longo do tempo. Pode significar também níveis absolutos de expressão a qualquer dado momento, ou pode significar a perfeita oposição, mas bem coreografada no modelo de expressão. Pensando em métodos comparativos, qual medida de discrepância exatamente escolhida para medir os modelos de expressão influenciará o tipo de agrupamento funcional esperado. Métodos confiáveis e exeqüíveis em escala genômica para medição absoluta da expressão gênica precisam ainda ser desenvolvidos.

Corretamente interpretados ou não, dados de expressão gênica vêm sendo acumulados em volume e variedade cada vez maior. Um ensaio isolado de hibridação com *DNA Microarrays*, por exemplo, fornece na melhor das hipóteses uma visão estática do nível de expressão comparativo entre os genes amostrados. Seria como a fotografia do evento. Mas dificilmente uma fotografia consegue mostrar todo o panorama. Uma nova fotografia, tomada de um outro ângulo,

pode mostrar nuances que não haviam sido captadas anteriormente, e assim por diante. Conhecer as mudanças é diferente de percorrer o caminho que leva aos estados diferenciados. Por exemplo, entender a trajetória da ocorrência de vários RNAs mensageiros em vez de conhecer apenas valores absolutos ou comparativos em um dado momento, proporciona muito mais informação sobre a operacionalidade do sistema.

A vida é essencialmente dinâmica. Apenas o filme, isto é, a análise dinâmica do sistema, pode dar suporte para o entendimento completo dos processos biológicos. E aí está o grande desafio da bioinformática. A integração comparativa dos dados precisa ser realizada *in silico*, transformando o conjunto de imagens estáticas no filme da vida.

As ômicas

Antes da era da bioinformática, somente duas maneiras de fazer experimentação em biologia eram disponíveis: utilizando um organismo vivo (também chamado *in vivo*) ou em um sistema artificial (também chamado *in vitro*). Seguindo essa analogia, podemos dizer que a bioinformática é de fato a biologia *in silico*. A bioinformática veio para facilitar o uso de computadores no sentido de organizar e analisar integradamente uma montanha de dados complexos e variados, possibilitando enfrentar o desafio de decifrar componentes importantes dentro de um universo crescente de informações. Isso somado ao desenvolvimento de equipamentos poderosos para a miniaturização e automação da aquisição de dados biológicos em larga escala, deu campo para o surgimento de uma lista de novos termos, que não pára de crescer. Estamos entrando na era das ômicas (Palsson,2002). Com centenas de milhares de proteínas para identificar, correlacionar e entender, por exemplo, não é suficiente estudar um gene, um produto gênico ou um processo de cada vez. Por outro lado, estudar em larga escala um conjunto de moléculas com o objetivo de entender mecanismos celulares, dificilmente podem responder questões interessantes sem a assistência da informação gerada pela pesquisa tradicional dirigida por hipóteses. Por isso, os dois tipos de ciência atualmente disponíveis, as ômicas e as pesquisas dirigidas por hipóteses (Weinstein, 2001), são sinérgicas e devem ser utilizadas de modo a se complementarem.

Genômica

A genômica se caracteriza pelo estudo dos genes e suas funções. A sua chegada, com o projeto genoma humano no final da década de 1980, alavancou toda a revolução atual no campo da biologia. Muitas expectativas e investimentos têm sido empregadas na genômica, visando aplicações nas áreas da indústria farmacêutica, agricultura, produção de energia e proteção do meio ambiente. Mas a determinação da seqüência completa de vários genomas não é o final da história. É apenas o começo, principalmente pelo fato de que mecanismos biológicos não podem ser inferidos simplesmente a partir do conhecimento da seqüência sem o auxílio de outras estratégias de estudo, as ômicas em geral.

Genômica comparativa. Esse novo ramo da genômica, que vem se tornando cada vez mais comum dada a quantidade de seqüências de genomas sendo produzidas, tem o objetivo de comparar todo o conteúdo de DNA do genoma de um organismo particular com outros genomas já conhecidos. Através dessa análise pode ser possível identificar diferenças, tanto no conteúdo gênico quanto não-gênico, que podem ser responsáveis por importantes propriedades fenotípicas ou evolutivas, como patogenicidade, reações a condições ambientais adversas, proximidade taxonômica entre grupos e até mesmo a aquisição (ou manifestação?) de determinados comportamentos individuais.

Transcriptômica (ou genômica funcional)

O produto inicial da expressão gênica em um organismo é conhecido como transcriptoma e se caracteriza por uma coleção de moléculas de RNA mensageiro cuja informação biológica é requerida pela célula em um determinado momento. Essas moléculas de mRNA são sintetizadas a partir de genes que codificam proteínas e, assim, direcionam a síntese do produto final da expressão gênica, o proteoma, que especifica a natureza das reações bioquímicas que a célula está apta a realizar. Um ponto importante a notar é que o transcriptoma nunca é sintetizado *de novo*, isto é, não começa do zero. Cada célula recebe parte de seu transcriptoma materno quando é formada pela divisão celular, e depois é responsável pela manutenção e adaptação do transcriptoma conforme os diferentes estágios de sua vida e o tipo de diferenciação tomado.

