

Bioinformática como ferramenta nas pesquisas atuais*

Wagner Arbex[†]

Vítor Manuel Morais Santos Costa[‡]

Marcos Vinícius G. Barbosa da Silva[§]

Agosto 2006

1 Introdução

O objetivo do presente texto é encaminhar o participante do “III Encontro de Genética e Melhoramento da Universidade Federal de Viçosa”, que comemora os “30 Anos de Inovação e História” do Programa de Pós-graduação em Genética e Melhoramento, dentro das possibilidades da bioinformática na pesquisa, sendo simplesmente um agente provocador para que o leitor, a partir do que será apresentado, busque um conhecimento maior.

O texto encontra-se dividido em três partes. A primeira trata de uma definição de bioinformática. Obviamente, o leitor tem exata noção do que vem a ser bioinformática, porém como é uma área de pesquisa recente, aproximadamente entre 20 e 30 anos, existem muitas e diferentes definições e abordagens para o mesmo conceito.

O que se propõe, então é uma definição para o termo “bioinformática” que seja adequada ao escopo deste trabalho. Na verdade, não será proposta nenhuma “nova” ou “revolucionária” definição, mas uma simplificação e uma generalização, partindo da diferenciação entre “bioinformática” de “biologia computacional”.

Uma vez definido o escopo do termo bioinformática, na segunda parte do texto, será feita uma breve discussão dos trabalhos de bioinformática. Como praticamente qualquer área do conhecimento, a bioinformática pode ser uma “simples” ferramenta para a pesquisa em si, ou a responsável pela interpretação do resultado, ou melhor, em última análise, pela geração do conhecimento a ser aplicado ou utilizado por outras disciplinas.

Na terceira parte do trabalho serão apresentados alguns casos, onde a bioinformática pode ser vista como ferramenta nas pesquisas, gerando conhecimento, e como suas ferramentas podem ser utilizadas nas pesquisas, ou seja, como seus métodos, procedimentos etc. podem ser o instrumento de prospecção.

Especificamente, no primeiro caso, será discutido como a bioinformática pode contribuir com a geração de conhecimento para diversas áreas de pesquisa, tais como biologia molecular, bioquímica e antropologia, auxiliando nas investigações destas. Como exemplo, serão apresentados alguns resultados que podem ser obtidos a partir do estudo de polimorfismos de base única (*single nucleotide polymorphisms* - SNPs).

*ARBEX, W.; SILVA, M. V. G. B. da; COSTA, V. M. M. S. Bioinformática como ferramenta nas pesquisas atuais. In: III ENCONTRO DE GENÉTICA E MELHORAMENTO DE UNIVERSIDADE FEDERAL DE VIÇOSA, **Anais do III Encontro de Genética e Melhoramento da Universidade Federal de Viçosa**. Viçosa: Universidade Federal de Viçosa, 2006.

[†]Analista “A” da Embrapa Gado de Leite. Rua Eugênio do Nascimento, 610, CEP 36038-330, Juiz de Fora, MG, e-mail: arbex AT cnp.gl.embrapa.br

[‡]Professor Adjunto da Universidade Federal do Rio de Janeiro

[§]Pesquisador “A” da Embrapa Gado de Leite

No segundo caso, será exemplificado o que pode ser feito por meio da bioinformática na análise de dados de microarranjos (*microarrays*) para identificação e determinação de vias ou redes metabólicas.

O enfoque dos trabalhos selecionados para ilustrar este último caso, mostra a bioinformática “como ferramenta” da biologia molecular e da bioquímica e, ao mesmo tempo, como as “ferramentas de” bioinformática podem ser utilizadas.

Como foi dito, espera-se que o leitor entenda a abordagem do assunto, onde a idéia é contextualizar a bioinformática dentro da pesquisa, e aceite a provocação de investigar na profundidade adequada o tema tratado.

2 O que é bioinformática?

Uma das referências que marca a origem da bioinformática é o início dos projetos genomas. Todos sabem o grande volume de dados gerados por tais projetos e, portanto a necessidade de sistemas de computação com grande aporte computacional e de profissionais especializados para a manipulação dos mesmos.

Nesse contexto, foi possível perceber o surgimento da figura de um novo profissional, ou melhor, de um novo perfil de profissional que deveria entender o problema biológico, tratar o problema de forma a criar um modelo de representação do mesmo, implementar o modelo e, por fim analisar o resultado concebido pelo modelo.

Era necessário um profissional que possuísse conhecimento suficiente de:

- biologia molecular, para, no mínimo, entender o problema;
- matemática e probabilidade, de forma que fosse capaz de estabelecer um modelo de representação do problema;
- computação, possibilitando a implementação da solução correta e apropriada, sendo esta inicialmente conhecida ou não;
- estatística, para que as soluções fossem comprovadamente confiáveis.

