UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS AGRÁRIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DOS ALIMENTOS

Carla Souza de Mello

APLICAÇÃO DE TRANSCRIPTÔMICA E PROTEÔMICA COMO AVALIAÇÃO COMPLEMENTAR DE ALIMENTOS ATRAVÉS DE ANÁLISE MULTIVARIADA

Florianópolis
2014

CARLA SOUZA DE MELLO

Aplicação de Transcriptômica e Proteômica como Avaliação Complementar de Alimentos Através de Análise Multivariada

Tese submetida ao Programa de Pós Graduação em Ciência dos Alimentos da Universidade Federal de Santa Catarina como um dos pré-requisitos para a obtenção do Grau de Doutor em Ciência dos Alimentos.

Orientadora: Profª. Drª. Ana Carolina Maisonnave Arisi.

Florianópolis
2014

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Souza de Mello, Carla
Aplicação de Transcriptômica e Proteômica como Avaliação Complementar de Alimentos Através de Análise Multivariada / Carla Souza de Mello ; orientadora, Ana Carolina Maisonnave Arisi - Florianópolis, SC, 2014.
143 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências Agrárias. Programa de Pós-Graduação em Ciência dos Alimentos.

Inclui referências

1. Ciência dos Alimentos. 2. Transcriptômica. 3. Proteômica. 4. Análise multivariada. 5. Análise de alimentos. I. Maisonnave Arisi, Ana Carolina. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência dos Alimentos. III. Título.

Folha de aprovação

AGRADECIMENTOS

A Deus, pela iluminação e direção do caminho.

À minha amada mãe Lídice Souza, meu modelo de vida, meu porto seguro, que sempre tem a palavra certa na hora certa. Ao meu segundo pai Arnizaut, que sempre me apoiou e me fez rir nas horas mais inesperadas. Ao meu falecido pai, Carlos, que sempre me protege e apoia, de onde quer que esteja. À minha família toda (!), especialmente à tia Arminda e cia, pelo suporte emocional e companheirismo.

I thank my beloved husband, Marco, who is always there for me, on the good and the bad times, either being really close together or far apart, for giving me strength to go on, a perfect advice for my work or simply the best hug in the world; the greatest partner I could ever imagine I would have.

À prof. Dra. Ana Carolina Maisonnave Arisi pelos seis anos de orientação e apoio, compartilhando novas ideias e incentivando o desenvolvimento da pesquisa.

I thank all the co-authors of this work, especially to Esther Kok and Marleen Voorhuijzen, for the possibility of sharing the knowledge and experiences. Special thanks to Jeroen van Dijk, who dedicated considerable time to guide me along the new world of "omics" and multivariate analysis, and encouraged me to pursue my professional development.

Aos amigos do laboratório de Biologia Molecular (CAL, UFSC), com quem aprendi muito e compartilhei inúmeros momentos ótimos dentro e fora do lab, o que foi fundamental para meu equilíbrio emocional e profissional. São minha segunda família. Tenho muito a agradecer a todos: desde o pessoal das antigas (a sorridente, conselheira e amiga Andréia, o Fábio com seus sábios conselhos, a Diana dando o exemplo de organização, Deise, sempre carinhosa e questionadora); assim como aos "proteômicos" (a orientação/diversão/companheirismo constante dos amigos Geisoca e Pedro, o auxílio indispensável da doce Gabrieeela, a tranquilidade da Cibe e por pouco, mas precioso tempo, apoio da Mirella e Naíra); aos amigos-parceria incondicional (Tomás, seeempre bem disposto resolvendo todos os pepinos, as amadas, carinhosas Pam e Ferdi, duas pequenas-grandes companheiras, a Van com seu ombro amigo sempre ali, os companheiros Gustavo, Jojô e Kelly) e a todos os colegas com quem convivi durante estes anos que de alguma forma, seja com vidraria, reagente, uma conversa ou uma parceria pro café ou chimarrão, contribuíram nessa caminhada.

Ao prof. Dr. Miguel Pedro Guerra, à Dra. Gabriela Cangahuala Inocente e outros integrantes do LFDGV (Laboratório de Fisiologia do

Desenvolvimento e Genética Vegetal), UFSC, pelo auxílio e empréstimo de equipamentos para execução de parte das análises proteômicas.

Aos amigos que mesmo de longe (e de perto) tiveram participação em momentos cruciais durante esses quatro anos, entre eles: Lusa, Carol, Elena, Maia, Fezoca, Ale (Nine), Beta, Agata, Victor, Carols, Sandri, Anabele, Vitão, Clarissa. A eles, o meu carinho.

À Universidade Federal de Santa Catarina, ao Departamento de Ciência e Tecnologia de Alimentos e ao Programa de Pós-Graduação em Ciência dos Alimentos, pela oportunidade de realizar este curso.

Ao projeto de cooperação internacional CAPES/Wageningen 005/09, ao Instituto de Segurança de Alimentos (RIKILT) da Universidade de Wageningen (WUR), Holanda, e ao Ministério Holandês de Assuntos Econômicos pelo suporte financeiro para o desenvolvimento de parte deste trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos e ao CNPq (Edital Universal 2012) pelo financiamento do projeto.

Agradeço também ao governo federal e a todos os residentes do território brasileiro que pagam seus impostos em dia e, consequentemente, contribuem para que o governo possibilite o acesso ao ensino superior gratuito no Brasil. Eles indiretamente colaboraram com meus estudos e desenvolvimento profissional, assim como o fizeram para milhares de pessoas nesta e em outras instituições públicas.

"A mente que se abre a uma nova ideia jamais voltará
ao seu tamanho original."

Every now and then a man's mind is stretched by a new idea or sensation and never shrinks back to its former dimensions.

Oliver Wendell Holmes, Sr.
The Autocrat of the Breakfast Table (1858)

RESUMO

A crescente presença de novos produtos alimentícios no mercado desperta discussões relacionadas à segurança de alimentos. Cada país ou região tem suas próprias leis para liberação de novos alimentos para consumo, porém existe um consenso internacional no que diz respeito às regulamentações para segurança do consumo destes produtos, inclusive de alimentos provenientes de tecnologia de DNA recombinante. O conceito de equivalência substancial tem sido utilizado com este fim, fundamentado no fato de que os alimentos já existentes no comércio são admitidos como seguros para o consumo e servem como base para comparação por meio de análises de componentes específicos. Apesar de convenientes, estas análises-alvo são bastante limitadas por pesquisarem a presença de somente alguns elementos previamente conhecidos. Deste modo, abordagens mais completas para avaliação de alimentos têm sido propostas, como análises transcriptômica e proteômica. Como estas análises geram um grande volume de dados, enfoques utilizando análise estatística multivariada têm sido sugeridos para interpretação dos resultados. Neste trabalho verificou-se a aplicação de ferramentas estatísticas multivariadas para análises transcriptômica (microarranjo) e proteômica (eletroforese bidimensional) com vistas à sua utilização como análise complementar na avaliação de segurança de novos alimentos. Para isso, analisaram-se cinco variedades de batatas reconhecidas como seguras para consumo (Biogold, Fontane, Innovator, Lady Rosetta e Maris Piper). As análises do transcriptoma das amostras revelaram que foi possível a classificação das amostras utilizando a ferramenta SIMCA com uma classe. Foram desenvolvidos dois cenários contendo um conjunto de cinco classificadores e, em cada cenário, foram testadas duas amostras independentes sabidamente seguras, porém analisadas em diferentes momentos (incluindo, assim, variabilidade técnica no teste). Em cada conjunto de classificadores, as amostras teste que foram mais vezes classificadas como não pertencentes aos modelos (ou seja, não classificadas como seguras) representam as amostras com maior variabilidade técnica, pois foram cultivadas e analisadas em tempos diferentes daquelas utilizadas para a construção dos classificadores. Já as amostras que foram reconhecidas como seguras na maioria dos classificadores possuem menor variabilidade técnica. Foi também realizada a análise proteômica por eletroforese bidimensional destas amostras. Utilizou-se tiras de gradiente de pH imobilizado (IPG) de dois comprimentos diferentes, 13 e 24 cm, todas na faixa de pH de 4-7, e os conjuntos de dados gerados, representando a porcentagem de volume de *spots* (tendo os valores omissos substituídos ou simplesmente eliminados),

foram visualizados por diagramas de análise de componentes principais (PCA). Foi verificada clara separação das variedades já nos dois primeiros componentes principais do conjunto de dados contendo valores omissos substituídos. Estes resultados revelam a possibilidade de se construir ferramentas de classificação por técnicas de análise ampla de perfil como a transcriptômica e proteômica, explorando assim uma nova abordagem para avaliação de segurança de alimentos. Para aprimorar o trabalho, a análise de um maior número de amostras permitirá maior precisão dos resultados, incluindo-se assim um alto nível de variabilidade técnica na construção dos classificadores. Desta forma, será possível a reprodução em pequena escala de situações reais de avaliação de segurança de alimentos.

Palavras-chave: Transcriptômica. Microarranjo. Proteômica. Eletroforese bidimensional. Avaliação de segurança de alimentos. Quimiometria. PCA. SIMCA.