Como regra geral, RNAs mensageiros bacterianos têm meias-vidas de não mais de poucos minutos e em eucariotos a maioria dos mRNAs são degradados poucas horas após a sua síntese. O "turnover" rápido significa que a composição do transcriptoma não é fixa e pode ser rapidamente reestruturada pela mudança no nível de síntese de mRNAs específicos. Assim, a transcrição não resulta na síntese do transcriptoma, mas apenas o mantém pela reposição de mRNAs que foram degradados, e promove mudanças na composição do transcriptoma ligando ou desligando os diferentes genes ou conjuntos de genes.

Avanços tecnológicos baseados na PCR, intenso sequenciamento de cDNA e síntese *de novo* de ácidos nucléicos, têm contribuído para o desenvolvimento de técnicas de quantificação de mRNA em larga escala, em muitos casos em escala genômica, possibilitando que centenas ou milhares de genes sejam estudados em paralelo em vez de um gene de cada vez. Métodos como *Differential Display* (DD), *Serial Analysis of Gene Expression* (SAGE) e *DNA array hibridization* ou *DNA microarray*, todos trouxeram benefícios significativos em relação ao *Northern blotting* em termos de sensibilidade e número de ensaios. Entre essas tecnologias, a que vem ganhando preferência para estudar a composição de um transcriptoma, e fazer comparações entre diferentes transcriptomas, é a técnica de *DNA microarray*,

que se baseia na hibridação em paralelo de ácidos nucléicos. Experimentos de expressão gênica com *DNA microarrays* vêm sendo largamente utilizados para explorar o modelo de expressão simultânea e em paralelo de milhares de genes. Isso requer ferramentas poderosas de correlação computacional.

Um *DNA microarray* consiste de uma coleção de sequências parciais de genes (normalmente cDNAs) que são espotados individualmente em locais específicos de uma lâmina. Essas sequências geralmente variam de 500 a 4000 bases (idealmente 500 a 2000 bases) e podem ser escolhidas a partir de diferentes regiões do gene dependendo do objetivo do projeto. Uma variação da técnica, chamada DNA chip, é baseada na deposição ou síntese *in situ* de oligonucleotídeos para a geração de alvos. Esses chips contêm oligômeros curtos variando de 25 a 80 bases como seqüências-alvo. Enquanto essas sequências curtas podem conferir alta sensibilidade, elas podem apresentar baixa especificidade de ligação comparada com *DNA microarrays*, uma vez que as seqüências são curtas e usualmente não representam genes conhecidos.

O uso de *DNA microarrays* para o estudo do modelo de expressão gênica baseia-se em dois princípios. Primeiro, considera-se que cada gene é expresso ou não e as diferenças no seu nível de expressão em uma célula ou tecido, em determinado momento, são um reflexo de quais mRNAs estão presentes e a sua abundância, e; segundo, as fitas de DNA podem hibridar-se com seqüências complementares formando uma molécula estável em fita dupla.

Tipicamente, a primeira face dos dados experimentais de *DNA microarrays* é uma lista de genes/sequências ou números de identificação e o seu perfil de expressão. Modelos de correlação dentro do conjunto massivo de dados de pontos não são óbvios por uma inspeção visual. Diferentes algoritmos de agrupamento computacional precisam ser usados simultaneamente para reduzir a complexidade dos dados e para encurtar a relação entre genes de acordo com o seu nível de expressão ou mudanças nos níveis de expressão. Problemas relacionados com as técnicas de agrupamento são considerados na seção anterior.

Uma das maiores vantagens da utilização da técnica de *DNA microarray*, comparando-a com outros métodos, é a facilidade da análise simultânea e em paralelo de um grande número de genes e de um grande número de amostras. Deve ser notado, entretanto, que todas essas técnicas usadas para a quantificação de mRNA proporcionam um nível de informação empírica e não uma condição estável absoluta. Além disso, sabe-se que a detecção de uma diferença na abundância de um mRNA específico entre duas amostras biológicas não é necessariamente refletida por uma diferença quantitativa equivalente no nível de abundância da proteína, o que muitas vezes está implícito nos estudos.

Existem, portanto, limitações intrínsecas da técnica, entre as quais (i) a abundância do mRNA nem sempre é bem correlacionada com a abundância da proteína, (ii) a sensibilidade e variação dinâmica dos métodos existentes são tais que os mRNAs menos abundantes, potencialmente codificando as proteínas regulatórias mais importantes, não são facilmente medidos como acontece com os mRNAs mais abundantes, e (iii) a atividade das proteínas codificadas pelos mRNAs é regulada a vários níveis após a sua expressão. Por exemplo, a localização subcelular e/ou a extensão em que as proteínas são pós-traducionalmente modificadas, não são reveladas pela medição da abundancia do mRNA.