Em resumo, por Prodocimi et al. (2002), “esse profissional deveria ter conhecimento suficiente para saber quais eram os problemas biológicos reais e quais seriam as opções viáveis de desenvolvimento e abordagem computacional dos problemas em questão”. Logo, a figura desse profissional, que será discutida mais tarde, ficou conhecida como “biólogo computacional” ou “bioinformata”.

Com esse profissional surgiu uma nova área de conhecimento que recebeu algumas denominações, tais como, “biologia computacional” e “bioinformática”, sendo essas as mais comuns, e também outras não muito utilizadas, mas, bastante interessantes, como “biologia molecular in silico”.

É fato que todos os termos utilizados na denominação desta nova área não trouxeram junto uma definição precisa do conceito a ser apresentado. Além disso, chegavam a ser contraditórias, visto que por algum tempo a bioinformática foi definida como um subconjunto da biologia computacional, mas, em outros momentos, enunciava-se o inverso.

Atualmente, e a cada dia com mais intensidade, percebe-se que não existe uma “predominância” da biologia sobre a informática, ou vice-versa, mas a natureza do problema a ser tratado é que vem determinar o enfoque inicial a ser adotado na abordagem do problema. As definições estão cada vez mais se fundindo e se agrupando em torno do termo “bioinformática”, sem distinção de áreas entre biologia computacional e bioinformática. O que, em parte, pode ser observado pelo simples uso da palavra “bioinformática”, ou da equivalente em inglês “bioinformatics”, estar sendo muito mais utilizada do que a expressão “biologia computacional” ou “computational biology”.

Tal observação pode ser comprovada em uma rápida pesquisa (Figura 1) por meio da máquina de busca mais utilizada na atualidade, o Google, pelos termos “bioinformática” e “biologia computacional” e por seus equivalentes “bioinformatics” e “computational biology” que retornou, em número aproximado de referências, respectivamente, 1,45 milhão, 41,8 mil, 98,3 milhões e 16,7 milhões.

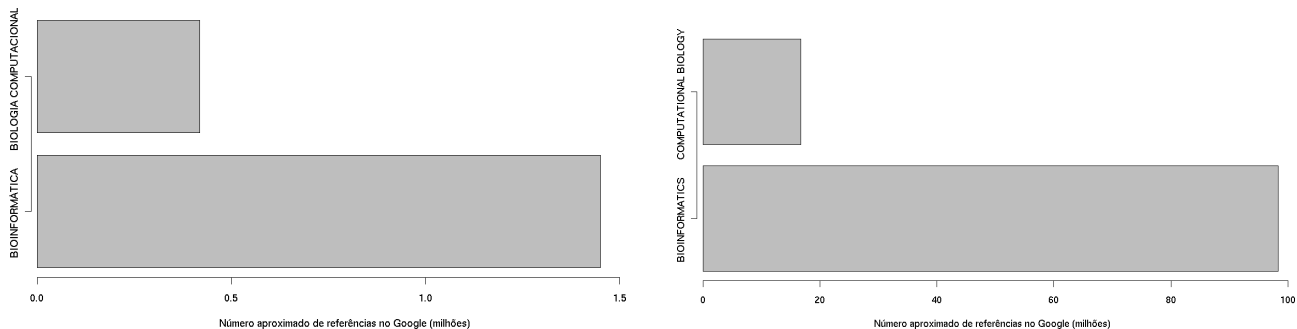


Figura 1: Pesquisa, feita no Google, pelos termos “bioinformática”, “biologia computacional”, “*bioinformatics*” e “*computational biology*”, em 13/07/2006.

Segundo a Wikipedia, em Wikipedia (2006a), os termos “bioinformática” e “biologia computacional” são usados frequentemente de forma permutável e envolve o uso de técnicas aplicadas de matemática, ciência da computação e estatística para resolver problemas biológicos.

Como área de conhecimento, a bioinformática é formada (Figura 2) principalmente por duas grandes disciplinas das ciências biológicas e da computação: a biologia molecular e a inteligência artificial, além de agregar muitos elementos de outras disciplinas destas duas ciências.

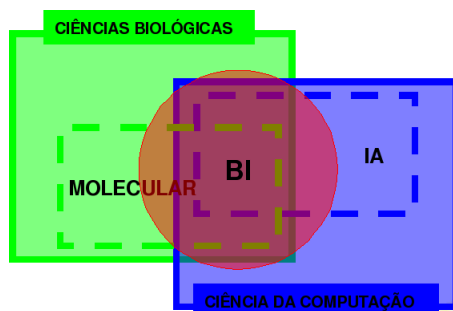


Figura 2: Representação simplificada da área de conhecimento da bioinformática

Ainda conforme Wikipedia (2006a), também inclui subdivisões de outras ciências igualmente importantes, tais como bioquímica computacional e biofísica computacional.