ABSTRACT

The increasing occurrence of new food products in the market stimulates ever more discussions related to food safety. Each country or region possess their own laws for releasing new foods for consumption, but there is an international consensus regarding the regulations for the safety of consumption of these products, including foods derived from recombinant DNA technology. The concept of substantial equivalence has been used for this purpose, based on the fact that food already commercialized are accepted as safe for consumption and serve as basis for comparison through analysis of specific components recognized as toxic. These targeted analyzes are convenient, but they are rather limited because they search for the presence of only a few elements which are previously known. Thus, more comprehensive approaches such as transcriptomics and proteomics analyses have been proposed for food safety evaluation. Multivariate statistical approaches have been suggested for interpretation of results, given that these analyzes generate a large amount of data. On this study the application of multivariate statistical tools for analysis of data from transcriptomics (microarray) and proteomics (two-dimensional electrophoresis) techniques was verified, aiming its use as a complementary tool in safety assessment of novel foods. For that, five varieties of potatoes recognized as safe for consumption (Biogold, Fontane, Innovator, Lady Rosetta and Maris Piper) were analyzed. Transcriptome analysis of samples showed that it was possible to classify the samples using the SIMCA one class. Two scenarios containing a set of five classifiers have been developed, and each set of two independent samples considered as safe were tested, but analyzed at different time points (including technical variability in the test). In each set of classifiers, the test samples which were most often classified as not belonging to the models (i.e. not classified as safe) represent the samples with higher technical variability, given they were grown and analyzed at different time points from those used to construct the classifiers. However, the samples that have been recognized as safe in most classifiers have lower technical variability. In addition, proteomic analysis using two-dimensional electrophoresis was performed with these samples. Immobilized pH Gradient (IPG) strips pH 4-7 of two different lengths, 13 and 24 cm, were used, and the generated datasets representing percentage of volume of spots (missing values have been replaced or simply removed) were visualized by PCA. Clear separation of the varieties was verified already in the first two principal components of the dataset containing replaced missing values. These results reveal the possibility of building classification tools through profiling techniques such

as transcriptomics and proteomics, thus exploring a complimentary approach for food safety assessment. To improve the work, analysis of an increased amount of samples will enable more accurate results, thus including a high level of technical variability in the construction of classifiers. As a result, it is possible to represent real situations of food safety assessment in small-scale.

Key words: Transcriptomics. Microarray. Proteomics. Two-dimensional electrophoresis. Food safety evaluation. Chemometrics. PCA. SIMCA.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

2-DE	Eletroforese bidimensional
cDNA	DNA complementar
CHAPS	(3-[(3-Colamidopropil)dimetilamônio]-1-propanossulfonato)
cRNA	RNA complementar
CTAB	Brometo de cetiltrimetilamônio
Cy3	Fluoróforo cianina-3
DNA	Ácido desoxirribonucleico
DP	Desvio padrão
DTT	Ditiotreitol
EDA	Análise exploratória de dados (*Exploratory data analysis*)
GM	Geneticamente modificado
IC	Intervalo de confiança
IEF	Focalização isoelétrica
IPG	Gradiente de pH imobilizado (*Immobilized pH gradient*)
LiCl	Cloreto de lítio
MMLV-RT	Transcriptase reversa do vírus da leucemia murina de Moloney
M_R	Massa molecular
PAGE	Eletroforese em gel de poliacrilamida
pI	Ponto isoelétrico
PC	Componente Principal
PCA	Análise de Componentes Principais (*Principal Component Analysis*)
PCR	Reação em cadeia da polimerase (*Polymerase chain reaction*)
PMSF	Fluoreto de fenilmetanossulfonila
r	Coeficiente de correlação
RNA	Ácido ribonucleico
mRNA	RNA mitocondrial
rRNA	RNA ribossomal
SDS	Dodecil sulfato de sódio
SDS-PAGE	Eletroforese em gel de poliacrilamida com SDS
SIMCA	Modelagem Independente Flexível por Analogia de Classes (*Soft Independent Modelling of Class Analogy*)
TCA	Ácido tricloroacético
Tris-HCl	Hidroximetilaminometano e ácido clorídrico

SUMÁRIO

INTRODUÇÃO

A crescente oferta mundial de novas variedades de plantas para o consumo humano tem causado impacto nas formas de controle de segurança de alimentos adotadas pelos órgãos responsáveis. Tradicionalmente, a avaliação de segurança de novos alimentos baseia-se na análise comparativa, que inclui a determinação dos compostos nutricionais e de fatores antinutricionais relevantes em cada espécie vegetal. Os resultados dessas análises são comparados às quantificações desses compostos previamente conhecidos em referências próximas como, por exemplo, em espécies equivalentes, em isolinhas (variedades não transgênicas, no caso de plantas geneticamente modificadas), ou mesmo em dados de literatura, levando em conta a variação composicional natural, desde que esta seja conhecida (RESENDE et al., 2013). No entanto, as informações disponíveis são limitadas devido ao grande número de variáveis envolvidas e a análise comparativa de novos alimentos, derivados de modificação genética ou não, deve ser feita com algumas ressalvas (KOK et al., 2008). A Organização para Cooperação e Desenvolvimento Econômico (OECD) desenvolveu documentos que servem como diretrizes para a investigação de importantes cultivares, resumindo o conhecimento existente sobre estas variedades e seus componentes-chave a serem investigados durante a avaliação de um determinado produto para alimentação humana ou animal (OECD, 2006). Este tipo de avaliação resulta na análise individual de um grande número de compostos gerados em diferentes rotas metabólicas, abrangendo a análise do metabolismo da planta na busca por possíveis efeitos não intencionais na nova espécie. No entanto, outras rotas metabólicas de mesma importância provavelmente não estarão sendo avaliadas.

Neste sentido, as análises amplas de perfil (conhecidas como "ômicas") como a transcriptômica e a proteômica têm sido sugeridas como alternativas mais completas para a avaliação de alimentos, pois fornecem informações mais detalhadas e completas. As análises ômicas geram um grande volume de informações que podem levar a conclusões inclusive sobre os efeitos não esperados nas mais diversas formas de melhoramento de plantas usadas como alimento (DAVIES, 2010).

Assim, novas abordagens para interpretação deste grande volume de dados têm sido exploradas, como análises estatísticas multivariadas. A aplicação destas análises no campo de avaliação de segurança de alimentos pode auxiliar e fornecer informações mais completas sobre o uso seguro de culturas alimentares já presentes no mercado e de novos alimentos (CELLINI et al., 2004; BERRUETA, 2007). Análise de componentes

principais (PCA) foi utilizada para investigar possíveis mudanças não intencionais na composição de mamão papaia transgênico em relação à sua isolinha convencional (JIAO, 2010). Um sistema de classificação multivariada foi analisado por meio de um estudo de caso sobre detecção de fraude em alimentos (LÓPEZ, 2014). Métodos de modelagem de classe também já foram largamente utilizados em estudos de autenticação de alimentos (OLIVERI; DOWNEY, 2012).

Neste trabalho o objetivo foi verificar a possibilidade da utilização de duas técnicas ômicas (transcriptômica e proteômica) como instrumentos complementares de avaliação de alimentos por ferramentas de estatística multivariada utilizando variedades de batatas reconhecidas como seguras para consumo.

Esta tese é apresentada em forma de capítulos, sendo o primeiro capítulo uma revisão da literatura sobre aspectos de segurança de alimentos no Brasil e no mundo. Em seguida, no mesmo capítulo, é feita uma breve abordagem sobre as análises amplas de perfil transcriptômica e proteômica e, finalmente, discorre-se sobre a análise multivariada como forma de avaliação de dados provenientes de análises ômicas em alimentos. Os demais capítulos são apresentados sob a forma de artigos científicos.

No Capítulo 2, é abordada a análise transcriptômica pela técnica de microarranjo como possível instrumento de avaliação de segurança de alimentos pelo desenvolvimento de um sistema de classificação utilizando a ferramenta de Modelagem Independente Flexível por Analogia de Classe (SIMCA) com uma classe. Diferentes variedades de batatas reconhecidamente seguras para consumo foram usadas como base para a investigação. Este trabalho teve o apoio financeiro do Ministério Holandês de Assuntos Econômicos e foi realizado no RIKILT Instituto de Segurança de Alimentos, Wageningen University and Research centre, Holanda, com bolsa de doutorado sanduíche do projeto de cooperação internacional CAPES/Wageningen (Projeto N.005/2009) sob supervisão de Dra. Esther J Kok.

No Capítulo 3, é apresentado um estudo sobre a aplicação da análise proteômica pela técnica de eletroforese bidimensional (2-DE) em dois tamanhos diferentes de gel de poliacrilamida (*strips* de 13 e 24 cm) nas mesmas variedades de batatas. Apresenta-se uma análise de como as variedades são separadas quando os dados de acúmulo de proteínas são plotados em diagramas de PCA tendo seus valores omissos substituídos ou eliminados do conjunto de dados. Este trabalho teve apoio financeiro do CNPq (Edital Universal 2012) e foi realizado no Laboratório de Biologia Molecular, Departamento de Ciência e Tecnologia de Alimentos, UFSC.

Por fim, foram feitas considerações finais sobre os resultados obtidos neste trabalho e as possíveis perspectivas no abrangente campo de análise de segurança de alimentos.

CAPÍTULO 1

REVISÃO BIBLIOGRÁFICA

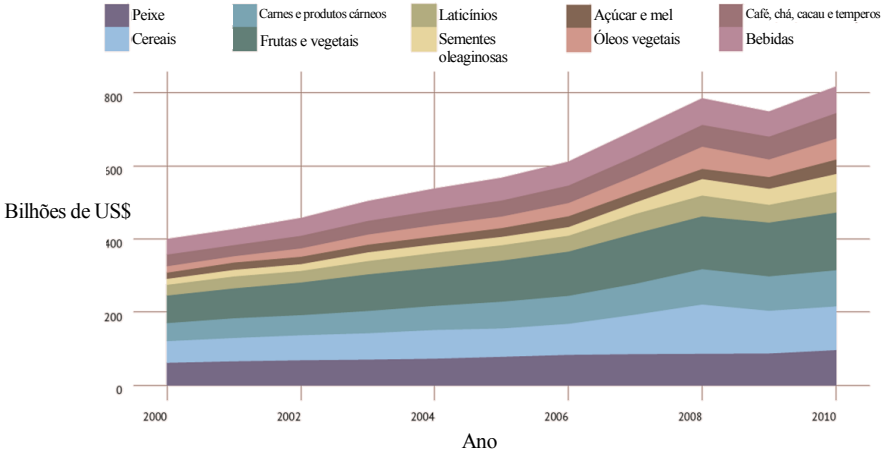
1 REVISÃO BIBLIOGRÁFICA

1.1 SEGURANÇA DE ALIMENTOS

Tópicos relacionados à avaliação de segurança de alimentos têm sido extensivamente discutidos nos últimos anos. Os rumos da produção global de alimentos, assim como seu processamento, distribuição e preparo, apresentam novos desafios para a segurança dos alimentos. Plantas cultivadas em um país podem agora ser transportadas e consumidas do outro lado do mundo. As pessoas exigem hoje uma maior variedade e disponibilidade para consumo do que no passado, além de buscarem alimentos mais saudáveis, seguros e nutritivos. Com essa introdução de novas variedades no mercado, aumentam as preocupações relacionadas aos critérios de liberação destes alimentos, sejam eles provenientes de modificação genética ou não.