Proteômica

Para entender a função de todos os genes em um organismo, é necessário conhecer não só quais genes são expressos, quando e onde, mas também quais são os produtos da expressão e em que condições esses produtos (proteínas) são sintetizados em certos tecidos. A proteômica tenta descrever o conjunto completo de proteínas produto da expressão do genoma (James, 1997), e fornece informações importantes para complementar os estudos de transcriptômica e metabolômica.

Os organismos podem sintetizar muitos milhares de proteínas ao mesmo tempo, e a diversidade potencial de tipos de proteínas no proteoma certamente excede o número estimado de genes no genoma. Isso ocorre porque os produtos de um gene podem diferir devido a *splicing* alternativo e uma variedade de modificações pós-traducionais possíveis, como apresentado acima. O crescente interesse no campo da proteômica vem concentrando esforços para acelerar o desenvolvimento e implementação de estratégias mais apropriadas para a análise de expressão e função de proteínas em escala genômica.

Esse interesse tem ocorrido, em parte substancial, devido ao sucesso dos projetos de sequenciamentos genômicos, considerando que a realização bem sucedida desses projetos tem resultado em uma apreciação mais extensa de que, por si só, eles revelam menos do que se esperava sobre a biologia do organismo. Os dados de sequências genômicas proporcionam uma plataforma essencial para um conhecimento mais amplo das estratégias experimentais complementares que darão suporte à caracterização dos genes contidos nos genomas. A utilização integrada dessas ferramentas possibilitará o entendimento de como os produtos desses genes atuam conjuntamente para regular as atividades do organismo.

A proteômica depende da extração, separação, visualização, identificação e quantificação das proteínas presentes em um organismo ou tecido, em um determinado momento. Todos esses estágios têm limitações. Portanto, atualmente, é impossível descrever o proteoma completo de um organismo.

Atualmente, o ponto de partida para muitas tentativas na investigação das mudanças na expressão protéica envolve a resolução das proteínas de uma mistura complexa por eletroforese 2-D e a sua subsequente identificação usando métodos analíticos cada vez mais precisos e poderosos. Eletroforese 2-D, complementada com HPLC, permite

separar e purificar vários milhares de proteínas extraídas de um tecido ou células, em um determinado momento ou condição. Embora a eletroforese 2-D apresente significantes limitações, parece ser o melhor método até o momento para resolver um grande número de proteínas de uma mistura, ao mesmo tempo em que permite acessar as mudanças no nível de expressão e a purificação de proteínas chave para subsequente caracterização.

Avanços relativamente recentes na caracterização de proteínas têm surgido da automatização de métodos como *matrix-assisted laser desortion-ionization* (MALDI) e *eletrospray ionization* (ESI) *mass spectrocopy* (MS) para se obter o fingerprinting de massa e sequenciamento de peptídeos.

Metabolômica

A metabolômica é uma área da genômica funcional que estuda as mudanças na expressão de pequenas moléculas orgânicas, conhecidas como metabólitos, em sistemas biológicos. Ela promete complementar a genômica por permitir avaliações objetivas do fenótipo (Weckwerth, et al, 2004).

Grande importância vem sendo dada para a combinação de dados de metabolômica com dados de expressão gênica e proteômica. A metabolômica ajudará na revelação de como os genótipos são associados com os fenótipos e fazer simulações de mecanismos celulares em larga escala. Em uma escala maior, o **fenomenoma** (Schilling et al, 1999; Palsson, 2000**)** ajudará a materializar métodos de análise com a melhor tecnologia para estudos [e interpretações] do metaboloma.

O fenomenoma requer uma organização de descobertas biológicas, quantificando e identificando todos os metabólitos em um complexo de amostras biológicas, rápida e simultaneamente. Isso deve ser obtido sem qualquer seleção a priori dos metabólitos de interesse, para evitar tendenciosidades. Softwares de bioinformática são necessários para organizar e facilitar a visualização dos dados de modo a auxiliar na sua interpretação (Steuer et al, 2003; Covert et al, 2004). Os softwares devem combinar dados obtidos por *DNA microarrays*, proteômica e metabolômica numa mesma visualização.

Essa tecnologia permitirá, em última instância, a integração e correlação das mudanças globais no metabolismo e expressão gênica. Uma análise quantitativa de todos os metabólitos em uma célula pode ajudar no entendimento de problemas como, por exemplo, os efeitos pleiotrópicos, em que um único gene determina um número de características não relacionadas. Problemas assim podem ser mais bem entendidos se uma alteração detectada no conteúdo de um metabólito, utilizado em vias metabólicas diferentes, estiver relacionado com uma mutação no gene ou a sua sobre-expressão ou inibição.

