Desvendando o Código Genético

Um quebra-cabeça que começa a ser montado

Newton Portilho Carneiro
Ph.D. Biologia Molecular
newtonc@cnpms.embrapa.br
Embrapa Milho e Sorgo - Sete Lagoas, MG

Andréa Almeida Carneiro
Ph.D. Biologia Molecular
andreac@cnpms.embrapa.br
Embrapa Milho e Sorgo - Sete Lagoas, MG

Cláudia Teixeira Guimarães
D.S. Biologia Molecular
claudia@cnpms.embrapa.br
Embrapa Milho e Sorgo - Sete Lagoas, MG

Edilson Paiva
Ph.D. Biologia Molecular
edilson@cnpms.embrapa.br
Embrapa Milho e Sorgo - Sete Lagoas, MG

Os genes são as unidades biológicas responsáveis em determinar as características de um organismo. Apesar de atualmente conhecermos como as informações contidas nos genes são codificadas em proteínas, muitas dessas proteínas / genes não possuem uma função conhecida. Como descobrir a(s) função(ões) de uma proteína codificada por um determinado gene? Proteínas podem ter funções enzimáticas, estruturais ou de reserva, a que iremos nos referir nesta discussão, como funções bioquímicas. Contudo, funções bioquímicas estão encaixadas em um contexto mais amplo, como por exemplo: proteínas participam de processos de regulação celular, de defesa, de tolerância a estresses, etc, os quais serão referidos neste texto como função biológica de uma proteína. Este artigo tem por objetivo descrever alguns dos mecanismos utilizados para o estudo da função bioquímica e biológica de proteínas codificadas por um gene ou em grande escala, por grupos de genes.

Seqüenciamento de Genes – Projetos Genomas

Existem basicamente dois tipos de projetos genoma. Um chamado estrutural que é o seqüenciamento total do genoma, e outro funcional, que se baseia no seqüencimento apenas dos genes expressos. A estratégia mais utilizada para genomas estruturais é a chamada "shotgun", que é uma seqüência em grande escala de subclones de fragmentos de DNA já mapeados. Estudos estruturais de genomas estruturais têm a principal vantagem de combinarem o seqüenciamento e o mapeamento físico dos genes, mas a desvantagem de um elevado custo. Considerando esse aspecto, está sendo feito o seqüenciamento total do genoma, apenas de organismos com genomas pequenos ou daqueles que já têm grande financiamento. Por outro lado, o genoma funcional é baseado apenas no seqüenciamento dos genes expressos e tem a vantagem de poder caracterizar a expressão temporal e local dos genes. O seqüenciamento no genoma funcional é feito em cerca de 800 pares de bases (bp) de apenas uma das fitas de cDNA. Considerando que as bibliotecas de cDNA são montadas direcionalmente, a maior parte do seqüenciamento é feito a partir da extremidade 5' do cDNA, devido ser a mais informativa e conservada. Embora possa haver erros de leitura nas seqüências geradas, estes não comprometem, na maioria dos casos, a identificação dos correspondentes genes. Desta forma, milhares de seqüências podem ser determinadas com um limitado investimento. A terminologia do inglês para etiqueta dos genes é chamada de "Express Sequence Tags" (ESTs).

O banco de dados de ESTs tem mostrado ser uma grande fonte de identificação, principalmente da função bioquímica de muitos genes e de funções biológicas que podem ser inferidas com base na freqüência de certos genes, que são identificados em bibliotecas de cDNAs construídas sob diferentes condições. Exemplificando: uma proteína X pode ser identificada como uma quinase (função bioquímica) através do seqüenciamento do gene que a codifica e da comparação com o banco de dados. Se essa mesma proteína for identificada apenas em tecidos ou orgãos que sofreram o ataque de um patógeno Y , esse fato pode levantar a suspeita de que a proteína

X esteja relacionada com os processos de defesa contra o organismo Y (função biológica).

Atualmente (Outubro/2000), mais de 5 milhões de seqüências, correspondendo a genes de vários organismos estão disponíveis em bancos de dados públicos. Os projetos genomas, no início dos anos 90, começaram seqüenciando principalmente genes abundantes e indicavam que cerca de 60% das seqüências eram desconhecidos. Como já era de se esperar, pesquisadores das mais diversas áreas foram depositando informações de seqüências e suas funções no banco de dados. Para muitos desses genes, as funções foram determinadas pela utilização da combinação de uma série de estudos que envolveu caracterização de mutantes, níveis e localizações da expressão gênica, modificações de substratos específicos, hibridação *in situ*, mapeamento, interação proteína-proteína *in vitro* e *in vivo* entre outros. Essas seqüências gênicas, cujas funções já estão determinadas, servem de suporte para a identificação da função gênica de novas seqüências depositadas em bancos de dados. Quantos genes são desconhecidos hoje, comparando-se com o início dos anos 90? A comparação de seqüências pode ser feita a nível de nucleotídeos, aminoácidos ou de domínios funcionais. As análises de comparação são feitas enviando-se a seqüência editada para o banco de dados e os resultados são devolvidos em uma forma chamada de "e value". Quanto menor esse valor, maior a similaridade da seqüência com aquela presente no banco de dados. Cabe ao pesquisador responsável determinar qual é o limite mínimo para considerar que uma seqüência está qualitativa ou quantitativamente representada no banco de dados públicos. Apesar do sistema de bioinformática estar bastante sofisticado nesses projetos, é difícil para um computador indicar o que pode ser considerado um novo gene, um membro de uma família gênica ou variações de um alelo. Hoje, projetos genomas descrevem que para um "e value" de 10^{-25} existem cerca de 35% de genes desconhecidos.