Diferentes grupos de pesquisa visualizam de diferentes formas a abrangência da bioinformática. Prodocimi e Santos (2004) observam que o surgimento da bioinformática clássica ocorreu com o seqüenciamento de biomoléculas e, assim:

“destas deve permanecer inseparável, sendo possível definir de forma razoavelmente clara: a bioinformática consiste em ‘todo tipo de estudo ou de ferramenta computacional que se pode realizar e/ou produzir de forma a organizar ou obter informação biológica a partir de seqüências de biomoléculas’”.

Setubal (2003) propõe duas abordagens. Uma interpretação estreita, segundo a qual bioinformática seria basicamente o uso de ferramentas computacionais para problemas da biologia molecular, e a interpretação ampla, em que: “a bioinformática seria a nova ciência que procura interpretar o ‘mundo dos seres vivos’ através dos conceitos da informática”.

Em adição, explica que o “mundo dos seres vivos” é uma hierarquia de camadas de complexidade crescente, começando pelas moléculas e seqüências biológicas (DNA, RNA e proteínas), as quais são estudadas e analisadas individualmente.

A partir do primeiro nível desta hierarquia, segue-se a interação entre moléculas, momento em que se torna importante entender a função de cada molécula; a célula, em que se busca entender como funcionam os processos celulares; e ainda os tecidos e órgãos, os indivíduos, as populações e a biosfera.

Na atualidade, pela natureza dos trabalhos que estão sendo desenvolvidos, ainda não é possível concordar com a amplitude dessa definição e, notadamente, a bioinformática atua maciça e predominantemente nas “primeiras camadas”, ou seja, entre moléculas e seqüências, e interação entre moléculas e processos celulares.

Assim, para este texto e por uma visão objetiva e pragmática, a bioinformática consiste no desenvolvimento e no uso de técnicas de informática, de modelagem matemática e computacional e de modelagem probabilística e estatística para resolver problemas de biologia molecular, estabelecendo uma “convergência tecnológica” a partir do uso do conhecimento proveniente dessas áreas que a fundamentam.

Voltando ao bioinformata, seria inocência idealizar que, para atender a tantas exigências, com o nível de conhecimento e complexidade esperados, esse profissional devesse ter obrigatoriamente uma formação tão ampla e profunda quanto se queira imaginar em todo o conhecimento envolvido.

Ocorre, porém, como já foi enunciado de várias formas, que a bioinformática é multidisciplinar, o que normalmente provoca a especialização do profissional em determinada subregião ou, eventualmente, em um grupo de disciplinas correlacionadas. O perfil e as necessidades do bioinformata promovem, naturalmente, o trabalho não em equipes e entre equipes, devido à amplitude do campo de conhecimento, formando as redes de pesquisa.

3 “Faz DNA?”: as ferramentas de bioinformática e a bioinformática como ferramenta

A bioinformática pode ser utilizada sob dois diferentes aspectos. Em uma primeira abordagem, a “ferramenta de bioinformática” pode ser complementar ou basilar e fornecer infraestrutura para o desenvolvimento dos trabalhos de pesquisa. Já a “bioinformática como ferramenta” nas pesquisas é responsável pela interpretação do conhecimento gerado, que será utilizado por outras áreas de pesquisa. Na verdade, essa discussão vem contribuir com a definição já discutida na seção anterior, querendo apenas destacar as formas com as quais a bioinformática é utilizada.

Provavelmente, o exemplo mais notório de utilização das ferramentas de bioinformática é o seu uso no armazenamento de dados, que permite o tratamento a priori de diversos tipos de dados da “era genômica”. Ou seja, após o trabalho nos laboratórios e nas bancada para obtenção dos mesmos, ficou a cargo da bioinformática a identificação, a organização, o armazenamento, a recuperação, a classificação, a apresentação etc. das informações geradas.

O grande volume de informações é um dos desafios do bioinformata. Em agosto de 2005, o International Nucleotide Sequence Database Collaboration (INSDC), formado pelo DNA DataBank of Japan (DDBJ) o European Molecular Biology Laboratory (EMBL) e o GenBank (National Center for Biotechnology Information - NCBI), anunciou ter ultrapassado a marca de 100 gigabases (MEHNERT; CRAVEDI, 2005), ou seja, 100.000.000.000 de pares de bases¹.

Na época em questão, o GenBank era responsável por aproximadamente 87 gigabases e, seis meses depois, em fevereiro de 2006, já haviam sido computadas cerca de 122,9 gigabases (NCBI, 2006b).

¹Considerando o número de pares de bases de seqüências obtidas por registro tradicional e o número de pares de bases de seqüências registradas por projetos que utilizaram o método “shotgun” (*whole genome shotgun* - WGS)

Ou seja, um crescimento de aproximadamente 40% no volume de dados armazenados, somente no GenBank. Outro exemplo pode ser visto na Figura 3, que apresenta o crescimento do GenBank, considerando o número de seqüências depositadas no período de 1982 à 2005 (NCBI, 2006c).