O Quadro 2 mostra a evolução das principais novas áreas da pesquisa biológica no últimos anos, baseada no número de ocorrências de termos relacionados na literatura científica.

Além dessas, uma variedade de ômicas vem surgindo e uma sobreposição de propósito é inevitável. Entre outras tantas, a **farmacogenômica** (Marshall, 1997) visa entender a interação da constutuição genética de um indivíduo com a resposta a drogas.

A **fisiômica** (Sanford et al, 2002) se dedica a fazer uma descrição quantitativa das funções fisiológicas de um organismo intacto. É necessário predi- zer o fenótipo a partir do genótipo, mas isso é difícil por causa das influências do ambiente e as circunstâncias do crescimento, desenvolvimento e doenças. O objetivo é obter o um discernimento de toda a fisiologia de um organismo, incluindo as vias metabólicas e todas as moléculas e suas interações, que fazem o organismo completo. Uma das primeiras iniciativas nesse campo é o Projeto Fisioma (http://physiome.org/), cujo principal objetivo é entender o organismo humano, descrevendo quantitativamente a sua fisiologia e patofisiologia, utilizando inclusive informações provenientes dos fisiomas de outros organismos, para melhorar a saúde humana (Bassingthwaighte, 2000).

A **regulômica** (Werner, 2004) é o estudo das instruções bioquímicas da rede de interação gênica que controla os mecanismos de regulação da expressão dos genes para fazer todos os tipos de célula necessários para construir organismos completos (Kondro, 2004; Gao et al 2004; Roven & Bussemaker, 2004).

A **peptidômica** se dedica a estudar peptídeos pequenos (0,5 a 15 kDa), como hormônios, citoquinas, fatores de crescimento, venenos, toxinas, peptídeos antimicrobianos etc. Essas moléculas têm papel fundamental em muitos processos biológicos (Schulz-Knappe et al, 2001; Prates & Bloch, 2002).

A **degradômica** é a aplicação de dados gerados pela genômica e proteômica para identificar as proteases

Quadro 2 – Número de ocorrências de referências no PubMed (http://www.ncbi.nlm.nih/) em algumas novas áreas da pesquisa biológica, desde 1998. Busca limitada para os campos Título e Abstract.

Palavra chave	1988	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	Abril2004
"Genomics"	3	12	23	38	52	64	90	130	208	386	678	1263	2081	3104	4199	4660
"Comparative genomics"	—	—	—	—	—	—	4	8	18	37	69	126	192	291	427	503
"Functional genomics"	—	—	—	—	—	—	—	—	10	46	131	277	480	736	1016	1127
"Transcriptomics"	—	—	—	—	—	—	—	—	—	—	1	3	7	23	41	63
"Proteomics"	—	—	—	—	—	—	—	—	1	20	67	277	631	1254	2022	2444
"Pharmacogenomics"	—	—	—	—	—	—	—	—	1	11	37	136	249	472	702	795
"Metabolomics"	—	—	—	—	—	—	—	—	—	—	—	2	7	28	59	81
"Peptidomics"	—	—	—	—	—	—	—	—	—	—	—	—	5	8	18	23
"Bioinformatics"	—	—	—	—	3	12	20	44	78	144	230	420	657	1058	1604	1852

e os seus substratos em escala genômica, para descobrir novos papéis para proteases in vivo. O objetivo é facilitar a identificação de novos alvos para o desenvolvimento de fármacos visando o tratamento de doenças (Lopez-Otin & Overall, 2002).

A **epigenômica** busca esclarecer como o genoma funciona como um todo. Ela combina a genética com o ambiente para buscar uma compreensão dos sistemas biológicos complexos como a plasticidade do genoma. Embora todas as células nucleadas de um organismo levem o mesmo genoma, elas expressam diferentes genes em diferentes momentos e condições. Esses mecanismos de regulação da expressão gênica são complexos, e um dos principais fatores envolvidos são as mudanças epigenéticas resultantes da metilação diferencial do genoma. Daí, diz-se que resultam diferentes epigenomas. Alguns estudos têm demonstrado o envolvimento da metilação do DNA num processo chamado *imprinting* genômico, que controla a expressão de alguns genes em mamíferos, podendo ter efeito no surgimento de doenças, especialmente o câncer. Novik et al (2002) apresenta uma revisão sobre o assunto.

A **toxicogenômica** (Kramer & Kolaja, 2002 e Guerreiro et al, 2003) marca um novo paradigma no desenvolvimento de drogas e análise de risco, que promete gerar uma enorme quantidade de informação na direção de aumentar o entendimento do mecanismo molecular que leva à toxicidade da droga e eficiência. É esperado que a toxigenômica seja mais e mais integrada com todas as fases do processo de desenvolvimento de drogas, particularmente na toxicologia mecanística e preditiva, e descobrimento de biomarcadores, buscando identificar polimorfismos no DNA relacionados com a suscetibilidade individual à toxicidade em relação a uma determinada droga. O objetivo é a seleção de candidatos no sentido de ajudar a desenvolver e utilizar drogas que produzam menor toxicidade.