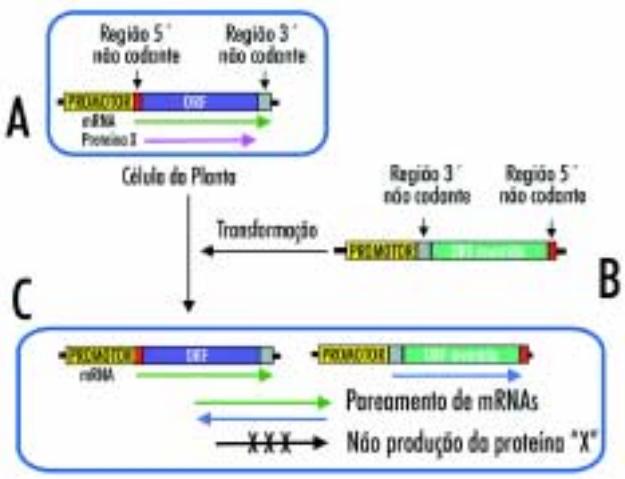

Figura 1 - Mutação por antisenso. A – Célula não mutante contendo o gene "X"; B – Construção contendo o gene "X" na orientação antisenso; C - Célula "A" transformada com construção "B". Célula da Planta Mutante na Proteina X

Existe um grande número de projetos genomas sendo desenvolvidos no mundo. Um resumo dos ESTs depositados em banco de dados públicos pode ser acessado no endereço internet http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html. No Brasil, projetos genoma têm sido principalmente realizados no Estado de São Paulo, com o auxílio da Fapesp (Fundação de Apoio à Pesquisa, do Estado de São Paulo).

Como um seqüenciamento *per si* pode ser útil na determinação da biologia de um gene? Dependendo do número de clones seqüenciados e das variedades das bibliotecas de cDNAs construídas, é possível tirar conclusões relacionadas à abundância, expressão temporal e espacial de muitos genes. Apesar de muitos desses seqüenciamentos serem feitos a partir do final 5' do gene, é possível ter toda a região codante do gene devido à presença de clones de cDNAs derivados de mRNA de diferentes tamanhos. A bioinformática pode reconhecer um bom seqüenciamento, retirar regiões dos vetores, submeter automaticamente a análise do "*Blast*" e montar "*contigs*" ou grupos de seqüências que tenham "*overlaps*" e, com isso, caracterizar membros de uma família, alelos, "*single nucleotideo polymorphism*" (SNPs) etc.

Nem sempre os projetos genomas públicos descrevem as seqüências desconhecidas. A vantagem de um banco de dados disponibilizar as seqüências desconhecidas é a oportunidade que um pesquisador poder comparar as expressões de um gene em outros organismos e de sugerir prováveis funções. As seções a seguir descrevem processos usados para auxiliar a melhor descrição da(s) função(ões) gênica(s).

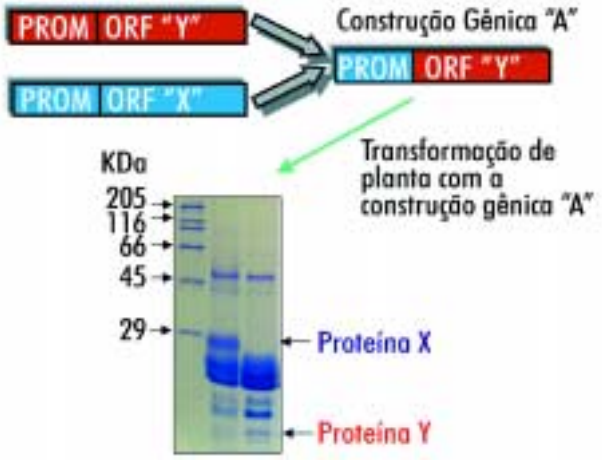

Figura 2 - Co-supressão de promotor. O experimento de SDS-PAGE demonstra que a proteína "X" não existe mais na planta transgênica contendo a construção gênica

Genética Direta e Reversa

Genética direta é um processo que envolve, inicialmente, a análise do fenótipo e depois o isolamento e a identificação do gene responsável pela característica. Genes têm sido isolados por clonagem posicional, mutagênese com inserção de transposons ou de T-DNA, por escrutínio de bibliotecas de

DNA expresso e por técnicas de expressão diferencial (*differental display*).

Com o aumento da capacidade de clonar, modificar e examinar as atividades biológicas de segmentos de DNA, a caracterização de um gene pode também ser realizada pela utilização de uma rota inversa, denominada de genética reversa. Esse novo processo parte de um gene cuja estrutura molecular é conhecida e procede por explorar a contribuição do gene para um determinado fenótipo. Assim sendo, um caminho experimental parte do gene como uma seqüência de nucleotídeos para a sua função correspondente. Sendo a seqüência do gene conhecida, sua função pode ser desvendada pela sua inativação por meio da recombinação homóloga com uma seqüência defeituosa, ou pela diminuição da sua expressão por técnicas de antisenso ou de co-supressão.

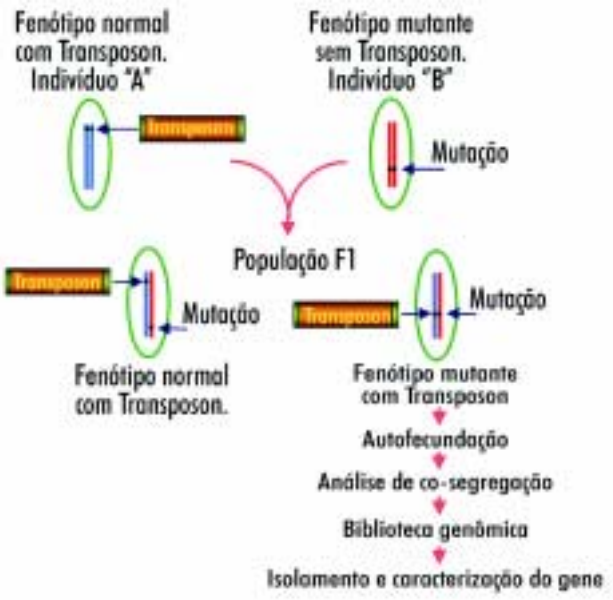

Figura 3 - Identificação de um gene utilizando transposon por meio de genética direta. A grande maioria dos indivíduos F1 do cruzamento entre um indivíduo "A" contendo um transposon ativo com o indivíduo mutante de interesse "B" serão normais devido à complementação do alelo normal da linhagem "A", contudo, em uma freqüência baixa na população F1, o transposon estará inserido no locus do indivíduo "A" corresponde ao gene mutado do indivíduo "B". Nesse caso, o indivíduo F1 será semelhante ao indivíduo "B". Dessa forma o alelo mutado do indivíduo "B" pode ser substituído pelo transposon, um marcador conhecido, através de cruzamentos, e o gene de interesse identificado através da construção de uma biblioteca genômica

Técnicas de Genética Reversa Usadas para Manipular Genes e para Produzir Organismos Mutantes

Mutação por Perda da Função Gênica

Nesse próximo seguimento demonstraremos como técnicas de recombinação homóloga, antisenso e co-supressão podem ser empregadas para mutar um gene conhecido, fazendo com que ele se torne não funcional e a importância das plantas transgências como importantes ferramentas cientifícas para auxiliar na identificação da função gênica.