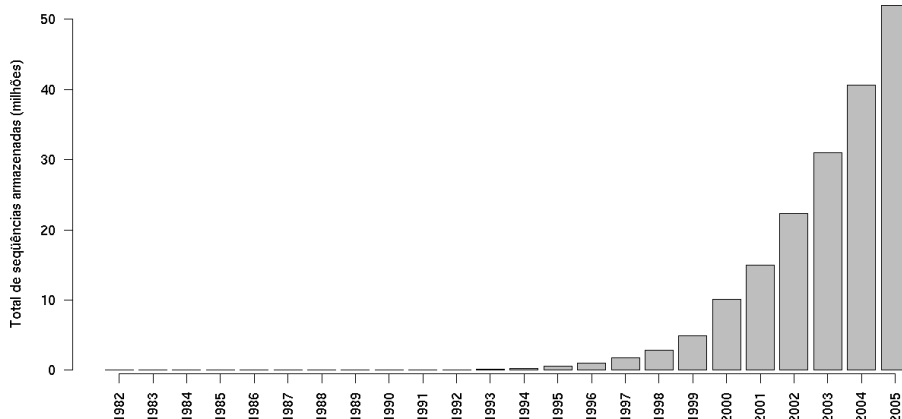


Figura 3: Crescimento do GenBank (NCBI, 2006c).

Os dois exemplos delineiam a questão principal da bioinformática como ferramenta para tratamento de dados. Somente com os números do GenBank, sem considerar todo o INSDC ou outras bases de dados genômicos espalhadas pelo mundo, é possível perceber que o crescimento exponencial dos dados, em muitos períodos, ultrapassa a previsão do aumento da capacidade de processamento (Lei de Moore²), o que ilustra o problema do tratamento dos dados, isto é, a quantidade de dados cresce mais do que da capacidade de processamento dos mesmos.

Ainda como ilustração, seguem os números relacionados ao armazenamento de dados de etiquetas de seqüências expressas (expressed sequence tags - EST). O dbEST, a base de dados EST do GenBank, registra o total de 37.642.134, considerando o total de 1.167 diferentes organismos (NCBI, 2006a).

O segundo aspecto de utilização da bioinformática é na geração de conhecimento para outras áreas: a bioinformática como ferramenta nas pesquisas atuais. Não se discute estreita ligação da bioinformática com os projetos genomas, entretanto a estes não se restringe. No decorrer da última década os trabalhos de bioinformática estão cada vez mais diversificados e mais próximos da área em que estão sendo aplicados os resultados dos trabalhos.

A primeira vista, pode-se pensar que esta discussão trata “somente” de definir pesquisa aplicada, ou melhor, na aplicação direta dos resultados nas áreas de estudo, mas o diferencial está na capilaridade dos trabalhos e nos números do mercado de bioinformática. Em julho de 2006, a Research & Consultancy Outsourcing Services - RNCOS, publicou em seu relatório Bioinformatics Market Update 2006, que apenas nos processos de seleção, identificação e validação de drogas, o mercado, em todo o mundo, crescerá cerca de três vezes em oito anos. Partindo de US\$ 6.3 bilhões, em 2002, e chegando US\$ 19,5 bilhões em 2010 (RNCOS, 2006).

Ainda de acordo com a RNCOS, somente em 2005 o mercado de bioinformática cresceu aproximadamente 16% e deve manter essa taxa de até 2010.

Os números desse mercado só não são mais expressivos do que a inserção e o impacto das aplicações dos trabalhos nos meios acadêmicos e até mesmo no cotidiano e no dia a dia popular.

Há pouco tempo, ninguém seria capaz de imaginar que um apresentador, em um programa de auditório em televisão, desafiaria um participante mediante a meios e ferramenta científico. Hoje, contudo, é comum ver apresentadores de sucesso, “organizando” um circo em que, juntamente com a

²O fundador da Intel, Gordon Moore, constatou que a cada 18 meses a capacidade de processamento dos computadores dobra, enquanto os custos permanecem constantes. A Lei de Moore está em vigor há mais de 30 anos e a maioria dos especialistas acredita que deve durar pelo menos mais cinco gerações de processadores (WIKIPEDIA, 2006b).

platéia aos brados, desafia um possível progenitor, que parece estar entregue aos leões - ou seria, aos ratinhos: “faz DNA?”. Sem dúvida, isto é pesquisa científica aplicada como poucas vezes se vê.

4 O que está sendo feito?

De forma ampla e não aprofundada, serão apresentados dois casos na tentativa de exemplificar a bioinformática como ferramenta nas pesquisas, ou seja, na geração do conhecimento para outras áreas, e o uso das ferramentas de bioinformática utilizadas em trabalhos característicos da própria bioinformática.