Embora a maioria dos alimentos consumidos em todo o mundo seja cultivada localmente, em alguns lugares não há produção suficiente para atender a demanda. Em virtude disso, o comércio global tem sido fundamental para suprir as exigências do mercado, causando avanço constante no comércio agrícola e alimentar. O valor dos fluxos internacionais aumentou cerca de cinco vezes ao longo dos últimos 50 anos, refletindo as tendências mundiais no volume global do comércio (FAO/WHO, 2013). Na figura 1.1 é apresentado um gráfico com o crescimento no volume de exportações de diversos produtos alimentícios de 2000 a 2010, o que caracteriza o aumento do comércio global de alimentos.

Figura 1.1: Composição das exportações globais, entre os anos de 2000 e 2010 (em bilhões de dólares).



Fonte: adaptado de FAO/WHO, 2013.

A globalização do comércio de alimentos oferece benefícios aos consumidores, uma vez que resulta em uma maior variedade de alimentos acessíveis, de alta qualidade e disponibilidade, atendendo às exigências do consumidor. A diversidade de alimentos em uma dieta balanceada melhora o estado nutricional e de saúde. No entanto, essas mudanças quanto à disponibilidade também apresentam novos desafios para produção e distribuição de alimentos seguros, ocasionando amplas repercussões sobre a saúde (RESENDE et al., 2013; WHO, 2002).

A disponibilidade de alimentos seguros, além de beneficiar a saúde, é um direito humano básico. Embora se tenha observado um aumento dos esforços de diversos países para a fabricação de alimentos mais seguros, ainda milhões de pessoas adoecem todos os anos de doenças transmitidas por alimentos. A introdução de novas tecnologias, incluindo a engenharia genética e a irradiação, representa um desafio especial. Algumas novas tecnologias podem ser benéficas para o aumento da produção agrícola e suprir as demandas da população, mas sua eficiência e segurança devem ser demonstradas para que possam ser aceitos pelos consumidores (WHO, 2002).

Dentre as estratégias utilizadas para a geração de novas variedades de plantas que podem ser usadas como alimentos, encontram-se o

melhoramento clássico e a modificação genética. Ambas objetivam causar alterações em rotas metabólicas para que se obtenham plantas com características mais favoráveis, por exemplo, com maior produtividade, adaptadas a diferentes agroecossistemas, resistentes a doenças e pragas ou com melhor qualidade nutricional ou de processamento (CELLINI et al., 2004).

O melhoramento clássico fundamenta-se na diversidade genética das espécies. Se os traços desejáveis já estão presentes em determinada variedade, estes podem ser mantidos nas próximas gerações por cruzamento entre uma linhagem de alto desempenho e outra contendo o traço esperado. No entanto, se a característica almejada não está presente na espécie, mas sim em uma variedade relativamente distante, algumas técnicas de cultura de tecidos podem ajudar o produtor a obter gerações férteis a partir de cruzamentos usualmente estéreis. Em uma terceira situação, em que as propriedades desejadas não se encontram nem em linhagens distantes, pode-se obter linhagens mutantes contendo as características por mutagênese química (KOK et al., 2008). Já a modificação genética ocorre pela inserção de um ou mais genes que causarão a indução ou efetivamente a formação de novas proteínas ou redução da expressão de transcritos já presentes nas rotas metabólicas. Além disso, mais de uma característica pode ser alcançada pelo cruzamento de linhagens geneticamente modificadas contendo diferentes genes (KOK et al., 2008).

O melhoramento de plantas gera modificações esperadas (efeitos intencionais) nas novas gerações, tais como ativação ou silenciamento de genes, alteração dos níveis de determinadas proteínas, formação de novos metabólitos ou alteração no nível de metabólitos já existentes. Entretanto, efeitos não esperados (efeitos não intencionais ou pleiotrópicos) também podem ocorrer, tanto no melhoramento clássico/convencional como em técnicas que utilizam a tecnologia de DNA recombinante. No caso da transgenia, os efeitos não intencionais podem ser parcialmente previsíveis quando se conhece o local específico de inserção do fragmento do DNA recombinante, a função do traço inserido ou seu envolvimento em rotas metabólicas. No entanto, outros efeitos podem ser imprevisíveis devido ao limitado conhecimento sobre a regulação do gene ou as interações intergênicas (KOK; KUIPER, 2003; KUIPER et al., 2001).

Segundo D'Alessandro e Zolla (2012), na segurança de alimentos de origem vegetal deve-se considerar três pontos:

(i) A determinação de biomarcadores proteicos para rastreamento dos alimentos, que poderá identificar a presença de compostos alergênicos em cultivares de regiões específicas;

(ii) A possibilidade de ressaltar possíveis diferenças entre as espécies transgênicas e suas equivalentes convencionais, no caso de transgenia; e

(iii) A ausência de patógenos provenientes de contaminantes bióticos ou abióticos, como contaminantes microbiológicos (toxinas) e químicos (metais pesados).

Geralmente, novos alimentos são considerados seguros desde que cuidados apropriados sejam tomados durante seu desenvolvimento, produção, processamento, armazenamento, manuseio e preparo. Em muitos casos, admite-se que o conhecimento necessário para controlar os riscos associados com alimentos tradicionais já tenha sido adquirido durante o curso de seu longo histórico de consumo (CONSTABLE et al., 2007). Assim, evidências que caracterizem o histórico de uso do produto podem contribuir para sua comprovação de segurança.

O termo "histórico de uso seguro" vem sendo discutido desde o começo da década de 1990, e não existe um conceito especificamente definido. Segundo o Guia Canadense de Novos Alimentos (*Canadian Guidelines on Novel Foods*) (HEALTH CANADA, 2003), entende-se que um alimento tenha um histórico de uso seguro se, durante muitas gerações geneticamente diversas, ele tem sido consumido em largas quantidades e que existem dados toxicológicos e alergênicos apropriados para confirmar que nenhum mal será resultado a partir do consumo deste. Uma vez verificado que tem histórico de consumo seguro, este alimento ou planta tradicional será referência para avaliação comparativa do novo produto. A abordagem comparativa com vistas à avaliação de segurança de um novo alimento é conhecida como "equivalência substancial" (ES) (CONSTABLE et al., 2007).

O conceito de ES foi formulado inicialmente em 1993 pela OECD e serve como uma ferramenta-guia para avaliação de alimentos. No documento, define-se que se o novo alimento é reconhecido como sendo substancialmente equivalente ao alimento já existente, então outras preocupações em relação à segurança ou ao conteúdo nutricional provavelmente serão insignificantes; tais alimentos passam a ser tratados da mesma forma que sua equivalente tradicional. Quando novos alimentos, classes de alimentos ou componentes não são tão conhecidos, torna-se mais difícil aplicar o conceito de ES; eles são então avaliados quanto à experiência obtida nas avaliações de compostos similares. Já no caso de um novo produto não ter ES, as diferenças identificadas devem ser o foco para outras avaliações. Quando não se tem base alguma para comparação de um novo alimento, ou seja, não existem produtos similares (ou no caso de plantas GM, não há uma isolinha da variedade) que tenham sido utilizados

como alimento, então o novo produto ou componente deve ser avaliado com base nas suas próprias características (OECD, 1993b).

Apesar de ter sido originalmente introduzido para alimentos GM, este conceito também é aplicado para avaliação de segurança de alimentos de novas fontes e produzidos por novos processos. O uso do conceito de ES destina-se aos testes analíticos e toxicológicos, evita o uso desnecessário de experimentos em animais e explora os dados históricos, além de incentivar uma abordagem mais abrangente da avaliação de segurança (DYBING et al., 2002).

Esta abordagem é tradicionalmente utilizada com a maioria das novas variedades de plantas utilizadas para alimentação obtidas por melhoramento clássico; portanto, estas não são sistematicamente avaliadas quanto à segurança antes da sua introdução no mercado, a menos que existam indicações claras de que a composição do vegetal é significativamente alterada. Apenas um número limitado de vegetais é rotineiramente analisado quanto aos antinutrientes e substâncias específicas. Assim, novas cultivares de soja, milho, batata e outras plantas de cultivo comuns são avaliadas pelos próprios produtores para caracterização agronômica e fenotípica (KOK et al., 2008).

O conceito de ES não constitui, no entanto, uma ferramenta específica de avaliação de segurança de alimentos. Representa o ponto inicial para estruturar a investigação de segurança de um novo alimento em relação à sua isolinha pela identificação de similaridades e diferenças. Esta é considerada uma abordagem conveniente para a avaliação de segurança de alimentos provenientes de plantas contendo DNA recombinante, porém é reconhecidamente limitada e se sugere que novas abordagens sejam feitas de maneira mais global (KLETER; KOK, 2010; KOK; KUIPER, 2003; KOK et al., 2008; KUIPER et al., 2001), e que possam inclusive avaliar a segurança para consumo de novos alimentos, que não tenham referência de histórico de uso seguro, tais como novas variedades produzidas por melhoramento clássico ou mutação química.