Recombinação Homóloga

A recombinação homóloga é a alteração de uma pequena parte da seqüência de um gene de interesse que geralmente é feita com a incorporação de um gene marcador e a reintrodução desse gene mutado no organismo de origem. Uma vez dentro do organismo de interesse, o gene mutado tem a capacidade de substituir o gene nativo pela recombinação homóloga, criando um organismo mutante. Um gene marcador pode ser o gene *bar* de seleção para o herbicida fosfinotricine (PPT). As plantas contendo o gene modificado com o marcador são selecionadas na presença do herbicida e, após ficar demonstrado a incorporação do gene modificado no genoma da planta transformada, essa é autofecundada e, por estudos moleculares, é possível identificar a planta que tenha apenas o alelo modificado (homozigose). A planta homozigota para o gene modificado de interesse pode apresentar ou não um fenótipo aparente, mas as plantas com os fenótipos mutantes evidentes podem proporcionar uma forma rápida de correlacionar o fenótipo com o gene modificado. Métodos de distúrbio da função gênica, via recombinação homóloga têm sido descritos para leveduras e ratos, sendo poucos os casos descritos em plantas. Uma aplicação com sucesso da recombinação homóloga, em plantas foi a demonstrada para o gene AGL MADS-box, em *Arabidopsis*.

Antisenso

A metodologia do antisenso envolve a introdução, na célula, de moléculas de RNA ou de DNA, construídas artificialmente, que sejam complementares (antisenso) ao RNA mensageiro (mRNA) do gene de interesse. Uma das hipóteses para explicar o motivo pelo qual o RNA antisenso pode causar mutações no organismo transformado é o fato de que estas moléculas de RNA ou de DNA artificiais se ligam ao mRNA celular, inativando-o (Fig. 1). Usando essa tecnologia, o antisenso do gene da *chalcone sintase* (*chs*), responsável pela pigmentação das flores, foi introduzido em plantas de petúnia e de tabaco, resultando em plantas com alteração na pigmentação das suas flores (fenótipo mutante). Sheehy *et al.* (1988) também usaram a tecnologia do antisenso para inibir a produção de enzimas responsáveis pelo desenvolvimento do fruto do tomate, produzindo, assim, um tomate mutante com o amadurecimento retardado.

Co-supressão

Na técnica de co-supressão, um gene de interesse é engenheirado por meio de técnicas de biologia molecular, para superexpressar uma proteína de interesse. Uma vez que a célula possui uma maquinaria minuciosamente ajustada, qualquer alteração nos níveis de expressão de uma proteína produz uma grande confusão

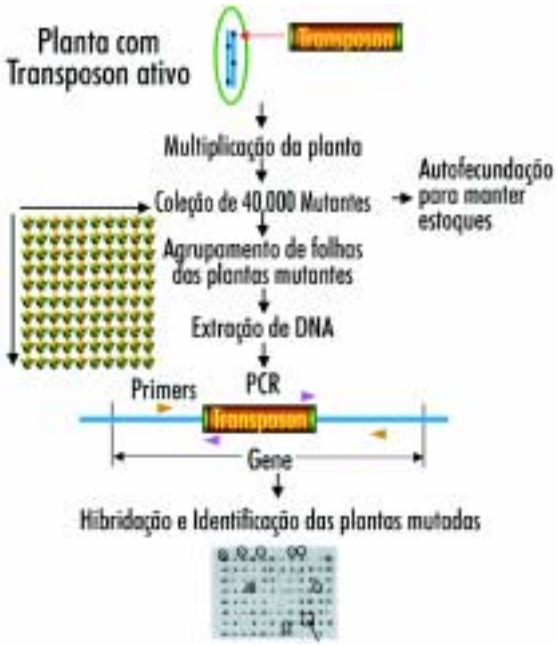

Figura 4 – TUSC (Trait Utility System in Corn). Plantas contendo transposons ativos são multiplicadas para a montagem de uma biblioteca de mutantes. Conhecendo a freqüência com que o transposon se multiplica e se insere em regiões codantes, pode-se calcular teoricamente o número de plantas necessárias para se ter um transposon em cada gene. Cada planta mutante é autofecundada e as sementes estocadas. Uma reação de PCR é feita com DNA extraído de folhas de grupos de plantas dessa biblioteca de mutantes, usando-se um primer do transposon e um primer do gene (a seqüência de ambos é conhecida). A hibridação é feita para auxiliar no escrutínio de um grande número de plantas. Os grupos de plantas cujo sinal foram positivos são subdivididos até que seja encontrada a planta que contenha o transposon inserido no gene. Para aumentar a chance de encontrar a planta mutante, testa-se uma série de *primers* de um único gene simultaneamente

nos mecanismos de regulação celular. Como conseqüência, na maioria das vezes, a superexpressão de uma proteína na célula faz com que a produção dessa proteína seja desligada e, ao contrário do que seria esperado, forma-se, então, um organismo mutante, que não produz a proteína de interesse. Isso pode acontecer a nível de gene ou a nível de promotor. Um exemplo de co-supressão a nível de promotor é demonstrado na Figura 2 onde um mutante foi produzido não na proteína que estava sendo superexpressa, mas na proteína de onde foi retirado o promotor. Embora seja difícil compreender como a superexpressão de um gene possa ocasionar a diminuição da síntese de sua proteína, vários experimentos têm sido realizados demonstrando a ocorrência desse fato.

O primeiro resultado de co-supressão foi obtido por meio de um estudo envolvendo a variação da coloração de flores de petúnia pela introdução do gene da *chs* sob o controle do promotor 35S.

Mutação por Inserção de Transposon ou T-DNA

Transposons

Transposons são elementos de DNA que têm a capacidade de sair de uma região do genoma e se incorporar em outra. Sendo o movimento do transposon um processo aleatório, quando estes elementos "pulam" em um genoma, podem se inserir no meio de um gene, tornando-o inativo. Em plantas, uma série de transposons têm sido usados como ferramenta, tanto na genética direta quanto na reversa, para isolamento e caracterização de vários genes ou fenótipos. O princípio da técnica está descrito na Figura 3. O primeiro gene de planta clonado por transposon, via genética direta, foi o gene "bronze", de milho, que codifica UDP-glucose:flavonóide 3-*O*-glucosiltransferase, uma enzima da via metabólica das antocianinas (Fedoroff *et al.*, 1984).