Os casos apresentados a seguir são apenas dois exemplos que servem exclusivamente para ilustrar a abordagem proposta e, conseqüentemente não encerram a discussão.

4.1 Investigação de polimorfismos de base única e sua utilização em outras áreas do conhecimento

Na proposta desse texto, qualquer tentativa de conceituação de polimorfismos de base única (*single nucleotide polymorphisms* - SNP) torna-se desnecessária, porém serão apresentados alguns pontos básicos que ajudarão na linha de raciocínio em que a bioinformática é utilizada como ferramenta para a geração de conhecimento de outras ciências.

Os projetos genoma trouxeram muitas revelações para a humanidade. Especificamente, uma das descobertas do projeto genoma humano foi verificar que o código genético humano mostrou-se mais variado e complexo do que propriamente maior, quando comparado ao de outras espécies.

Uma das variações e particularidades do genoma, humano ou de qualquer espécie, são os chamados polimorfismos de base única, modificações de um único nucleotídeo, em uma dada seqüência quando comparada a outra, que podem alterar a formação de uma certa proteína e, se for o caso, o conjunto dessas alterações pode provocar variações de características individuais.

De acordo com Brondani e Brondani (2004), a maior parte do genoma entre os indivíduos de uma mesma espécie é idêntica, porém existe a variabilidade genética, que são as diferenças encontradas em algumas regiões do genoma. A variabilidade consiste da alteração nas seqüências de bases ao longo do DNA e ocorrem por substituição, ausência ou duplicação de bases.

Essencialmente, o uso de ferramentas de bioinformática no estudo de polimorfismos de base única busca fornecer meios para que seja possível esclarecer as seguintes questões:

- Como identificar um polimorfismo de base única em uma seqüência?
- Como comprovar se o “nucleotídeo trocado” que caracteriza a seqüência como polimórfica, é realmente um caso de polimorfismo? A “base diferente” pode ser um simples erro de identificação.
- O polimorfismo provocará alteração na seqüência de bases a ponto de alterar a conformação de uma proteína, formando uma “nova” proteína?
- A nova proteína, se esta realmente foi formada, quando combinada com as demais provocará ou suprirá a manifestação de alguma característica específica no indivíduo?

Modificações originadas por polimorfismos também poderiam ser vistas como mutações. Em termos gerais, a diferença entre o que é um SNP e o que é uma mutação é determinada em função do número de ocorrências da alteração de base, mais especificamente, em função da freqüência alélica.

Assim, caso uma alteração de base, em uma determinada população, ocorra com freqüência superior a 1%, fica caracterizado a ocorrência de SNP, caso contrário, a alteração caracteriza uma mutação (GUIMARÃES; COSTA, 2002; BARNES; GRAY, 2003).

Como esclarecimento, deve ser ressaltado o uso do termo “mutação”. Atualmente, a definição apresentada vem sendo negligenciada e as alterações de base com frequência menor do que 1% estão sendo chamadas de “variações de baixa frequência”, enquanto o termo mutação está sendo utilizado para denominar variações genômicas que estejam relacionadas com doenças no indivíduo (BARNES; GRAY, 2003).

Além das aplicações mais comuns de SNPs, tais como, nos estudos de correlações entre genótipo e fármacos, ou seja, interações entre drogas e uma proteína em particular (LESK, 2005; BALDI; BRUNAK, 2001), a identificação de resistência ou susceptibilidade de indivíduos em relação a certas doenças, definição de marcadores de predisposição a determinadas patologias e de prognóstico a diferentes tratamentos (GUIMARÃES; COSTA, 2002), outras ciências não muito próximas da bioinformática também utilizam a bioinformática - e nesse caso, a análise de SNPs - como ferramenta em suas respectivas áreas de pesquisas.

Atualmente, SNPs podem ser empregados, entre outras, em áreas como medicina forense, antropologia molecular, evolução, genética de populações, conservação e manejo de fauna (GUIMARÃES; COSTA, 2002).

Por exemplo, estudos antropológicos e sociológicos utilizam a alteração de bases em seqüências genéticas para determinação do padrão genético de populações, do indicativo de séries históricas de variação do tamanho de populações e de padrões de migração de populações (PENA et al., 2000; LESK, 2005).

Já se sabe que a evolução desse tipo de polimorfismo é lenta, além disso, é possível estabelecer períodos prováveis em que uma determinada população manifestou ou perdeu um SNP. De acordo com Barnes e Gray (2003), estudos relatam que existe 94% de probabilidade de que uma população venha a perder um SNP, ou mesmo uma mutação, em 10 gerações, cerca de 200 anos.

Assim sendo, uma vez estabelecido o período em que a seqüência polimórfica acompanhou a população e sabendo que a seqüência está restrita à mesma, é possível, com os dados e as ferramentas corretas, mapear a população que se quer estudar.