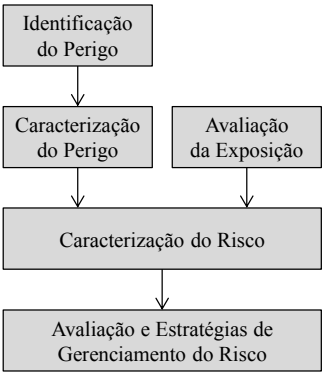
Programas nacionais de segurança de alimentos estão cada vez mais focados em uma abordagem desde o campo até a mesa como um meio eficaz de reduzir os riscos de origem alimentar. Esta abordagem holística para o controle dos riscos relacionados com os alimentos envolve a consideração de todas as etapas da cadeia, desde a matéria-prima até o consumo. Para isso, tem-se tomado medidas baseadas em sólidas informações científicas a nível nacional e internacional e, assim, sugerem-se novos métodos para identificar os potenciais perigos e reduzir a carga de doenças transmitidas por alimentos.

A avaliação de efeitos não intencionais pode ser feita pela análise das características agronômicas da nova planta ou pela avaliação ampla de nutrientes-chave, compostos antinutricionais ou componentes tóxicos típicos da planta. Alguns exemplos de efeitos não desejados que foram relatados incluem a inserção do gene de soja para expressão de glicinina que causou aumento no teor de glicoalcalóides em batatas (HASHIMOTO et al., 1999) e também originou um aumento inesperado no conteúdo de vitamina B6 em arroz (MOMMA et al., 1999); já a expressão de provitamina A em endosperma de arroz livre de carotenóide causou a formação não prevista de alguns derivados de carotenóides (YE et al., 2000).

A avaliação de risco global dos compostos alimentares ou das misturas de componentes dos alimentos deve ser realizada para se assegurar fornecimento de alimentos adequados para a saúde humana. A Análise Comparativa de Risco é um processo que inclui a identificação e caracterização do perigo, avaliação da exposição e caracterização do risco (Figura 1.2).

De acordo com Codex Alimentarius (FAO/WHO, 2011), a identificação de perigo requer a identificação de agentes biológicos, químicos e físicos capazes de causar efeitos adversos à saúde e que podem estar presentes em um grupo de alimentos. Para a caracterização do perigo deve ser feita a análise qualitativa e/ou quantitativa da natureza desses efeitos adversos possivelmente presentes no alimento, e também a avaliação do seu histórico de consumo. O cálculo de ingestão provável de alimento que pode conter substâncias que representam perigo à saúde constitui a avaliação da exposição, que é realizada em conjunto com a caracterização do perigo. Com base nesses dados é possível fazer a caracterização do risco, em que é estimada a probabilidade e severidade de ocorrência de potenciais efeitos adversos à saúde em uma determinada população. A etapa final consiste em gerenciar o risco pela ponderação de políticas alternativas levando em consideração o conjunto de fatores relevantes para a proteção da saúde dos consumidores além da adoção de medidas para prática de comércio justo e, se necessário, seleção de medidas de controle e prevenção adequadas (RENWICK, 2004; FAO/WHO, 2011; RESENDE et al., 2013).

Figura 1.2: Etapas da análise comparativa de risco, incluídas na avaliação global de segurança dos alimentos.



Fonte: adaptado de Renwick (2004).

Contudo, as técnicas comparativas apresentam relevantes limitações analíticas, uma vez que é possível a ocorrência de compostos tóxicos desconhecidos ou inesperados, principalmente na ausência de espécies similares com histórico de consumo seguro. Além disso, existe uma carência em metodologias adequadas para detecção. Em relação a estas alterações inesperadas, diferentes estratégias podem ser utilizadas para identificar possíveis efeitos secundários como a utilização de técnicas de análise-alvo (composto-específico) ou não-alvo (análise ampla de perfil) (KOK et al., 2008).

No caso das análises-alvo, para cada evento de modificação genética devem ser estabelecidas referências para a quantificação de alguns nutrientes-padrão como proteínas, carboidratos, gorduras, vitaminas e outros compostos que podem afetar a composição nutricional das novas plantas. Abordagens que utilizam análises-alvo, entretanto, possuem limitações cruciais com relação aos antinutrientes e toxinas naturais, especialmente em variedades menos conhecidas. Análises amplas de perfil como transcriptômica, proteômica e metabolômica são ferramentas complementares para a avaliação de segurança, inclusive de variedades GM. Elas permitem medições simultâneas e comparação de milhares de componentes sem a necessidade de sua identificação prévia (CELLINI et al., 2004).

A combinação destas abordagens não específicas gera novo enfoque de maior compreensão em relação às análises-alvo, criando assim, oportunidades para a identificação de efeitos não intencionais. A ocorrência de efeitos pleiotrópicos não é exclusiva de novas plantas contendo DNA recombinante, mas um fenômeno que ocorre frequentemente em novas plantas produzidas por melhoramento tradicional. Devido à prática comum de selecionar linhagens favoráveis e descartar as não desejadas ao longo das práticas de melhoramento, é raro encontrar relatos de efeitos não intencionais (CELLINI et al., 2004; COCKBURN, 2002).

1.2 ANÁLISE TRANSCRIPTÔMICA

Estudos do conjunto de níveis de transcritos, favorecidos pelo aumento do número de sequencias de DNA depositadas em banco de dados específicos, estão tornando-se cada vez mais importantes para o entendimento dos sistemas biológicos. A transcriptômica é uma análise adequada para este fim, a exemplo de tecnologias como microarranjo de DNA e RNAseq.

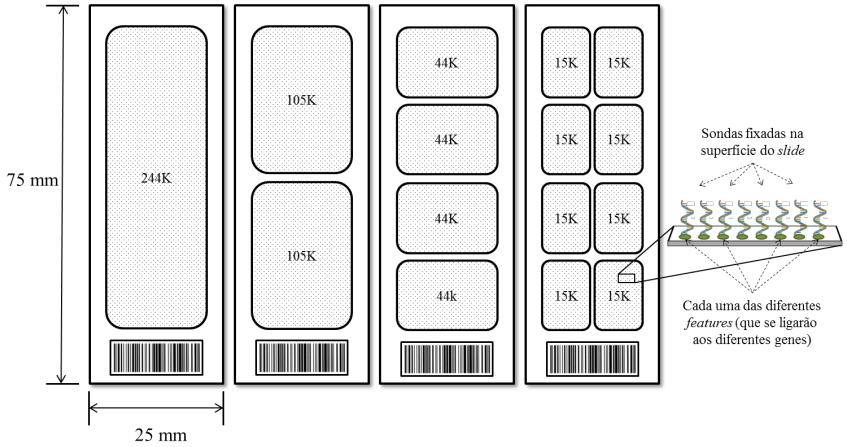
A tecnologia de microarranjo tem sido uma abordagem muito utilizada para análise de expressão gênica (DAVIES, 2010). O princípio de todos os tipos de microarranjo é o mesmo: grande quantidade de material biológico pode ser simultaneamente analisada em pequena escala. No caso de microarranjos de DNA para avaliação de transcritos, por vezes chamados *chips* de DNA, a técnica apresenta bom rendimento e se tornou ferramenta padrão para perfil de expressão gênica, uma vez que níveis de mRNA de um grande número de genes podem ser medidos simultaneamente em um único ensaio. Essa é a grande vantagem sobre os métodos tradicionais mais limitados para expressão gênica como qRT-PCR ou *Northern Blotting*, apesar de utilizarem o mesmo fundamento de que uma fita simples de DNA hibridiza em sua fita complementar sob condições adequadas de reação (BATISTA et al., 2008; SPIELBAUER; STAHL, 2005).

Os microarranjos de DNA são coleções de sondas (segmentos de fitas simples de DNA) distribuídas ordenadamente sobre uma superfície rígida feita de vidro ou silicone, chamada lâmina ou *slide*. Cada lâmina comercialmente disponível pode conter milhares de sondas distribuídas em um ou até oito arranjos, dependendo do fabricante. Cada sonda representa um único gene, porém coletivamente elas podem representar o genoma inteiro de determinado organismo. Sobre a superfície do *slide*, as sondas são posicionadas em locais específicos, os *spots*. O número de diferentes

spots por unidade de área define o volume de informação do arranjo (também chamado "complexidade"). A complexidade do *slide*, juntamente com o número de moléculas de sondas por unidade de área dentro de um *spot* (ou seja, a densidade) são parâmetros-chave de um microarranjo. Quanto mais sondas por arranjo, maior a complexidade e densidade do arranjo e, então, mais informação poderá ser obtida e mais robusta será esta ferramenta analítica. Para minimizar o tamanho dos arranjos, os locais (posições) das sondas (chamados *features*) e seu espaçamento devem ser os menores possíveis, porém ainda suficientemente precisos para o reconhecimento das moléculas (PIRRUNG, 2002).

As principais tecnologias utilizadas para fabricação automatizada de microarranjos de DNA são a fotolitografia, impressão por jato de tinta e impressão de contato. Para isso, se utilizam basicamente dois tipos de microarranjos: os *chips* de DNA e os arranjos em lâminas de vidro. Na fotolitografia, cada sonda (oligonucleotídeos de 25 bases) é sintetizada diretamente na superfície do *slide*, neste caso chamado de *chip* (fixação *in situ*). Estes *chips* de DNA são conhecidos como arranjos de DNA de alta densidade. É uma metodologia robusta e capaz de comportar até 500.000 sondas por *chip*. Possui alta especificidade e reprodutibilidade, no entanto é uma técnica bastante onerosa e com flexibilidade limitada, uma vez que os equipamentos de fixação e detecção são geralmente de uso restrito dos fabricantes. Já nas tecnologias que utilizam a impressão a jato de tinta ou de contato, as sondas são pressintetizadas para somente depois serem ligadas à superfície da lâmina de vidro (hibridização por *spotting*). Primeiramente é realizada a seleção dos genes de interesse a partir de bancos de dados específicos e a seguir os DNAs (sondas) podem ser sintetizados por PCR e purificados para, então, serem colocados no *slide* de vidro (figura 1.3). Dentre as vantagens desta tecnologia está a praticidade de fabricação e escolha das sondas. Também apresenta alta qualidade e especificidade, pois é possível customizar as sondas, tanto em relação ao seu tamanho (até 2 kpb) como aos alvos. Contudo, é uma metodologia bastante laboriosa durante as etapas de síntese e purificação, prévias à fabricação dos microarranjos. Essas técnicas de dispensação cada vez mais disponíveis tornaram os *chips* de DNA mais acessíveis aos laboratórios de pesquisa acadêmica, e ainda permitem alta precisão da análise das hibridizações (PIRRUNG, 2002; SPIELBAUER; STAHL, 2005).