A genética reversa usando mutantes por inserção de transposons foi descrita, inicialmente, em *Drosophila melanogaster*. Essa metodologia é baseada em uma reação em cadeia da polimerase (PCR – Polymerase Chain Reaction), que utiliza um *primer* complementar ao final do transposon e outro complementar ao final do gene. Dessa forma, os produtos de Polimerase Chain Reaction (PCR) só serão obtidos se um transposon estiver inserido no gene de interesse. Genes que tiveram sua expressão alterada pela inserção do transposon são recuperados por meio da amplificação por PCR do DNA extraído do indivíduo com fenótipo mutante. Esse processo tem sido utilizado na identificação de genes em *Caenorhabditis elegans* e milho.

A família dos transposons *Mutator* (*Mu*) em plantas apresenta altas taxas de mutações e tem alto grau de conservação nas extremidades das seqüências invertidas. Essas duas características são bastante interessantes, pois ajudam a selecionar alelos que contenham os elementos *Mu* cuja seqüência é previamente conhecida. Inserções dos elementos *Mu* podem ser identificadas pela amplificação por PCR. O gene mutado pode, então, ser recuperado e propagado em sementes F_2. Outros aspectos importantes desse processo, e essenciais para a análise de um menor número de plantas, são o número de transposição e o número de cópias do transposon *Mu* na planta. A tecnologia de caracterização de funções gênicas através do *Mu* foi utilizada pela primeira vez em milho pela Pioneer Hi-Bred Co., sendo denominada *Trait Utility System* (TUSC) (Fig. 4). Bensen *et al.* (1995) utilizaram a técnica do TUSC para caracterizar o mutante *Anther ear1* (*An1*), cujo produto gênico está envolvido na síntese do primeiro intermediário tetracíclico na via da biossíntese de giberelina (GA). A mutação *An1* resultou em um fenótipo responsivo à GA, que inclui uma altura reduzida de planta, atraso na maturidade e desenvolvimento de flores perfeitas em espigas normalmente pestiladas.

Um projeto financiado pela *National Science Foundation* (NSF) coordenado pela Dra. Virginia Walbot (Universidade de Stanford, EUA) tem por objetivo demonstrar a funcionalidade de genes de milho usando dois métodos complementares: o seqüenciamento de DNA genômico flanqueando as inserções do transposon *Mu* e a identificação e caracterização dos indivíduos mutantes contendo esses

transposons. Um banco de mutantes está sendo criado usando plantas transformadas com os elementos *Mu* contendo o plasmídeo Bluescript (Fig. 5). A nova inserção contendo o transposon poderá ser seletivamente clonada direto em *E. coli*, gerando uma biblioteca de mutação insercional para análise de DNA. Cada planta F_2 será autofecundada e as sementes serão estocadas no *"Maize Genetics Cooperative Stock Center*", um órgão especialmente criado para armazenar os estoques de sementes mutadas. Os usuários poderão utilizar a técnica de PCR para seleção de uma coleção de plasmídeos que foram inseridos em genes de interesse. Cerca de 50.000 ESTs, flanqueadas pelo transposon *Mu,* já foram completamente seqüenciados durante o primeiro ano do projeto. O objetivo final é seqüenciar cerca de 150.000 segmentos de DNA genômico contendo inserções do transposon Mu.

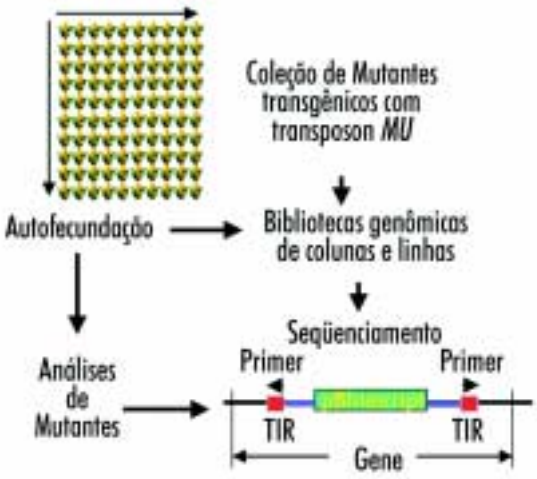

Figura 5 – Projeto NSF coordenado pela Drª. Virginia Walbot (University of Stanford). O transposon contendo uma marca de seleção para resistência à ampicilina em bactéria é transferido para o milho por transformação. As plantas são multiplicadas e são feitas bibliotecas genômicas a partir de colunas e fileiras de plantas contendo esses transposons engenheirados, inseridos aleatoriamente no genoma. Os plasmídeos originados da biblioteca genômica contendo esse cassetes são seqüenciados e as plantas que contêm os transposons inseridos nos genes de interesse são identificadas

T-DNA

O T-DNA é um segmento de DNA presente no plasmídeo Ti (DNA não cromossômico) de *Agrobacterium tumefaciens*, uma bactéria de solo que causa tumores quando infecta plantas. O tumor é causado pela capacidade da bactéria de transferir seu T-DNA para as células vegetais. O T-DNA contém genes que codificam hormônios e aminoácidos necessários à sobrevivência da Agrobactéria dentro da planta. Cientistas descobriram que podiam alterar este fragmento de DNA substituindo os genes nativos por qualquer gene de interesse e, que esses novos genes continuariam sendo transferidos para as plantas pelo sistema mediado pela bactéria. A partir dessa descoberta, o T-DNA passou a ser usado como uma ferramenta na genética direta e reversa de maneira semelhante aos transposons. O T-DNA se insere aleatoriamente no meio de genes tornando-os não funcionais e produzindo um organismo mutante. A função do gene *Agamous* de *Arabidopsis* foi caracterizada como sendo um regulador transcricional necessário para o desenvolvimento floral utilizando essa metodologia (Yanofsky *et al.*, 1990).