4.2 Análise de dados de microarranjos para identificação e determinação de redes de regulação gênica e vias metabólicas

Pelos mesmos motivos apresentados no caso anterior, não cabe no contexto tratar de fundamentos de microarranjos (*microarrays*), mas, da mesma forma, será necessário pontuar algumas questões para facilitar o entendimento dos exemplos.

Os trabalhos, Perkins, Hallett e Glass (2004), Nachman, Regev e Friedman (2004) definem modelos de interpretação e análise de dados de microarranjos em estudos de regulação gênica, que se diferem pela abordagem.

Em síntese, enquanto Perkins, Hallett e Glass (2004), propõem um modelo lógico, estabelecendo três regras de inferência, sendo uma delas uma função de estimação, Nachman, Regev e Friedman (2004), propõem um modelo de aprendizado, baseado em redes bayesianas dinâmicas (*bayesian dynamic networks* – BDN).

Perkins, Hallett e Glass (2004) observam que as novas tecnologias para monitoramento de expressão gênica, tais como ensaios com microarranjos, permitem medir a atividade de milhares de genes simultaneamente e, então possibilitam avaliar, por exemplo, reações ao longo do tempo à variações genéticas e ambientais.

Entretanto, surgem problemas que são comuns sob o ponto de vista dos modelos matemático e computacional, mas tomam uma nova dimensão pelo volume de dados a serem tratados, tais como “ruído” nas medidas ou medidas não anotadas, não computadas, não observadas etc. e a natureza complexa e estocástica do processo de regulação dos genes.

Descrevem o “modelo de regulação transcricional” baseado no comportamento dos agentes reguladores ao longo do tempo, ou melhor, nos parâmetros cinéticos destes, que são as taxas de transcrição de níveis de atividade reguladora, que atuam sobre os alvos.

O modelo é composto por três regras:

1. a espécie j regula a espécie i , se em dois diferentes momentos na mesma série, ou em séries temporais diferentes, todas as espécies, exceto j , têm o mesmo estado lógico, e taxa de produção de i é alterada;
2. a espécie j regula a espécie i , se em uma das séries temporais existir um momento de troca de estados lógicos em que j é uma das espécies cujo estado é modificado e, no momento da troca de estados, a taxa de produção de i é alterada;
3. a função de estimação de regulação estabelece a que taxa produção da espécie i é determinada em função dos estados lógicos das espécies que formam o conjunto de espécies reguladoras de i , sendo:

$$h(X_{\hat{R}_i}) = \begin{cases} 1 \\ 0 \\ ? \end{cases}$$

A análise dos dados de ensaios com microarranjos pelo modelo proposto permite o estabelecimento de funções lógicas que descrevem redes de regulação gênica associadas aos dados.

Considerando a via metabólica vista na Figura 4, a aplicação do modelo de regulação transcricional resultaria em doze funções, uma para cada proteína envolvida, das quais são apresentadas, como exemplo, as funções relativas às proteínas AG , $EMF1$ e LUG , sendo as duas últimas, respectivamente, auto-reguladora e sem reguladores:

$$\begin{aligned} f_{LUG} &= 0 \\ f_{EMF1} &= X_{EMF1} \\ f_{AG} &= \neg X_{TFL1} \wedge \neg X_{AP1} \wedge (X_{LFY} \vee \neg X_{LUG}) \end{aligned}$$

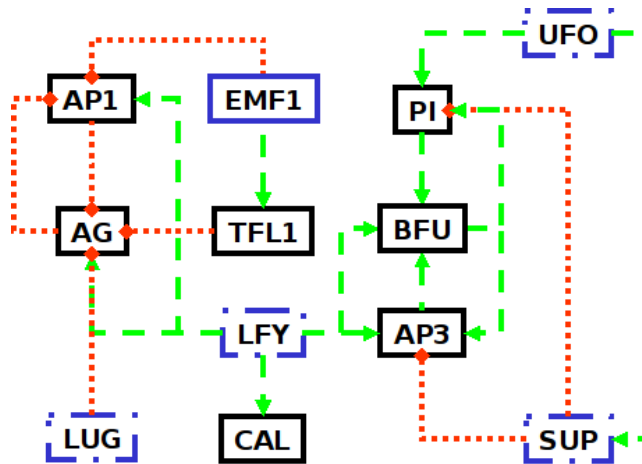


Figura 4: Exemplo de uma via metabólica da *Arabidopsis thaliana*, onde existem 7 proteínas comuns, representadas por retângulos em linha contínua preta, uma proteína auto-reguladora, representada por um retângulo em linha contínua azul, e 4 proteínas sem reguladores, representadas por retângulos em linha interrompida azul. Também são representados os sinais promotores, em linha tracejada verde, e repressores, em linha pontilhada vermelha (PERKINS; HALLETT; GLASS, 2004).

Deve ser notado que os conectores lógicos de conjunção (\wedge) e disjunção (\vee) só podem ser tomados a partir da interpretação dos dados pelas regras do modelo.