Antes e depois da genômica: a velha e a nova biologia

Depois do descobrimento da dupla fita de DNA, do código genético, enzimas de restrição, PCR e tantos avanços na biologia molecular durante a segunda metade do século passado, na última década experienciamos uma nova revolução no campo da biologia com a era da genômica, e com ela muitas outras ômicas, como apresentado acima. Nesse contexto, muitas perguntas surgiram e permanecem ainda sem respostas satisfatórias, como: quais os impactos da genômica nos projetos de pesquisa nas diversas áreas das ciências biológicas? o método científico

Figura 2. Ilustração do processo de obtenção de novas descobertas nos diversos campos da ciência.

ainda é relevante? a bioinformática é uma disciplina separada? como pode ser melhorada a comunicação entre as culturas científicas atuais e a tecnologia da informação (IT) para solucionar a necessidade da integração dos dados disponíveis, que apresentam-se em fontes e formatos tão variados? perguntas como essas são chaves para as ações futuras nas biociências.

Fazendo um paralelo entre a velha biologia e a situação atual, podemos notar que o predomínio de pesquisadores mais ou menos independentes e profundamente especializados em um domínio estreitamente focado, não é adequado para a nova ciência cada vez mais integrada e ampla. Os estudos voltados para um gene ou uma função de cada vez dão lugar para a análise quantitativa de centenas de milhares de genes, e não mais focalizando apenas uma espécie, mas com uma abordagem de integração comparativa de dados interespecíficos. Os grandes investimentos voltados para enfoques científicos muitas vezes pouco abrangentes e hipóteses dirigidas pela pesquisa são substituídos pela automação e miniaturização, reduzindo o custo e aumentando a velocidade da coleta de dados. A necessidade da busca de ferramentas computacionais básicas e somente para analisar conjuntos de dados é suplantada pela rápida disponibilidade de bancos de dados, grandes demais para um pesquisador conseguir analisar os dados sozinho. E, assim, onde estão as hipóteses? poderíamos caracterizar essa revolução como uma grande expedição para o acabamento da ciência da vida? quais são os impactos para a sociedade?

Embora se tenha observado uma grande mudança no tipo e quantidade de dados obtidos, e a validade do método científico ser colocado em xeque, o plano clássico no curso da ciência continua sendo válido. Os dados geram informação, que gera novos conhecimentos, que proporcionam o caminho para novas descobertas. No final, algumas vezes, paradigmas são transpostos (Figura 2). A principal diferença é que até algumas décadas atrás, esse processo requeria somente poder de raciocínio, lápis e papel. Agora requer tecnologia computacional sofisticada. Para isso, os centros de pesquisa e universidades cada vez mais terão que ter seus próprios grupos de bioinformática, mantendo equipes multidisciplinares com atividades que de um lado promovam uma melhor exploração dos dados biológicos através de ferramentas de bioinformática e, por outro lado, as questões geradas pelos dados biológicos obtidos possibilitem melhorar as ferramentas de bioinformática. A bioinformática será cada vez mais importante em termos de integração da informação, buscando impulsionar a aquisição de conhecimento sobre os sistemas biológicos para a geração de novas saídas para problemas na agricultura, medicina, produção de energia e conservação do meio ambiente.

O papel da bioinformática em expansão

Os projetos genoma transformaram a biologia em muitos sentidos, mas

o mais impressionante avanço foi a emergência da bioinformática e o treinamento dos cientistas em tecnologias modernas de pesquisa. Inicialmente a bioinformática teve como aplicação principal facilitar o manuseio da grande quantidade de dados gerados pelos projetos genoma, como a montagem de contigs e fechamento de seqüências genômicas, além de dar suporte para outras estratégias experimentais no campo da biologia molecular.

De lá para cá, muitas informações foram disponibilizadas em bancos de dados públicos de seqüências gênicas, proteínas, estruturas de macromoléculas, perfil metabólico, filogenia e outros, cujo valor ainda não pode sequer ser estimado. Hoje não é mais possível avançar em biotecnologia sem a integração da tecnologia da informação com a tecnologia experimental. As abordagens de estudos biotecnológicos atualmente buscam resolver questões específicas, optando-se normalmente por fazer uma análise computacional inicial com a utilização dessas informações para direcionar e selecionar as estratégias experimentais, com considerável economia financeira e de tempo, sem considerar a efetividade de tais procedimentos na aceleração da obtenção dos resultados e descobertas científicas.