Genética reversa em *Arabidopsis* usando grandes populações de plantas mutadas, com um elemento de inserção, taes como o T-DNA de *A. tumenfaciens.* Grandes populações de plantas mutantes foram geradas para constituir uma biblioteca de inserção. Em *Arabidopsis,* uma unidade de transcrição tem em média 2.5 kb (intron e exon), o que significa que um genoma de 100 Mb pode ser dividido em cerca de 40.000 genes. Para ter 95% de chance de acertar um dos genes, são necessárias 120.000 inserções aleatórias independentes. O DNA é extraído de plantas mutagenizadas e agrupadas em grupos maiores. Inserções simples podem ser detectadas por PCR em grupos de cerca de milhares de indivíduos. Dependendo da natureza da população usada, grupos ou supergrupos podem ser organizados em matrizes de duas ou três dimensões para facilitar a determinação final do indivíduo carregando a inserção desejada. As reações de PCR são realizadas em supergrupos de DNA usando-se combinações de oligonucleotídeos do gene e do elemento de inserção. Os produtos de PCR são carregados em um gel de agarose, transferidos para membrana e hibridizados com sondas produzidas a partir do gene de interesse e do elemento de inserção. Apesar do PCR apenas permitir a amplificação do gene mutado pelo elemento de inserção, "backgrounds" podem ocorrer. Desta forma, apenas as bandas que hibridizarem com ambas sondas serão levadas adiante. Uma vez que o produto de PCR tenha sido confirmado e seqüenciado, o supergrupo pode ser subdividido em grupos cada vez menores. A determinação final da linha mutante pode ser obtida em menor tempo, dependendo do esquema de agrupamento dos DNAs.

Uso de Microarrays de DNA para Estudar a Expressão de Genes em Todo o Genoma de Um Organismo

O *microarray* é uma metodologia utilizada para comparar a expressão de um grande número de genes, simultaneamente. Essa técnica emprega arranjos (*arrays*), que contêm um grande número de genes roboticamente distribuídos de forma ordenada sobre placas de vidro. A quantificação dos níveis de expressão na tecnologia de *microarray* é baseada em experimentos onde os milhares de clones de cDNA são hibridizados com duas sondas marcadas com diferentes fluorescências (geralmente uma que emite cor vermelha e outra, verde). As sondas podem ser conjuntos de cDNAs gerados a partir de células ou de tecidos em duas situações diferentes que se deseje comparar (por exemplo: resistência e suscetibilidade ao alumínio). Os resultados são produzidos sob forma de diferentes intensidades de fluorescência, que são captadas por microscopia de fluorescência

a laser, em função dos diferentes níveis de expressão de cada gene. A imagem dos pontos fluorescentes é processada por computadores e programas específicos, sendo gerada simultaneamente uma grande quantidade de informação (Fig. 6).

A tecnologia de *microarrays* não fornece apenas informações sobre a função de genes anônimos, o que favorecerá bastante os processos de genética reversa, mas também constitui uma ferramenta indispensável para estudos globais de expressão gênica, com grandes aplicações nos estudos de biologia molecular e fisiologia vegetal. Apesar do *microarray* ser um dot blot de RNA, grande número de conclusões podem ser tiradas com o uso desse processo.

Uma importante aplicação da tecnologia de *microarray* é o fato que muitos arrays podem ser produzidos para servir como uma plataforma comum entre vários pesquisadores. Se aclopado para o desenvolvimento de um banco de dados centralizado, os pesquisadores serão capazes de pesquisar em multiplos grupos de dados diferentes padrões de expressão de interesse. Com essa tecnologia, será possível acessar o impacto de específico tratamento, fatores ambientais, estágios de desenvolvimento e efeitos na expressão global de todos os genes em um transgênico, estudos envolvendo heterosis e produtividade, análise comparativa de organismos de genomas menores, como vírus, melhoramento assistido por marcadores, *fingerprinting* e escrutínio de germoplasma para identificar genes envolvidos em processos específicos, entre outros. Devido a informação gerada por estudos de *microarray*s ser quantitativa, mudanças súbitas na expressão de um gene podem ser detectadas e podem substituir metodologias de subtração e "differential display".

Apesar da obtenção de "cDNAs" de organismos procariotas ser difícil devido principalmente à falta do poli A+ e à alta instabilidade dos mRNAs, esses organismos ainda são um excelente sistema para análise de *microarray* devido principalmente ao seu genoma pequeno (em torno de 3 Mb) e a grande variabilidade de mutantes disponíveis. Estudos de *microarray*s podem ser usados para comparar o organismo selvagem com mutantes, em uma série de fatores específicos.

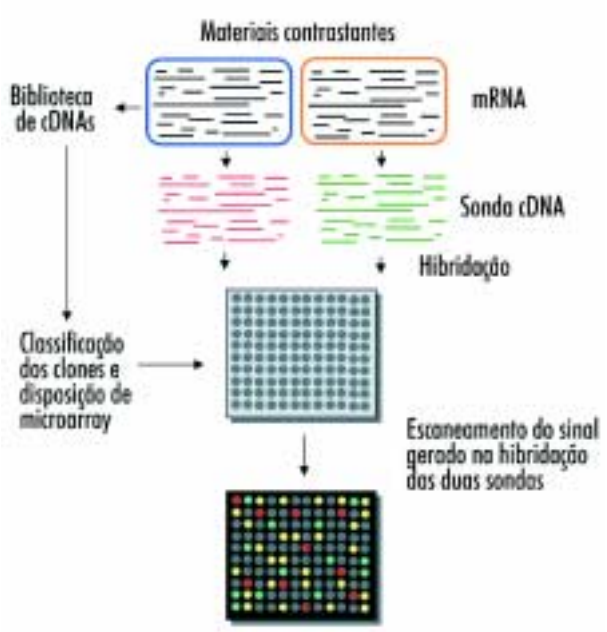

Figura 6 - Análise de expressão gênica usando *microarray*. O mRNA total de cada situação é usado para preparar sondas de cDNA usando a transcriptase reversa na presença de um nucleotídeo marcado com fluorescência. Um dos grupos de mRNA é marcado com um nucleotídeo com fluorescência verde e o outro, com fluorescência vermelha. As duas sondas são misturadas e hibridizadas com o DNA no *microarray*. Assim, a relativa abundância de cada mRNA é comparada por um analisador de imagem através do sinal gerado pelas duas sondas. O objetivo do processo é identificar genes cujos sinais gerados foram mais voltados para o verde ou para o vermelho. Aqueles clones, cujo mRNA não são diferencialmente expressos entre as duas populações de mRNA, terão uma cor intermediaria entre verde e vermelho, aqui representada pela cor amarela. Os clones representados pela cor cinza representam aqueles que estão representados em pequenas quantidades nas sondas. Toda a análise de imagem é feita em um computador que calcula intensidade de sinal gerado na hibridação