O trabalho de Nachman, Regev e Friedman (2004) procura o mesmo objetivo, que é entender a organização e a função de redes de regulação de genes, mas por um caminho complementamente diferente. Como motivação, os autores observam que o objetivo proposto é um desafio chave para a biologia molecular, tanto do ponto de vista experimental, quanto e computacional.

Em geral, modelos baseados em métodos probabilísticos não consideram taxas de transcrição e níveis de atividades reguladores, além disso, diversos trabalhos utilizam um diagrama de regulação fixo para identificar perfis e parâmetros de atividades de regulação não observadas.

A idéia do modelo de Nachman, Regev e Friedman (2004) passa pelo aprendizado de redes dinâmicas de transcrição, considerando o comportamento dos reguladores ao longo de um período e dos parâmetros cinéticos que podem afetá-los. Assim, propõem um modelo de aprendizado da estrutura da rede, das taxas de transcrição e níveis de atividade dos reguladores, a partir de modelos simplificados de redes de regulação e de dados de expressão gênica, obtidos por ensaios com microarranjos. Desta forma, estabelecem o “modelo temporal de regulons”, utilizando redes bayesianas dinâmicas (*dynamic bayesian networks* – DBN).

Regulons são vistos como uma “unidade genética”, com um ou mais “operons”, sendo destes últimos, os dados que permitem a diferenciação do modelo. Os operons são genes estruturais e, juntamente com suas seqüências, iniciam, atenuam ou terminam uma transcrição, ou seja, são “locais de controle”, que, corretamente interpretados, podem mostrar mais do que simplesmente detectar a atuação da proteína promotora ou repressora sobre o alvo (Figura 5).

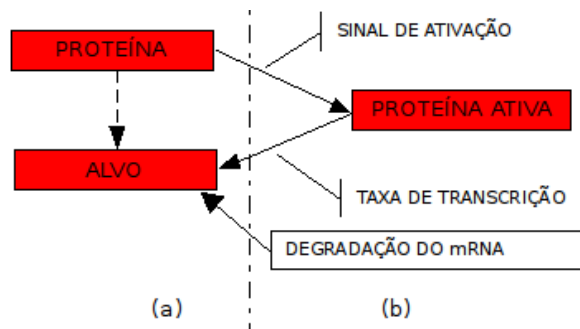


Figura 5: O modelo não se prende somente a ação da proteína sobre o alvo (a), mas descreve processo de regulação, considerando a atuação dos parâmetros cinéticos (b).

O modelo temporal de regulons considera a possibilidade de múltiplos reguladores sobre o alvo, múltiplos “genes alvos” e o comportamento dos alvos ao longo do período. Além disso, considera funções de regulação específicas para cada gene, apesar de, normalmente, um regulador poder atuar sobre vários alvos. Assim, o nível de atividade do regulador pode ser utilizado na função de regulação de todos os seus alvos.

Um dos elementos de aprendizado do modelo é a descrição das relações entre variáveis em um determinado ponto e no ponto consecutivo, assim para representar o ambiente do atributo da atividade reguladora H , suponha que o nível da atividade reguladora dependa do nível antecessor, como mostra a “equação de persistência”

$$H_i^{(t+1)} = H_i^t + \varepsilon_{H_i}^{(t+1)}$$

onde, $\varepsilon_{H_i}^{(t+1)}$ é o erro, com distribuição normal, média zero e variância σ_i .

Outro elemento do modelo de Nachman, Regev e Friedman (2004), é a taxa de transcrição de cada gene alvo, que depende do nível de atividade no momento do regulador que controla o gene. Então, supondo que a taxa R_k dependa de dois reguladores, H_1 e H_2 , pela função de regulação de Michaelis-Menten, pode-se obter:

$$R_k^t = g(H_1^t, H_2^t : \vec{\alpha}_k, \beta_k, \gamma_{k,1}, \gamma_{k,2})$$

onde, $\vec{\alpha}$ é o vetor dos sinais de ativação.

Portanto, taxa de transcrição I , por ser função temporal de atributos do nível de atividade reguladora, apresenta-se como resultado da soma de eventos estocásticos.

O modelo temporal de regulons é um modelo de aprendizado baseado em BDN, onde são considerados como elementos de entrada os dados de expressão gênica, obtidos nos ensaios com microarranjos, a taxa de transcrição I - juntamente com as informações referentes a taxa de decaimento do mRNA - e ainda um modelo simplificado da topologia entre a atividade dos reguladores e a taxa de transcrição.

A DBN, então, fornece uma saída composta dos parâmetros cinéticos, da rede de regulação e dos fatores de transcrição da atividade com as suas variações no tempo (Figura 6).