Além disso, muitas descobertas estão sendo feitas simplesmente pela análise sistematizada dessas fontes de dados, que não param de crescer tanto em volume como em complexidade e variabilidade. A tendência atual é para descobertas científicas e síntese sendo dirigidas pela informação emergindo intrinsecamente a partir da biologia em si e a partir da diversidade e heterogeneidade das observações experimentais. Um projeto típico de pesquisa pode começar com a coleção de sequências genômicas conhecidas ou não conhecidas. Para sequências não conhecidas, pode-se conduzir uma busca em bancos de dados por sequências similares ou usar algoritmos computacionais procurando predizer as suas possíveis identidades e funções.

Isso requer o acesso à versão mais atual da coleção de dados, em bancos de dados mundiais, e as ferramentas fundamentais da bioinformática agora são cada vez mais parte dos métodos experimentais. Entretanto, essas informações estão espalhadas em múltiplas fontes, impossibilitando que os cientistas obtenham direta e eficientemente a informação requerida para converter os dados complexos e heterogêneos em dados úteis, informação organizada e sistematizada conforme as linhas de pesquisa específicas.

Nesse ambiente, para responder uma simples questão pode ser necessário acessar várias fontes de dados e utilizar ferramentas de análise sofisticadas, como alinhamento de sequências, agrupamento, modelagem molecular etc. Enquanto a integração dos dados é uma área de pesquisa dinâmica, necessidades específicas dos biocientistas têm levado ao desenvolvimento de numerosos sistemas que acabam desconectando o acesso aos dados em um ambiente direcionado por resultados. O resultado é o crescente número de bancos de dados e web sites representando uma coleção confinada de dados, governada por sistemas próprios de gerenciamento e formatos particulares de input e output dos dados, apresentações gráficas dos resultados, e problemas sérios de compatibilidade e interoperabilidade com outros sistemas. Uma evidência disso é o número crescente de novos bancos de dados relatados a cada ano na edição de janeiro da *Nucleic Acids Research* (http://nar.oupjournals.org/). A edição atual lista 548 bancos de dados, 162 a mais em relação ao ano anterior (Galperin, 2004). Boa parte desses bancos ainda são construídos com enfoques extremamente limitados para aplicações restritas, sem a preocupação com relação à compatibilidade e troca de informações com outros sistemas. Adaptações são lentas e muitas vezes difíceis de implementar quando a filosofia básica do banco precisa ser mantida.

O acesso a esses dados precisa melhorar em termos de eficiência, velocidade e facilidade. Para facilitar o entendimento dos processos biológicos, é necessário fazer novos arranjos aos recursos de dados disponíveis. Por exemplo, o que se faz inicialmente em uma rota metabólica, uma rede de interações moleculares etc., é necessário generalizar para outros sistemas biológicos; a partir de *E. coli* para levedura, e chegar à biologia de organismos mais complexos, como o homem, animais e plantas economicamente importantes. Trabalhar toda essa informação conjuntamente é fundamental para a geração de novos insights. O rápido crescimento do volume de dados é um desafio para cada um, e com a produção de dados mais diversos e em larga escala (por exemplo, dados de *DNA microarrays*) esse crescimento está apenas começando.

As atividades de bancos de dados e desenvolvimento de algoritmos computacionais precisam estar integradas para produzir uma infra-estrutura de informação coesiva delimitando toda a biologia. Para isso é necessário o desenvolvimento de ferramentas para disseminar e analisar massivas quantidades de dados, inclusive literatura, e a construção de comunidades de bancos de dados baseadas em princípios operacionais padronizados e com padrões interoperacionais.

Muitos dos problemas da bioinformática são genéricos, por isso soluções em um domínio podem ser naturalmente aplicáveis para outros. O entendimento da informação molecular até a célula, órgão e o sistema biológico do organismo será o maior desafio (fenomenoma). A passagem do genótipo para o fenótipo requererá um novo conjunto de ferramentas computacionais altamente robustas. O principal enfoque da bioinformática para os próximos anos será integrar esses dados de modo a permitir buscas transparentes através dos dados. Fazer isso de forma robusta abrangendo todo o conjunto de dados é um desafio real.

Apesar do avanço já feito, é necessário continuar a pesquisa no campo da genômica, principalmente para microrganismos associados a plantas economicamente importantes, incluindo fungos, e buscar entender as interações hospedeiro-microrganismo ou planta-patógeno. No caso da medicina, a necessidade atual é por dados clínicos bem estruturados e consistentes sobre grandes populações. Tais dados, que são difíceis de coletar e caros, serão críticos para ligar os

dados moleculares com o fenótipo. Embora exista um crescente número de centros de bioinformática, a maior tendência é que ela esteja presente nos centros de pesquisa e nas universidades, em cada departamento de biologia ou biotecnologia, em cada faculdade na área das ciências biológicas em todo o mundo. Todos os grandes centros de pesquisa terão que ter profissionais especializados em bioinformática/biologia computacional. Hoje é consenso geral que essas instituições necessitam de pessoas com esse entendimento em seus departamentos de biologia e necessitarão formar os seus estudantes de graduação em biologia quantitativa em vez de somente biologia experimental. Os experimentos precisam ser feitos no contexto do conhecimento corrente, e os dados gerados precisam ser rapidamente armazenados e explorados computacionalmente juntamente com o universo de informação disponível.