A análise de expressão gênica em grande escala oferece grandes vantagens, porém esse processo possui limitações. Uma clara limitação na aplicação dessa tecnologia é a grande quantidade de RNA necessária durante a etapa de hibridação. A quantidade de RNA total para uma hibridação é de 50 a 200 μg (2 a 5 μg para mRNAs). Certamente, os resultados gerados por esse processo não determinam a função de um gene, mas fornecem uma forte ferramenta para a sua compreensão. Os resultados fornecidos por esse processo servem para indicar genes candidatos interessantes para estudos mais detalhados. Estudos adicionais terão de ser feitos para verificar se os níveis de transcrição alterados refletem mudanças na síntese ou "turnover". Além disso, uma resposta diferente pode estar relacionada com processos pós-transcricionais como fosforilação, metilação, glicolisação, etc.

Devido a essa grande capacidade de análise de transcritos ao mesmo tempo, a tecnologia de *microarray* pode, muitas vezes, antecipar os projetos genomas de seqüenciamento estrutural e funcional. Os *microarrays* hoje são construídos a partir de clones conhecidos. Essa etapa envolve seqüenciamento de genes, identificação de clones em placas de estoque, manipulações desses clones para novas placas de estoque, reações de PCR e eletroforese em gel de agarose para confirmar a eficiência do processo, etc. Uma nova idéia que tem surgido é o uso de *microarray*s sem o seqüenciamento prévio. Clones de bibliotecas normalizadas são submetidos a amplificação do inserto por PCR e automaticamente transferidos para o *microarray*. Assumindo que a quantidade de clones raros tenha aumentado em relação aos abundantes na normalização da biblioteca, e que, em cada lâmina de vidro, podem ser organizados pelo menos 10.000 clones, o processo pode tornar-se bem mais simples e interessante. Clones relacionados com o estresse poderiam ser isolados com base na confecção de uma biblioteca de cDNA de raiz de

uma planta tolerante ao estresse, e sondas de cDNAs da planta tolerante e da planta susceptível ao mesmo estresse marcados com duas fluorescências diferentes. Um robô de *microarray* é capaz de colocar 16 amostras em 48 placas de vidro, lavar e secar a sonda para o próximo grupo de cDNA em 70 segundos. Assim, cerca de 10.000 amostras podem ser impressas em, aproximadamente, 12 horas. Vários sistemas robotizados mais rápidos e eficientes para a impressão e processamento dos cDNA nas placas de vidro, além de metodologias cada vez mais sensíveis e precisas para detecção e análise dos resultados têm sido desenvolvidos e disponibilizados. No exemplo anterior, apenas clones expressos diferencialmente seriam seqüenciados. Portanto, mesmo que tenhamos clones repetidos, estaremos com um número de clones para seqüenciamento bastante reduzido em relação a um seqüenciamento aleatório. Se na prática de *microarrays* pode, ou poderá em um futuro próximo, ser usado sem auxílio de "classificação" prévia dos clones, grupos que tenham, ou estejam desenvolvendo, organismos contrastantes para diversas caraterísticas de interesse, estarão com excelentes ferramentas para ajudar a identificar os genes envolvidos em vários processos fisiológicos. Linhas recombinantes provinientes de cruzamentos de indivíduos contrastantes e mutantes são as melhores opções, atualmente, para as análises de *microarray*s.

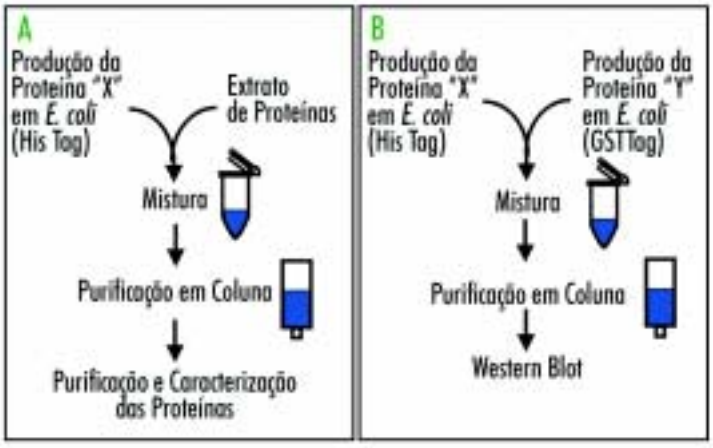

Figura 7 - Identificação de interação proteína-proteína (*in vitro*).
A) A proteína de interesse é clonada em um vetor de expressão em bactéria que permite a sua purificação em uma coluna de afinidade. Um extrato de proteínas produzido nas células vegetais é também submetido à coluna que já contém fixada a proteína de interesse. A coluna é lavada e as proteínas que interagirem com a proteína de interesse podem ser seqüenciadas ou usadas para produzir anticorpos que serão utilizados em uma biblioteca de expressão de cDNA.
B) Proteínas "X" e "Y" produzidas em *E. coli* são misturadas e passadas através de uma coluna de níquel. Se a proteína "Y" interagir com a "X", ambas ficarão retidas na coluna. O processo pode ser visualizado por anticorpos ou raio-X, caso a proteina "Y" esteja marcada com radioatividade

Vantagens e desvantagens do *microarray* baseado em fragmentos de DNA

Existem basicamente dois processos em que os *microarrays* podem ser confeccinados. Um é a impressão de oligos e o outro, os de framentos de DNA previamente isolados na lâmina de vidro. O *microarray* baseado em oligos, em particular aqueles produzidos por método de fotolitografia, tem alta densidade (250.000 olinucleotideos/cm2) e são mais consistentes nos arrays comparados com os *microarray*s baseados em fragmentos de DNA. O potencial de colocar clones em locais errados pode ser evitado no método do oligo porque são sintetizados *in situ,* baseado na seqüência do gene obtido diretamente do banco de dados. Além disso, devido à seqüência que os hibridiza no *microarray* baseado em oligo ser muito curta (20 a 25 nucleotídeos), as reações de hibridação são muito sensíveis na troca de um simples nucleotídeo, ao contrário do método baseado no fragmento de DNA. Isso faz *microarray*s baseado em oligonucleotideos particularmente apropriados para genotipagem e aplicações de seqüenciamento. A maior desvantagem do *microarray* baseado em oligos é que sua construção depende da disponibilidade de um banco de dados seguro. Ao contrário de *microarray*s baseados em fragmentos de DNA, informações de seqüências não são necessárias para a construção. Em comparação à dot blots de clones arranjados em membranas de filtro, *microarray*s são mais sensíveis, exibindo uma dinâmica mais ampla e permitindo uma análise simultânea de duas sondas complexas. Assim, apesar dos dot blots poderem ser apropriados para pequena escala, apenas *microarray*s baseados em fragementos de DNA podem fornecer análises em escala genômica. Uma desvantagem do processo, similar em qualquer outro método de hibridação, é o fato de que hibridação cruzada entre membros de uma família gênica pode ocorrer, ocasionando uma análise confusa de expressão gênica de membros de uma mesma família.