Figura 1.3: Esquema de exemplo de arranjos de vidro (usados na hibridização por *spotting*). São representados quatro *slides* de microarranjo e suas dimensões. O primeiro contém apenas um arranjo com 244.000 (244K) sondas, o segundo contém dois arranjos, cada um com 105.000 (105K) sondas, o terceiro tem quatro arranjos e 44.000 (44K) sondas em cada e o quarto contém oito arranjos, cada um com 15.000 (15K) sondas. À direita da figura uma pequena área do arranjo está ampliada, mostrando as sondas e os *features* de cada *spot*.



Fonte: autor.

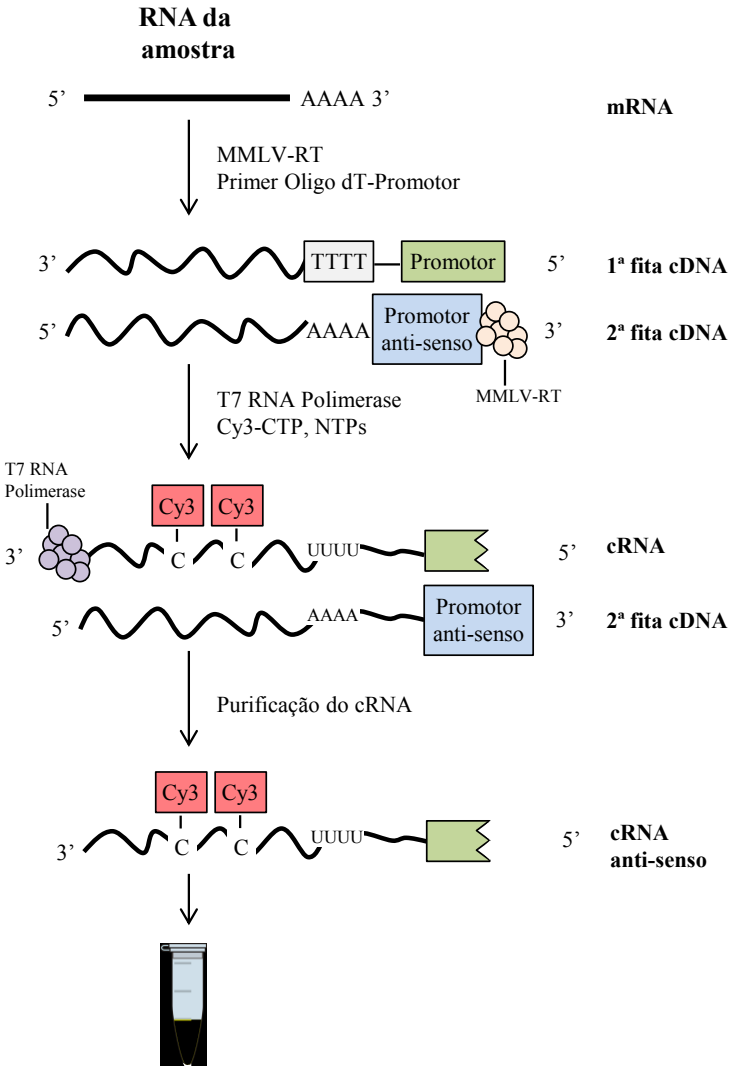
Do ponto de vista do delineamento do experimento, os diversos tipos de tecnologias de microarranjo podem ser divididos em sistemas de uma cor (*one-color* ou *single-channel*) e sistemas de duas cores (*dual-color*). Microarranjos de uma cor são aqueles em que todas as amostras são marcadas com somente um fluoróforo (geralmente cianina 3, Cy3, ou cianina 5, Cy5) e cada uma é hibridizada individualmente na lâmina. A principal vantagem é ser um método simples, porém variações entre as lâminas podem ser confundidas com efetivas diferenças entre as amostras. O microarranjo de duas cores é feito pela marcação das amostras com diferentes fluoróforos (Cy3 e Cy5) e hibridização destas juntas num único arranjo. Esse sistema minimiza as diferenças entre as lâminas e permite a comparação direta entre, por exemplo, amostra controle e amostra tratada. É, no entanto, uma abordagem mais cara e complexa, e exige que (o RNA ou cDNA da) a amostra controle mantenha-se disponível em qualidade constante ao longo de todo o experimento. A principal diferença nos resultados é que os arranjos de uma cor resultam em intensidades de

fluorescência absolutas, enquanto que os de duas cores apresentam uma proporção entre intensidades de fluorescência. Apesar destas diferenças, ambos os sistemas têm sido utilizados obtendo-se resultados similares entre si (OBERTHUER et al., 2010).

As etapas envolvidas em um experimento com microarranjos são: preparo das amostras; hibridização; e detecção, visualização e interpretação dos dados. Primeiramente o RNA das amostras é isolado e marcado com um ou dois marcadores de fluorescência. Por exemplo, fluorescência verde (Cy3) para RNAs da população tratada e fluorescência vermelha (Cy5) para as amostras controle. Em seguida, ambos os extratos são espalhados pela lâmina e as sequências de genes dos extratos hibridizam às suas sequências complementares nos *spots*. As marcações fluorescentes permitem que a quantidade de amostra ligada a um *spot* seja medida pelo nível de fluorescência emitida quando a lâmina é excitada por um laser. A detecção da fluorescência Cy3 é feita pela excitação das moléculas na faixa de comprimento de onda de 550 nm e da Cy5, de 649 nm. Por fim, verifica-se que, se o RNA em maior abundância for o da população tratada, o *spot* será verde; se o RNA do controle estiver em maior quantidade, será vermelho; havendo uma combinação de hibridização das duas populações, o *spot* será amarelo; já se nenhum dos RNAs hibridizou, o *spot* será preto. Assim, o nível de transcritos pode ser estimado pela intensidade relativa de fluorescência e pela cor de cada *spot* (BATISTA et al., 2008; SPIELBAUER; STAHL, 2005; VAN DIJK et al., 2010).

A amplificação e marcação com fluoróforo é feita com o RNA total isolado e iniciadores oligo (dT) ligados ao promotor da enzima T7 RNA polimerase, conforme esquema apresentado na figura 1.4 (exemplo de marcação com um fluoróforo, Cy3). As moléculas de mRNA presentes são convertidas a DNA dupla-fita pela ação da enzima transcriptase reversa do vírus da leucemia murina de Moloney (MMLV-RT). O material-alvo é então amplificado pela enzima T7 RNA polimerase, gerando uma fita de RNA complementar (cRNA) ligada à segunda fita de cDNA. A marcação com Cy3 é feita simultaneamente à amplificação. Os cRNAs marcados são então purificados e utilizados para a hibridização nos *slides* (placas de microarranjo) (DUGGAN et al., 1999).

Figura 1.4: Esquema ilustrativo do procedimento de amplificação e marcação fluorescente com fluoróforo Cy3 do cRNA das amostras para sua utilização na técnica de microarranjo



Fonte: modificado de DUGGAN et al. (1999).

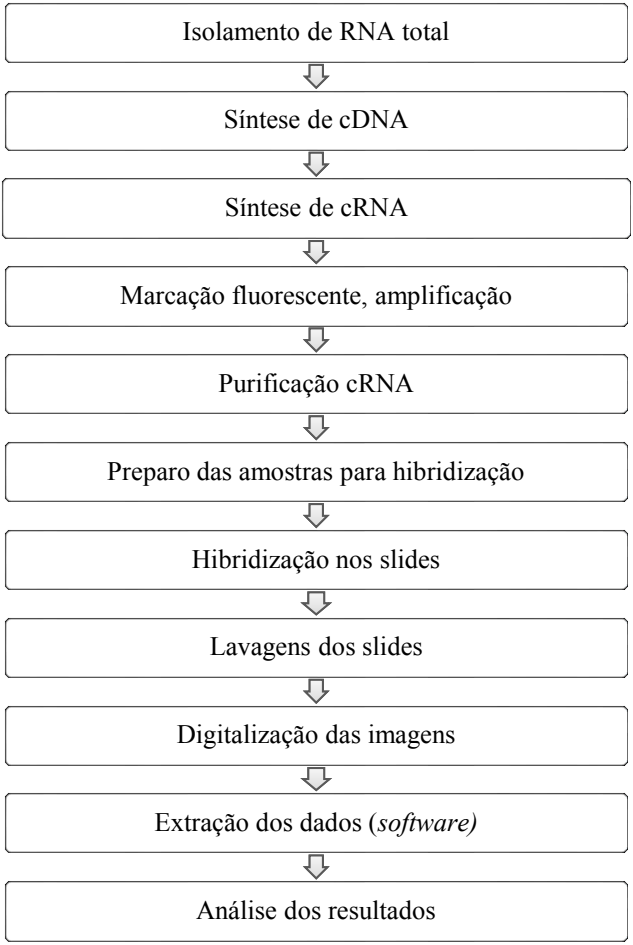
Os níveis de transcritos devem ser quantificados por meio de *scanner* que digitaliza os *slides* e mede a intensidade de fluorescência emitida em cada *spot*. Esses valores são extraídos por programas específicos que identificam o local de cada sonda, obtêm a intensidade daquela região registrada pelo *scanner* e aplicam a subtração da intensidade do fundo (*background*) local. Esses dados, corrigidos pelo programa, são então utilizados para análise (DUGGAN et al., 1999).

Um resumo das etapas da análise transcriptômica por microarranjo pode ser visualizado no fluxograma da Figura 1.5.