Nunca na história da ciência as informações foram tão democraticamente acessíveis como hoje. Especialmente as informações e ferramentas disponibilizadas pela bioinformática. Não importa quem e onde. O mesmo tipo de informação pode ser acessada por qualquer pessoa, em qualquer lugar do mundo. Praticamente todas as ferramentas de bioinformática e bancos de dados disponíveis podem ser dispostos de modo que possam ser acessadas e utilizadas na web. Basta fazer a pergunta correta e buscar a resposta.

Conclusão

O debate que está emergindo atualmente é se existe uma pletora ou escassez de dados experimentais proveitosos derivados pala plataforma das ômicas. O grande desafio, no entanto, é o que se pode fazer com esses dados. Não há dúvida de que a tecnologia da informação precisa ser tomada como parte integral do processo de descoberta pelos pesquisadores no campo da biologia. Este é o problema fundamental que precisa ser resolvido pela bioinformática, promovendo um profundo impacto no processo de descobertas biológicas. É necessário que ocorram discussões freqüentes entre todos os especialistas participantes de estudos relacionados, visando um emprego mais adequado da cultura científica dos participantes, já que, de modo simplificado, os biólogos querem entender como os organismos funcionam e os cientistas da computação querem fazer ferramentas que resolvam problemas. O estabelecimento de uma linguagem comum entre os especialistas em diferentes áreas, o monitoramento de quais ferramentas são mais usadas e importantes para o escopo do estudo, uma filosofia orientada para novas descobertas, não orientada por dogmas, são recomendações importantes para o sucesso dos empreendimentos científicos. Treinamentos constantes e workshops devem fazer parte dos investimentos previstos nos projetos.

O bom entendimento entre os pesquisadores de diferentes áreas é fundamental. Por exemplo, os cientistas da computação devem ser pacientes com o biólogo, já que este geralmente não sabe exatamente onde quer chegar ou o que espera dos dados (o que é natural nos estudos biológicos). Deve ensinar pelo menos os conceitos básicos de computação para estabelecer uma plataforma comum de comunicação, encorajar os biólogos a mostrar como eles estão realmente usando as ferramentas disponibilizadas e buscar sempre proporcionar o máximo de acesso aos dados. A retenção longa dos dados inibe o espírito de comunidade. Por parte do biólogo, espera-se que não espere muito ou tente fazer as coisas sozi - nho, fale com uma variedade de cientistas da computação, encontre aqueles mais interessados no seu problema, encontre aqueles com quem gosta de trabalhar, faça perguntas com freqüência e logo que surjam, use uma variedade de novas ferramentas, fazendo comentários/sugestões assim que puder e busque entender os desafios da computação para solucionar problemas novos. A obtenção de novos conhecimentos acelera quando todos contribuem.

Agradecimentos

Aos colegas Dr. Francisco Prosdocimi, Dr. Newton Portilho Carneiro e Dr. Alexandre Lima Nepomuceno pela revisão crítica deste artigo.

Referências

Bassingthwaighte JB. Strategies for the physiome project. Ann Biomed Eng. 2000, 28(8):1043-58. PMID: 11144666

Bernal A, Ear U, Kyrpides N. Genomes OnLine Database (GOLD): a monitor of genome projects worldwide. Nucleic Acids Res. 2001, 29(1):126-127. PMID: 11125068

Collins FS, Green ED, Guttmacher AE, Guyer MS; US National Human Genome Research Institute. A vision for the future of genomics research. Nature. 2003, 422(6934):835-47. PMID: 12695777

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. Nature. 2004, 429(6987):92-6. PMID: 15129285

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995, 269(5223):496-512. PMID: 7542800

Galperin MY. The Molecular Biology Database Collection: 2004 update. Nucleic Acids Res. 2004, 1;32 Database issue:D3-22. PMID: 14681349

Gao F, Foat BC, Bussemaker HJ. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics. 2004, 18;5(1):31. PMID: 15113405

Garavelli JS. The RESID Database of Protein Modifications: 2003 developments. Nucleic Acids Res. 2003, 31(1):499-501. PMID: 12520062