SAGE (Serial Analysis of Gene Expression)

SAGE é a extensão lógica do seqüenciamento de EST. Um inventário de mRNAs é feito com base no seqüenciamento de cDNAs curtos, clonados em tandem. O tamanho desses cDNAs é o suficiente para identificar genes correspondentes no banco de dados. O padrão de restrição desses genes diferentes está relacionado com a abundância relativa de cada cDNA.

Os padrões gerados pelo SAGE têm sido estudados em humanos e em levedura, contudo têm sido pouco utilizados em plantas. Um pré-requisito para a identificação dos ESTs é a disponibilidade de um banco de dados grande para a espécie estudada. Essa técnica é poderosa, mas não é conveniente para a comparação de muitas amostras diferentes e para o estudo de transcritos raros.

Proteomics – Estudo do Perfil Protéico de Um Organismo

Cientistas têm utilizado o termo "*proteomics*" para descrever as análises do perfil protéico de um organismo. O comportamento de muitas proteínas em um proteoma pode ser monitorado por meio de géis de eletroforese em duas dimensões. Nesse processo, proteínas provenientes de células ou de tecidos em duas condições fisiológicas diferentes (por exemplo, plantas resistentes e sensíveis a uma determinada doença ou condição de estresse) são extraídas, aplicadas em géis e comparadas qualitativa e quantitativamente. As proteínas presentes em apenas um dos géis ou que tenham a sua quantidade alterada são fortes candidatadas para atuarem no controle do processo em estudo. A partir da identificação das proteínas de interesse, pode-se, utilizando-se as técnicas de biologia molecular, isolar sua seqüência gênica e caracterizá-la.

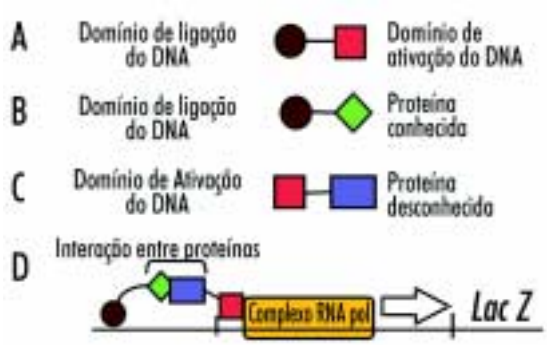

Figura 8 - Interação proteína-proteína *in vivo*. A- O ativador de transcrição GAL4 possui dois domínios: um que se liga ao DNA e outro que ativa a transcrição; B e C - Os dois domínios foram separados: o primeiro pode ser fundido à proteína de interesse (B) e o segundo, à proteína desconhecida (C); D - As colônias de levedura contendo os dois plasmídeos cujas proteínas interagem serão azuis em meio contendo X-GAL (substrato para a enzima), devido à resconstituição do fator de transcrição GAL4 e à ativação do gene da beta-galactosidase (lac Z)

Localização Celular de um mRNA ou Proteína para Inferir sobre sua Função

Os genes são transcritos em moléculas de RNAs mensageiros (mRNA), que são traduzidos em proteínas. Essas são transportadas para locais específicos dentro da célula de um determinado tecido ou organismo. Enquanto alguns genes codificam para proteínas que são expressas em todas as células de um indivíduo durante toda a sua vida (genes constitutivos), outros codificam para proteínas que são expressas apenas em algumas células, ou em um determinado período de desenvolvimento, ou em resposta a algum estímulo externo (genes tecido-específicos). A localização de uma proteína ou de seu mRNA nas células ou tecidos onde são produzidos é uma importante fonte de informação, que pode auxiliar na determinação da função de um gene ou proteína. Um dos princípios dessa metodologia é a identificação do mRNA ou da proteína de interesse por meio de sondas específicas. As sondas após se ligarem à proteína ou ao mRNA em estudo são detectadas com o auxílio de técnicas de microscopia. Uma vez conhecida a localização da molécula dentro de um organismo podem ser programados experimentos mais refinados para a definição de sua função. Técnicas de hibridizações *in situ* têm sido amplamente utilizadas para localizar genes e os seus produtos em cromossomos, tecidos e células de diversos organismos.

Interação Proteína-Proteína

A caracterização da função gênica pode ser auxiliada por meio de estudos de interação proteína-proteína *in vitro* e *in vivo*. A observação de que uma proteína, cuja função é desconhecida, interage com uma que é conhecida pode sugerir que as duas proteínas fazem parte do mesmo processo ou que podem estar localizadas no mesmo compartimento celular. Os dois tipos de testes (*in vitro* e *in vivo*) podem ser usados para verificar a existência de interação entre duas proteínas conhecidas, identificar novas proteínas que interagem com uma proteína em estudo e identificar regiões dentro de uma proteína que são importantes no processo de reconhecimento e da interação. Existe um grande número de alternativas e variações para estudar as interações entre proteínas (Fig. 7). Apesar de bastante informativas, as reações *in vitro* podem mascarar os resultados reais devido à impossibilidade de se reproduzir *in vitro* todas as condições existentes nas células *in vivo*. Por esse motivo, foram desenvolvidos protocolos que estudam a interação entre proteínas *in vivo*.