Sabe-se que as células respondem a mudanças no ambiente por meio de alterações na expressão gênica e isso pode influenciar na proliferação e diferenciação celular. Por isso, análises globais como as que envolvem a análise do transcriptoma, do proteoma e do metaboloma de plantas tornam-se importantes. Assim, a integração da técnica de microarranjo com a pesquisa básica e aplicada de alimentos e nutrição proporciona novas perspectivas nos efeitos nutricionais de alimentos e ingredientes como gorduras, carboidratos, proteínas, carotenoides, vitaminas, minerais e flavonoides em nível molecular. As mudanças causadas pelos macro e micronutrientes em leite, frutas e vegetais, por exemplo, ainda são pouco conhecidas. Neste sentido, a tecnologia de microarranjo auxilia na pesquisa mais aprofundada e na identificação de muitos outros componentes-alvo (DYBING et al., 2002; ROY; SEN, 2006; SPIELBAUER; STAHL, 2005).

A técnica de microarranjo tem sido amplamente utilizada para pesquisas relacionadas à segurança de alimentos. Por exemplo, foram correlacionados o genótipo e o fenótipo do patógeno *Salmonella enterica* sorotipo Enteritidis, microrganismo que causa salmonelose, doença de origem alimentar que gera grande preocupação para a saúde pública. Três cepas foram submetidas a análises de hibridização de microarranjo de DNA, ribotipagem e microarranjo de fenótipo. Os pesquisadores descreveram um levantamento sobre o tipo de características fenotípicas que estão associadas com os genótipos variantes deste microrganismo e, pelos resultados, eles propuseram alguns conceitos gerais sobre os fatores evolutivos que favoreceram o surgimento de salmonelose pandêmica associada ao consumo de ovos. Sugeriram que o problema da contaminação de ovos é derivado, principalmente, da radiação adaptativa que ajuda especificamente a sobrevivência e o crescimento do patógeno no trato reprodutivo das aves (MORALES et al., 2005).

Figura 1.5: Etapas da análise de microarranjo.



Fonte: autor.

Análise transcriptômica foi utilizada também para analisar a expressão gênica de quatro diferentes tipos de arroz, incluindo variedades mutagênicas ou transgênicas e suas respectivas isolinhas (BATISTA et al., 2008). Foram detectadas alterações nos perfis de expressão devido ao estresse causado pela modificação genética. Baudo e colaboradores utilizaram a técnica de microarranjo para comparar linhagens convencionais e transgênicas de trigo e, por meio da técnica foi possível verificar que

particularmente as amostras transgênicas analisadas eram substancialmente equivalentes às suas isolinhas (BAUDO et al., 2006). Em outro trabalho, foi verificada a eficácia da técnica de microarranjo para comparação de tomates em diferentes estádios de desenvolvimento (KOK et al., 2008), concluindo-se que o método é capaz de detectar pequenas diferenças, além de, juntamente com a proteômica, servir de importante ferramenta para a avaliação de segurança de alimentos. Também foram combinadas diferentes análises ômicas (transcriptômica, proteômica e metabolômica) para avaliação de segurança de alimentos na detecção de efeitos não intencionais derivados da modificação genética do milho GM Bt (BARROS et al., 2010). Constatou-se que os fatores ambientais, como locais e condições de cultivo e épocas de colheita, foram as principais causas de variação nos perfis transcriptômico, proteômico e metabolômico, e que as análises amplas de perfil apresentaram grande potencial para avaliação de segurança. Amostras de batatas também já foram testadas em placas de microarranjo para verificação da capacidade de detecção de diferenças nos genótipos e nas condições de cultivo; os autores confirmaram as diferenças detectadas por PCR em tempo real (VAN DIJK et al., 2009).

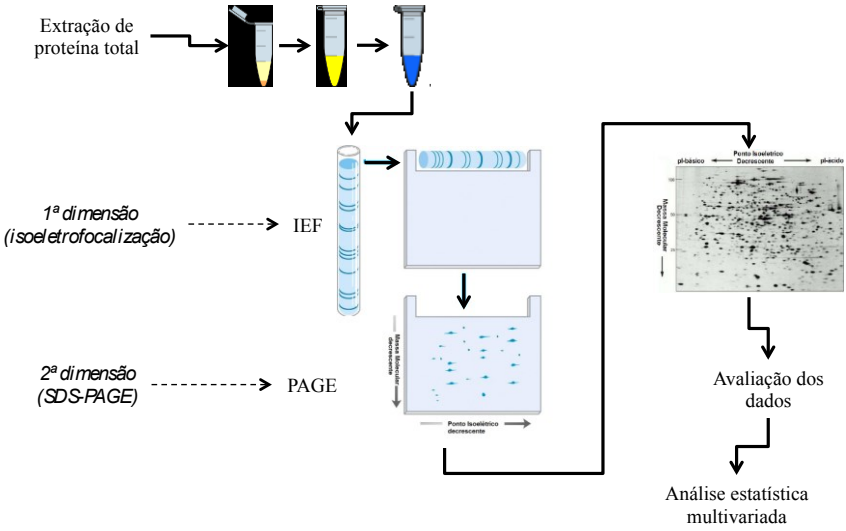
1.3 ANÁLISE PROTEÔMICA

Proteínas são as macromoléculas complexas mais abundantes nas células vivas, compostas por subunidades monoméricas de aminoácidos. As proteínas possuem uma grande diversidade de funções biológicas, e podem ser consideradas como os instrumentos moleculares por meio dos quais a informação genética é expressa (NELSON; COX, 2000). Elas apresentam interesse especial para avaliação de segurança pois elas podem formar toxinas, compostos antinutritivos ou alérgenos. A análise proteômica tem como objetivo descrever a presença de diversas proteínas e suas possíveis modificações causadas por perturbações biológicas, como mutações ou doenças, de modo abrangente e quantitativo (ANDERSON; ANDERSON, 1998; RUEBELT et al., 2006). Ao contrário do genoma, que expressa o perfil de DNA e é constante para todos os organismos, o proteoma de um organismo está em constante alteração e depende do ciclo celular, pode sofrer influência do ambiente e é tecido-específico (RUEBELT et al., 2006). Apesar de não existir um único proteoma fixo, ele é um produto que reflete diretamente o genoma. Portanto, uma pequena modificação no genoma (mutação por inserção) ou na regulação gênica de uma planta poderá causar modificação de alguma via metabólica e, por conseguinte,

poderá produzir uma nova proteína, e consequentemente o proteoma seria alterado (RUEBELT et al., 2006).

Uma das técnicas para análise proteômica mais comuns é a eletroforese bidimensional em gel de poliacrilamida (2-DE PAGE) (BINDSCHEDLER; CRAMER, 2011). A técnica envolve basicamente os passos de extração da proteína total do organismo ou tecido em estudo em um determinado momento de desenvolvimento, migração das proteínas em tiras para separação por ponto isoelétrico (pI; primeira dimensão) e migração em gel de poliacrilamida para separação por massa molecular (M_R, segunda dimensão). As proteínas são coradas e os géis contendo as duas dimensões de migração das proteínas são então digitalizados e cada proteína detectada (*spot*) é quantificada em equipamento *scanner* específico e um valor (dependendo do software utilizado, pode ser a porcentagem de volume, %Vol) é atribuído a cada *spot*. Na figura 1.6 é apresentado um esquema resumido das etapas da análise proteômica por 2-DE PAGE. Quando há interesse na identificação de proteínas diferencialmente acumuladas ou conservadas entre diferentes tratamentos, pode-se realizar a espectrometria de massas após a eletroforese bidimensional.

Figura 1.6: Esquema ilustrativo desde a extração de proteína total até a análise dos dados obtidos.



Fonte: adaptado de NELSON; COX, 2000.

A etapa de extração requer o estabelecimento de um protocolo bem definido para o sucesso da análise de proteínas. Elas devem ser desnaturadas, desagregadas, reduzidas e solubilizadas para que as interações moleculares sejam rompidas e, assim, cada *spot* represente um (ou um grupo de) polipeptídeo (s). As dificuldades na visualização das proteínas devem-se à alta faixa dinâmica de abundância e à diversidade de M_R, pI e solubilidade das proteínas (GORG; WEISS; DUNN, 2004). O método de extração escolhido influencia a isoeletrofocalização (IEF) que, por sua vez, afeta a qualidade do padrão dos geis. Isso é especialmente importante quando se trabalha com tecidos vegetais por causa da complexidade de sua matriz que contém baixos níveis de proteína, alta concentração de interferentes como compostos fenólicos, pigmentos, enzimas proteolíticas e oxidativas e de carboidratos. Não existe um único método aplicável a todos os organismos analisados por 2-DE que garanta a eficácia e reprodutibilidade das extrações. Por isso, alguns trabalhos dedicaram-se justamente à comparação de métodos de extração (DE LA FUENTE et al., 2011; DELAPLACE et al., 2006).

Os três passos fundamentais na preparação das amostras são (SHAW; RIEDERER, 2003):

(i) Rompimento celular;
(ii) Inativação ou remoção de substâncias interferentes;
(iii) Solubilização das proteínas.

O rompimento da célula pode ser feito por trituração (com ou sem nitrogênio líquido), lise osmótica, ciclos de congelamento-descongelamento, lise enzimática, sonicação, entre outros. Durante ou após a lise celular, pode haver liberação de compostos como enzimas proteolíticas e sais, que são prejudiciais à integridade da proteína extraída e por isso precisam ser removidos. As proteases podem ser inativadas pela adição de inibidores, prevenindo a degradação das proteínas. Já os sais podem interferir na separação eletroforética, por isso devem ser removidos se estiverem em alta concentração (> 100 mM). Sais podem ser retirados pela precipitação com ácido tricloroacético (TCA) e outros solventes orgânicos, ou com alternativas menos impactantes ao meio ambiente como de kits do tipo 2-D *clean-up* (GORG; WEISS; DUNN, 2004).