Genomics and Its Impact on Science and Society: The Human Genome Project and Beyond. U.S. Department of Energy Human Genome Program. 2003. Disponível http://www.ornl.gov/sci/techresources/Human_Genome/

publicat/primer2001/index.shtml
Goodsell DS. Inside a living cell. Trends Biochem Sci. 1991, 16(6):203-206. PMID: 1891800
Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res. 2002, 30(1):402-4. PMID: 11752349
Guerreiro N, Staedtler F, Grenet O, Kehren J, Chibout SD. Toxicogenomics in drug development. Toxicol Pathol. 2003, 31(5):471-9. PMID: 14692614
Hoersch S, Leroy C, Brown NP, Andrade MA, Sander C. The GeneQuiz web server: protein functional analysis through the Web. Trends Biochem Sci. 2000, 25(1):33-35. PMID: 10637611
James P. Protein identification in the post-genome era: the rapid rise of proteomics. Q Rev Biophys. 1997, 30(4):279-331. PMID: 9634650
Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004, 32 Database issue:D277-D280. PMID: 14681412
Kramer JA, Kolaja KL. Toxicogenomics: an opportunity to optimise drug development and safety evaluation. Expert Opin Drug Saf. 2002, 1(3):275-86. PMID: 12904143
Kondro W. MOLECULAR BIOLOGY: Consortium Tackles Mouse Regulome. Science. 2004, 304(5673):942A. PMID: 15143247
Lefranc MP. IMGT, The International ImMunoGeneTics Information System, http://imgt.cines.fr. Methods Mol Biol. 2004, 248:27-49. PMID: 14970490
Lopez-Otin C, Overall CM. Protease degradomics: a new challenge for proteomics. Nat Rev Mol Cell Biol. 2002, 3(7):509-19. PMID: 12094217
Marshall A. Genset-Abbott deal heralds pharmacogenomics era. Nat Biotechnol. 1997, 15(9):829-30. PMID: 9306389
Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, Warfsmann J, Ruepp A. MIPS: analysis and annotation of proteins from whole genomes.Nucleic Acids Res. 2004, 32 Database issue:D41-D44. PMID: 14681354
Novik KL, Nimmrich I, Genc B, Maier S, Piepenbrock C, Olek A, Beck S. Epigenomics: genome-wide study of methylation phenomena. Curr Issues Mol Biol. 2002, 4(4):111-28. PMID: 12432963
Palsson B. In silico biology through "omics". Nat Biotechnol. 2002, 20(7):649-50. PMID: 12089538
Palsson B. The challenges of in silico biology. Nat Biotechnol. 2000, 18(11):1147-50. PMID: 11062431
Pennisi E. Human genome. Reaching their goal early, sequencing labs celebrate. Science. 2003a, 300(5618):409. PMID: 12702850
Pennisi E. Human genome. A low number wins the GeneSweep Pool. Science. 2003b, 300(5625):1484. PMID: 12791949
Pennisi E. Bioinformatics. Gene counters struggle to get the right answer. Science. 2003c, 301(5636):1040-1. PMID: 12933991
Prates MV, Bloch C. Peptídeos antimicrobianos. Biotecnologia Ciência e Desenvolvimento. 2002, 29: 30-36.
Roven C, Bussemaker HJ. REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. Nucleic Acids Res. 2003, 31(13):3487-90. PMID: 12824350
Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Lewis S. Comparative genomics of the eukaryotes. Science. 2000, 287(5461):2204-2215. PMID: 10731134
Sanford K, Soucaille P, Whited G, Chotani G. Genomics to fluxomics and physiomics - pathway engineering. Curr Opin Microbiol. 2002, 5(3):318-22. PMID: 12057688
Schilling CH, Edwards JS, Palsson BO. Toward metabolic phenomics: analysis of genomic data using flux balances. Biotechnol Prog. 1999, 15(3):288-95. PMID: 10356245
Schulz-Knappe P, Zucht HD, Heine G, Jurgens M, Hess R, Schrader M. Peptidomics: the comprehensive analysis of peptides in complex biological mixtures. Comb Chem High Throughput Screen. 2001, 4(2):207-17. PMID: 11281836
Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV. The Ensembl web site: mechanics of a genome browser. Genome Res. 2004, 14(5):951-955. PMID: 15123591
Steuer R, Kurths J, Fiehn O, Weckwerth W. Observing and interpreting correlations in metabolomic networks. Bioinformatics. 2003, 19(8):1019-26. PMID: 12761066
Weckwerth W, Loureiro ME, Wenzel K, Fiehn O. Differential metabolic networks unravel the effects of silent plant phenotypes. Proc Natl Acad Sci U S A. 2004. PMID: 15136733
Weinstein JN. Searching for pharmacogenomic markers: the synergy between omic and hypothesis-driven research. Dis Markers. 2001, 17(2):77-88. PMID: 11673654
Werner T. Proteomics and regulomics: the yin and yang of functional genomics. Mass Spectrom Rev. 2004, 23(1):25-33. PMID: 14625890
Wright FA, Lemon WJ, Zhao WD, Sears R, Zhuo D, Wang JP, Yang HY, Baer T, Stredney D, Spitzner J, Stutz A, Krahe R, Yuan B. A draft annotation and overview of the human genome. Genome Biol. 2001, 2(7):RESEARCH0025. PMID: 11516338
Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS Lett. 2002, 513(1):135-40. PMID: 11911893