Uma maneira eficiente para detecção de interação *in vivo* é utilizar o sistema híbrido duplo de levedura. Esse processo constitui a construção de duas proteínas de fusão, por engenharia genética. Uma delas gera um híbrido entre seqüências para um domínio que se liga ao DNA do fator de transcrição Gal4 (aminoácidos 1-147) e a proteína de interesse. Um segundo plasmídeo de expressão contém uma seqüência que ativa o fator de transcrição Gal4 (aminoácidos 768-881) fundida com cDNAs (Fig. 8). Os cDNAs são provenientes de bibliotecas de onde existem genes candidatos. Dessa forma, se as duas proteínas expressas na levedura são capazes de interagir, o complexo resultante ativará a transcrição dos promotores contendo sítios de ligação para o Gal4, gerando uma colônia azul em meio contendo um substrato apropriado (X-GAL). O sucesso dessa metodologia foi primeiramente demonstrado pela interação Gal4-Gal80 (Fig. 8).

Conclusão

Desde a redescoberta das leis de Gregor Mendel, cientistas começaram a questionar a natureza do gene. Que tipo de molécula seria o gene? Como a informação contida em uma molécula poderia ser transmitida para as próximas gerações? Por que e como apareceriam indivíduos mutantes?

Apenas em meados do século XX, com a descoberta da estrutura do DNA por Watson e Crick, essas e muitas outras questões começaram a ser respondidas. Nas primeiras três décadas após a compreensão da estrutura do DNA, o conhecimento a

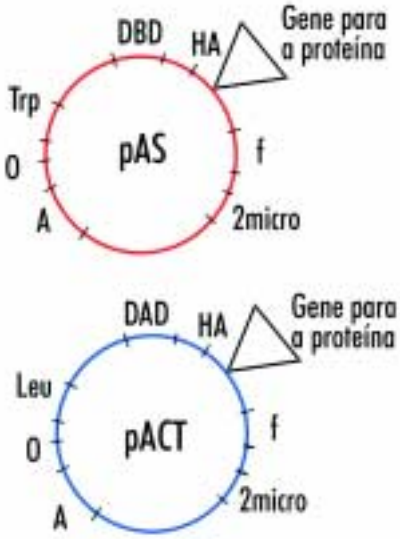

Figura 9 - Esquema dos plasmídios usados no sistema de levedura de híbrido duplo. Os plasmidios pAS2 e pACT2 contêm as proteínas de fusão de ligação no DNA (DBD) e ativação de transcrição (DAD), respectivamente. Os genes Trp e Leu permitem que a levedura cresça em meio sem triptofano e leucina. O gene Ap seleciona os plasmídios em *Escherichia coli.* HA é a proteína hemoaglutinina e pode ser usada como proteína repórter de leveduras transformadas

Figura 10 - Processo de seleção de leveduras contendo os plasmídios pAS2 e pACT2 e seleção das leveduras contendo interação entre as proteínas "X" e "Y"

respeito de sua biologia cresceu fantasticamente. Dentro desse período, ficou conhecida a natureza química dos genes, como a informação genética era armazenada, como as células respondiam a essa informação e como ela era transmitida de uma geração para outra.

A partir dos anos 70, cientistas aprenderam a manipular a molécula de DNA utilizando as técnicas de biologia molecular. Inaugurou-se a era da tecnologia do DNA recombinante, também denominada engenharia genética. Desde então, a biologia molecular tem experimentado grandes avanços tecnológicos que têm culminado em importantes fontes de conhecimento sobre a função, expressão e regulação de genes em diferentes organismos. Trabalhos de isolamento, seqüenciamento e caracterização de um ou poucos genes eram comumente publicados até a década de 90. Atualmente, um único laboratório pode seqüenciar 1000 (ou mais) genes por dia. Um pesquisador pode utilizar os dados gerados e chegar a conclusões muito mais precisas sobre a função de um gene em poucas semanas de trabalho, do que outro, em muitos meses de trabalho, há alguns anos atrás. Hipóteses podem ser testadas muito mais rápida e precisamente. Está havendo um redirecionamento na maneira como os grupos conduzem seus projetos de pesquisas na área da biologia molecular. A primeira fase dos chamados projetos genoma, seqüenciamento de genes, está gerando uma grande quantidade de informações, as quais serão provavelmente, multiplicadas com a implementação da segunda fase, ou seja, com a análise funcional do material seqüenciado. Neste texto foram mencionadas metodologias que têm por objetivo auxiliar na identificação da função gênica tanto em âmbito bioquímico como biológico. Entre elas, os *microarrays* têm revolucionado a análise funcional de seqüências gênicas em grande escala, uma vez que diferenças nos níveis de expressão de milhares de genes podem ser detectados simultaneamente. Desta forma, vários genes ainda desconhecidos poderão ter suas funções biológicas desvendadas utilizando esse processo. Em um futuro não muito distante, ao invés de se comprarem "*kits*" para construção de bibliotecas, serão compradas bibliotecas já arranjadas em matrizes, e o pesquisador testará sondas diferentes. Os resultados serão organizados não mais em ESTs, mas em níveis de expressão desses genes sob diferentes condições. As metodologias de análise gênica, tanto individuais quanto em larga escala, fornecerão um acesso a informações sem precedentes para todas as áreas da biologia. Entre elas, a agropecuária será amplamente beneficiada, uma vez que existe grande necessidade de identificar e de estudar genes que sejam responsáveis por conferir resistência a doenças, tolerância a estresses bióticos e abióticos, aumento da qualidade nutricional, entre outras características de interesse econômico.

Referências

BENSEN, R.J.; JOHAL, G.S.; CRANE, V.C.; TOSSBERG, J.T.; SCHNABLE, P.S.; MEELEY, R.B.; BRIGGS, S.P. (1995). Cloning and characterization of the maize *An1* gene. *Plant Cell* 7: 75-84.

FEDEROFF, N.V.; FURTEK, D.B.; NELSON, O.E. (1984). Cloning of the *bronze* locus in maize by a simple and generalized procedure using the transposable element Activator (*Ac*). *Proc. Natl. Acad. Sci. USA* 81: 3825-3829

SHEEHY, E.R.; KRAMER, M.; HIATT, W.R. (1988). Reduction of polygalacturonase activity in tomato fruit by antisense RNA. *Proc. Natl. Acad. Sci. USA* 85: 8805-8809.

YANOFSKY, M.F.; MA, H.; BOWMAN, J.L.; DREWS, G.N.; FELDMANN, K.A.; MEYEROWITZ, E.M. (1990). The protein encoded by the *Arabidopsis* homeotic gene *agamous* resembles transcription factors. *Nature* 346: 35-39.