A solubilização das proteínas é feita em tampões contendo agentes caotrópicos (desnaturantes), detergentes não-iônicos e/ou zwitteriônicos, agentes redutores, anfólitos carreadores e inibidores de protease. As substâncias desnaturantes como ureia e tioureia são comumente utilizadas em tampões de solubilização. A ureia é bastante eficiente na ruptura das ligações de hidrogênio, causando desdobramento e desnaturação das proteínas. Já a tioureia é capaz de desfazer interações hidrofóbicas. Geralmente, os tampões contêm esses dois compostos. Porém, devido à sua baixa solubilização, a tioureia é utilizada em menor concentração que a ureia. Dentre os detergentes zwitterionicos, é comum o uso de 3-[(3-colamidopropil)dimetilamônio]-1-propanossulfonato (CHAPS) e Triton X-100. Agentes surfactantes como o dodecil sulfato de sódio (SDS) também são utilizados nos tampões de extração para prevenir interações entre domínios hidrofóbicos, evitando agregação e precipitação das proteínas. A adição de agentes redutores é necessária para a quebra de pontes dissulfeto intra- e intermoleculares, favorecendo a desnaturação (GORG, WEISS; DUNN, 2004).

Após a corrida, a visualização das proteínas é feita pela coloração dos géis, que também representa uma importante etapa do processo da 2-DE. Tradicionalmente as proteínas são coradas com Coomassie Brilliant Blue (CBB) ou nitrato de prata. Apesar de a coloração por nitrato de prata ser bastante sensível (detecção de um mínimo de 0,1 ng de proteína), ela apresenta baixa reprodutibilidade entre as replicatas. A coloração por CBB, apesar de relativamente menos sensível (detecção mínima de 10 ng de proteína), correlaciona linearmente a intensidade de cada *spot* com a

quantidade de proteína (GORG; WEIS; DUNN, 2004; GOTTLIEB et al., 2004).

A 2-DE PAGE é uma ferramenta capaz de fazer a seleção simultânea de um conjunto de proteínas permitindo, por exemplo, a identificação de proteínas importantes durante o amadurecimento de frutas. Recentemente, a técnica foi utilizada para investigação de proteínas envolvidas no amadurecimento do mamão papaia. Foram identificadas 27 proteínas diferencialmente acumuladas entre a fruta sem tratamento poscolheita e com tratamento com 1-MCP, composto capaz de bloquear a produção de etileno e por consequência retardar o amadurecimento (HUERTA-OCAMPO et al., 2012). Proteínas também foram identificadas e classificadas em seis categorias relacionadas a modificações metabólicas ocorridas durante o amadurecimento, quando comparados frutos antes e depois de maduros (NOGUEIRA et al., 2012). Na análise de segurança de alimentos, a proteômica foi utilizada para avaliar perfis proteômicos de milho geneticamente modificado MON810 e sua isolinha, concluiu-se que havia equivalência substancial entre MON810 e suas variedades não-GM (BALSAMO et al., 2011; COLL et al., 2011).

1.4 ANÁLISE MULTIVARIADA

Técnicas de análise como transcriptômica e proteômica permitem a análise simultânea de milhares de variáveis que representam os níveis de genes ou de proteínas, respectivamente, gerando um grande volume de resultados. Como salientado por Scholz e colaboradores (2004), ferramentas de análise simultânea de todos os dados são necessárias para: manuseio eficaz dos dados, incluindo coleta e preprocessamento para facilitar comparações diretas dos conjuntos de dados de análise comparativa; processamento e exploração dos dados para extrair os componentes de interesse; apresentação de dados complexos sob uma forma de fácil compreensão usando estratégias visuais; e formação de um banco de dados mais eficaz.

A imensa quantidade de dados originados com essas análises não poderia ser avaliada por ferramentas estatísticas comuns, univariadas, pois estas não fornecem graus de liberdade suficientes para suportar todo o volume de dados gerados. Estatísticas univariadas poderiam ser utilizadas para se examinar, por exemplo, o nível de variação de expressão (ativação ou supressão) de determinadas moléculas de RNA ou a presença de proteína e apenas quando se tivesse uma quantidade relativamente grande de amostras. Em situações mais comuns, em que se tem um número

limitado de amostras como, por exemplo, arranjos (na técnica de microarranjo) ou de géis (na técnica de 2-DE), e uma grande quantidade de *spots* (variáveis), é preferível que se utilize a abordagem estatística multivariada. Assim, não se trabalha com hipóteses dirigidas, mas sim faz-se a geração das hipóteses após a análise multivariada (GOTTLIEB et al., 2004). Portanto, é possível explorar os dados sem perder informações importantes e, posteriormente, se aplicável, utilizar dados bioquímicos disponíveis para estabelecer as hipóteses relevantes ao trabalho (ENGKILDE; JACOBSEN; SØNDERGAARD, 2007) .

Métodos de análise multivariada requerem um conjunto de dados suficientemente amplo para que cubra a possível variação conhecida no domínio trabalhado. Lidar com amplos conjuntos de dados requer preprocessamento para visualização e entendimento dos dados, pois um grande volume de informação distinta e relevante está contido nesses conjuntos. Para se extrair esta informação, procede-se com a mineração dos dados (*data mining*). A mineração de dados é a exploração e a análise de uma grande quantidade de dados com o objetivo de descobrir padrões e regras significativas dentre eles (BERRY; LINOFF, 2004).

As principais tarefas da mineração dos dados são (BERRY; LINOFF, 2004):

(i) Classificação: é a inspeção dos registros, alocando-os em grupos ou classes previamente definidas de acordo com determinadas características observadas;

(ii) Estimação: é um tipo de classificação, em que são estimados valores para uma ou mais variáveis desconhecidas de acordo com os dados presentes;

(iii) Predição: também é uma forma de classificação, com a diferença de que os registros a serem preditos são classificados de acordo com algum valor futuro estimado;

(iv) Regras de associação: é uma tarefa que determina quais objetos tendem a ocorrer juntos;

(v) Agrupamento: faz a segmentação de uma população heterogênea em subgrupos (ou *clusters*) mais homogêneos não definidos previamente;

(vi) Representação/descrição (*profiling*): oferece uma descrição simplificada do conjunto de dados, e pode ser empregada na etapa inicial da análise de dados.

As três primeiras tarefas são exemplos de mineração direta de dados, em que o objetivo é encontrar o valor de um alvo específico. O processo de construção de um classificador começa com um conjunto predefinido de classes e cabe à mineração direta encontrar regras que expliquem essas

classes para poder classificar/estimar/prever as variáveis-alvo. Este é o chamado *método de aprendizagem supervisionado*. Já as regras de associação e agrupamento são tarefas que representam a mineração indireta, em que o objetivo é desvendar estruturas do conjunto de dados sem fazer relação com uma variável específica. A tarefa de representação/descrição pode ser tanto direta como indireta. A mineração indireta de dados é conhecida como *método de aprendizagem não supervisionado* (BERRY; LINOFF, 2004).

Os métodos de aprendizagem podem fazer essa distinção de classes pelo reconhecimento de padrões. Assim, têm-se técnicas de reconhecimento de padrões supervisionadas, em que são usadas informações sobre as associações de classe das amostras relacionadas a um grupo predeterminado para classificar novas amostras desconhecidas, e as não-supervisionadas, em que o resultado esperado não é especificado e as classes são definidas de acordo com a similaridade dos padrões (BERRUETA; ALONSO-SALCES; HEBERGER, 2007).

A análise exploratória de dados (*Exploratory Data Analysis* - EDA) é utilizada na mineração de dados para simplificar e adquirir um melhor conhecimento sobre o conjunto de dados, além de evitar conclusões erradas ou triviais. O principal desafio é remover a redundância e o ruído ou variações que não estão relacionadas às associações de classe dos dados e, ao mesmo tempo, reter informações significativas (SIEBERT, 2001). O pretratamento dos dados deve ser feito para a avaliação de valores discrepantes, ponderação e dimensionamento dos dados. Os métodos de dimensionamento mais comuns são: transformação logarítmica (quando existem grandes diferenças de intensidades); centralização na média (média é subtraída de cada variável); *auto-scaling* (cada variável é centralizada e dividida pelo seu desvio padrão); normalização (variáveis são divididas pela raiz quadrada da soma dos quadrados das variáveis); soma de linhas constantes (cada variável é dividida pela soma de todas as variáveis de cada amostra); variável de normalização (são normalizadas em relação a uma única variável); e faixa de transformação (o valor mínimo é definido como 0 e o máximo como 1 e todos os valores intermediários encontram-se ao longo de uma faixa linear entre 0 e 1) ou, ainda, ponderação de acordo com algum critério externo (BRERETON, 2003).

Em geral, o pretratamento dos dados é necessário antes da aplicação de quaisquer técnicas multivariadas. A análise de componentes principais (PCA) é considerada a principal técnica de EDA, reduzindo a dimensionalidade e permitindo a visualização dos dados retendo tanto quanto possível de informação dos dados originais. PCA foi aplicado para discriminar os perfis composicionais de mamão papaia transgênico e sua

isolinha convencional colhidos em tempos diferentes (JIAO et al., 2010). Neste trabalho foram investigados teores de compostos voláteis orgânicos, açúcares, ácidos orgânicos, carotenoides e alcaloides através de cromatografia líquida de alta eficiência (HPLC) e espectrometria de massa. Para todos os compostos analisados, a PCA apresentou forte agrupamento entre papaias transgênicos e suas isolinhas, e nítida separação destes em relação às diferentes épocas de colheita.

A PCA é uma ferramenta que serve de base para a modelagem independente flexível por analogia de classe (SIMCA), uma técnica de reconhecimento de padrões supervisionada para modelagem de classe (BERRUETA et al., 2007). Um estudo de caso sobre adulteração de pasta de avelã foi utilizado para o desenvolvimento de um modelo SIMCA de uma classe (LÓPEZ et al., 2014). As amostras foram analisadas por espectroscopia de infravermelho e utilizadas para construção e validação do modelo, que teve seu nível de significância otimizado neste estudo.