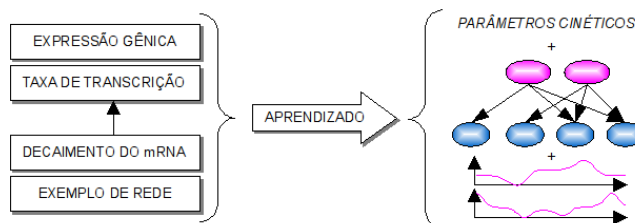


Figura 6: Representação do aprendizado da rede bayesiana dinâmica e os resultados esperados.

5 Considerações finais

Para atender ao objetivo do “III Encontro de Genética e Melhoramento da Universidade Federal de Viçosa”, o presente texto propôs definir o termo “bioinformática”, estabelecer seu escopo e sua importância no cenário da pesquisa e, por fim, como exemplo, discutir e apresentar trabalhos em assuntos de grande interesse na atualidade dentro do contexto da bioinformática. Desse modo, tem como finalidade e única pretensão estabelecer um ponto de partida para o leitor possa se aprofundar, discutir e criticar os pontos e questões apresentadas.

Referências

- BALDI, P.; BRUNAK, S. *Bioinformatics: the machine learning approach*. 2. ed. Cambridge: MIT Press, 2001.
- BARNES, M. R.; GRAY, I. C. *Bioinformatics for geneticists*. Chichester: John Wiley & Sons, 2003.
- BRONDANI, R. V.; BRONDANI, C. Germoplasma: base para a nova agricultura. *Ciência Hoje*, v. 35, n. 207, p. 70–73, Ago. 2004.
- GUIMARÃES, P. E. M.; COSTA, M. C. R. SNPs: sutis diferenças de um código. *Biotecnologia, Ciência & Desenvolvimento*, Brasília, v. 5, n. 26, p. 24–27, Mai. 2002.
- LESK, A. M. *Introduction to bioinformatics*. 2. ed. New York: Oxford University Press, 2005. 378 p.
- MEHNERT, R.; CRAVEDI, K. Public collections of dna and rna sequence reach 100 gigabases. Aug. 2005. Aug. 22, 2005.
- NACHMAN, I.; REGEV, A.; FRIEDMAN, N. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, School of Computer Science & Engineering, Hebrew University, Jerusalem 91904, Israel., v. 20, Aug. 2004. ISSN 1367-4803. Disponível em: <<http://view.ncbi.nlm.nih.gov/pubmed/15262806>>.
- NCBI. *dbEST: database of “Expressed Sequence Tags”: Summary by organisms*. Nov. 2006a. http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html. Nov. 24, 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html>.

NCBI. *GenBank overview*. Sep. 2006b. Sep. 26, 2006. Disponível em: <<http://www.ncbi.nlm.nih.gov/Genbank/index.html>>. Acesso em: 30 nov. 2006.

NCBI. *GenBank statistics*. Mar. 2006c. http://www.ncbi.nlm.nih.gov/Genbank/genbank_stats.html. Mar. 7, 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/Genbank/genbank_stats.html>. Acesso em: 30 nov. 2006.

PENA, S. D. J. et al. Retrato molecular do Brasil. *Ciência Hoje*, v. 27, n. 159, p. 16–25, Abr. 2000.

PERKINS, T. J.; HALLETT, M.; GLASS, L. Inferring models of gene expression dynamics. *Journal of Theoretical Biology*, McGill Centre for Bioinformatics, McGill University, 3775 University St. Montreal, Quebec, Canada H3A 2B4. perkins@mcb.mcgill.ca, v. 230, n. 3, p. 289–299, Oct. 2004. ISSN 0022-5193. Disponível em: <<http://dx.doi.org/10.1016/j.jtbi.2004.05.022>>.

PRODOCIMI, F. et al. Bioinformática: manual do usuário. *Biotecnologia, Ciência & Desenvolvimento*, Brasília, v. 5, n. 29, p. 12–25, Nov. 2002.

PRODOCIMI, F.; SANTOS, F. R. Sobre informática, genômica e ciência. *Ciência Hoje*, v. 35, n. 209, p. 54–57, Out. 2004.

RNCOS. *Bioinformatics Market Update*. [S.l.], July 2006. Product code: R459-779.

SETUBAL, J. C. Biologia computacional: panorama da bioinformática. Notas de aula de Biologia Computacional. 2003.

WIKIPEDIA. *Bioinformatics*. July 2006a. <http://en.wikipedia.org/wiki/Bioinformatics/>. July 20, 2006. Disponível em: <<http://en.wikipedia.org/wiki/Bioinformatics/>>. Acesso em: 20 jul. 2006.

WIKIPEDIA. *Lei de Moore*. Jun. 2006b. http://pt.wikipedia.org/wiki/Lei_De_Moore/. 27 de Jun. de 2006. Disponível em: <http://pt.wikipedia.org/wiki/Lei_De_Moore/>. Acesso em: 15 jul. 2006.