1.4.1 Análise de Componentes Principais

A análise de componentes principais (*Principal Component Analysis* – PCA) é uma técnica de reconhecimento de padrões não supervisionada geralmente utilizada para se ter uma visão geral do conjunto de dados e para se verificar o estabelecimento de possíveis conexões entre os dados (ENGKILDE et al., 2007). A ferramenta PCA permite a transformação de um grande número de variáveis possivelmente correlatas em um número menor de variáveis não correlatas, chamadas variáveis latentes, autovetores, fatores ou componentes principais (PC). As novas variáveis são sempre ortogonais (ou seja, perpendiculares, não correlacionadas) entre si, e sucessivos PCs descrevem quantidades decrescentes de variação do conjunto de dados. O primeiro componente principal cobre tanto o quanto é possível da variação do conjunto de dados, e cada componente subsequente cobre tanto quanto possível do restante da variação. Deste modo, a maioria da variação está contida logo no primeiro componente e progressivamente menos informação relevante estará contida quanto maior for o número de componentes (GOTTLIEB et al., 2004).

Essas variáveis não podem ser medidas diretamente, mas devem ser expressas como uma combinação linear de um conjunto de variáveis. A relação entre os PCs e as amostras é chamada *score* e a relação com as variáveis é chamada *loading*. A matriz original de dados X é decomposta em uma parte estrutural e uma parte de erro. A parte estrutural corresponde à matriz de *scores*, T, e à matriz transposta de *loadings*, P^T, enquanto a

matriz de erro é denominada *E* (ESBENSEN; SCHOENKOPF; MIDTGAARD, 1994), de acordo com a equação (1):

$$X = T \times P^T + E \quad (1)$$

A matriz de *scores T* indica as coordenadas dos dados originais no novo sistema de coordenadas dado pelos PCs e a matriz de *loadings P* contém os coeficientes destas variáveis no espaço original. Como resultado, essa combinação gera um diagrama de ordenação (*score plot*, ou diagrama de dispersão) onde os eixos originais são substituídos por eixos de componentes principais, e dependem dos valores atribuídos para cada amostra (*scores*) e para cada variável (*loadings*) (ENGKILDE et al., 2007). Os diagramas de dispersão, portanto, são gráficos que indicam a posição das amostras ao longo dos componentes principais. Amostras com maior similaridade em um determinado PC estarão mais próximas entre si no gráfico (ENGKILDE et al., 2007; GOTTLIEB et al., 2004).

1.4.2 Modelagem Independente Flexível por Analogia de Classe – SIMCA

A modelagem independente flexível por analogia de classe (SIMCA, do inglês *Soft Independent Modelling of Class Analogy*) é um método de reconhecimento de padrões supervisionado que modela a localização das classes no espaço multidimensional pelo cálculo de PCs separadamente para cada categoria. Nesse sistema, uma amostra desconhecida pode ser classificada como pertencente a uma, a mais de uma ou a nenhuma das categorias predefinidas (BRERETON, 2003). Uma classe é modelada a partir do número de PCs selecionados por validação cruzada.

O sistema SIMCA apresenta duas grandes vantagens em comparação a outros métodos de reconhecimento de padrões (HERRERO LATORRE et al., 2013). Cada classe de dados é modelada separadamente do resto das classes, o que evita a influência de uma sobre as outras no processo de definição das classes. Além disso, SIMCA é capaz de detectar *outliers*, o que significa que a técnica é capaz de identificar amostras que não pertençam a quaisquer das classes determinadas.

Assim como os outros procedimentos supervisionados de reconhecimento de padrões, a SIMCA consiste dos seguintes passos (BERRUETA et al., 2007; BRERETON, 2003):

(i) Seleção de um conjunto de treinamento, de validação e de teste, que consistem de objetos de associação de classes conhecidas para as quais as variáveis são medidas;

(ii) Seleção das variáveis, para eliminar variáveis que correspondam a ruído ou que não tenham poder discriminante;

(iii) Construção de um modelo utilizando o conjunto de treinamento, gerando categorias que agrupam as variáveis medidas;

(iv) Validação do modelo utilizando um conjunto independente de amostras, para avaliar a consistência da classificação.

O número de PCs em cada classe não deve ser muito alto, pois há risco de se modelar baseado em ruído ou em dados aleatórios, e nem deve ser muito baixo, pois se pode deixar de incluir variação suficiente, diminuindo o poder explanatório do conjunto de dados (BERRUETA et al., 2007). Também é possível a construção de modelo SIMCA contendo uma só classe, assim as amostras podem ser classificadas como pertencentes à classe predefinida ou como não pertencente a qualquer classe (BRERETON, 2003; XU; BRERETON, 2005).

O método é validado por estabelecimento de um conjunto de amostras de validação que seja independente do conjunto de amostras de treinamento que foi utilizado para construir o modelo. O conjunto de amostras para validação deve ser previamente conhecido para que se verifique a fidelidade de sua classificação. Se não houver amostras independentes suficientes para validação externa do modelo, uma validação cruzada interna pode ser realizada. Para este tipo de validação interna, o próprio conjunto de dados é dividido em dois subconjuntos: treinamento e validação (BERRUETA et al., 2007; FLATEN; GRUNG; KVALHEIM, 2004).

A validação cruzada interna é realizada pela segmentação do conjunto de dados. Um dos segmentos é deixado de fora e um modelo é construído com o restante dos segmentos, o conjunto de treinamento. Este modelo é então validado com o segmento externo, o conjunto de validação. O procedimento é repetido tantas vezes quantas forem os segmentos deixados de fora e, para cada modelo, este conjunto externo é utilizado para validação. Ao longo do processo teremos, portanto, diversos conjuntos de treinamento e de validação. Assim, ao final da validação cruzada, haverá não apenas um modelo, mas um grupo de modelos que descrevem o conjunto de dados (ENGKILDE et al., 2007).

CAPÍTULO 2

AVALIAÇÃO DE SEGURANÇA DE VARIEDADES DE PLANTAS UTILIZANDO ANÁLISE TRANSCRIPTÔMICA E CLASSIFICADOR DE UMA CLASSE

Artigo publicado (ver Apêndice C):
VAN DIJK, J. P., MELLO, C. S., VOORHUIJZEN, M. M., HUTTEN, R. C. B., ARISI, A. C. M., JANSEN, J. J., BUYDENS, L. M. C., VAN DER VOET, H., KOK, E. J. Safety assessment of plant varieties using transcriptomics profiling and a one-class classifier. **Regulatory Toxicology and Pharmacology,** v. 70, n. 1, p. 297-303, 2014.
DOI: http://dx.doi.org/10.1016/j.yrtph.2014.07.013.

2 AVALIAÇÃO DE SEGURANÇA DE VARIEDADES DE PLANTAS UTILIZANDO ANÁLISE TRANSCRIPTÔMICA E CLASSIFICADOR DE UMA CLASSE

SAFETY ASSESSMENT OF PLANT VARIETIES USING TRANSCRIPTOMICS PROFILING AND A ONE-CLASS CLASSIFIER

Jeroen P. van Dijk[1*†], Carla Souza de Mello[1,2†], Marleen M. Voorhuijzen[1], Ronald C. B. Hutten[3], Ana Carolina Maisonnave Arisi[2], Jeroen J. Jansen[4], Lutgarde M.C. Buydens[4], Hilko van der Voet[5], Esther J. Kok[2]

[†]The first two authors contributed equally to this manuscript.

[1] RIKILT, Wageningen UR, Wageningen, the Netherlands
[2] Federal University of Santa Catarina, Brazil
[3] Plant Breeding, Wageningen UR, Wageningen, the Netherlands
[4] Analytical Chemistry, Radboud University Nijmegen, the Netherlands
[5] Biometris, Wageningen UR, Wageningen, the Netherlands

RESUMO

Uma parte importante do atual procedimento de identificação de perigo de novas variedades de plantas é a análise-alvo comparativa, tanto para variedades novas como para variedades referência. A análise comparativa se tornará muito mais informativa com abordagens analíticas imparciais, como as análises ômicas. A análise de dados através da estimativa de similaridade de novas variedades com uma classe de referência contendo variedades conhecidas como seguras facilitaria muito a identificação de perigos. Análises biológicas e toxicológicas posteriores somente seriam necessárias para variedades que caíssem fora do padrão de valor estabelecido pelo sistema de classificação. Para esse propósito, um sistema de classificação de uma classe foi explorado para avaliar e classificar perfis de transcriptoma de variedades de batata (*Solanum tuberosum*) em um estudo modelo. Perfis de seis variedades, dois locais de cultivo, dois anos diferentes de colheita, incluindo replicatas técnicas e biológicas, foram utilizados para construir o sistema de classificadores. Foram feitas duas análises, uma representando a avaliação de uma variedade diferente (análise A) e outra representando uma variedade similar (análise B). Os resultados mostraram que a análise A apresentou maior distâncias de classe para o conjunto de teste comparado com a análise B. As observações reportadas neste estudo podem contribuir para uma abordagem mais global na identificação de perigos de novas variedades de plantas.

Palavras-chave: Transcriptômica. Classificação de uma classe. Avaliação de segurança de alimentos. Quimiometria.

ABSTRACT

An important part of the current hazard identification of novel plant varieties is comparative targeted analysis of the novel and reference varieties. Comparative analysis will become much more informative with unbiased analytical approaches, e.g. omics profiling. Data analysis estimating the similarity of new varieties to a reference baseline class of known safe varieties would subsequently greatly facilitate hazard identification. Further biological and eventually toxicological analysis would then only be necessary for varieties that fall outside this reference class. For this purpose, a one-class classifier tool was explored to assess and classify transcriptome profiles of potato (*Solanum Tuberosum*) varieties in a model study. Profiles of six different varieties, two locations of growth, two year of harvest and including biological and technical replication were used to build the model. Two scenarios were applied representing evaluation of a "different" variety (Analysis A) and a "similar" variety (Analysis B). Within the model, higher class distances resulted for the "different" test set compared with the "similar" test set. The present study may contribute to a more global hazard identification of novel plant varieties.

Key words: Transcriptomics. Profiling. One-class classifiers. Food safety evaluation. Chemometrics.

2.1 INTRODUÇÃO

O desenvolvimento de novas variedades de plantas tem conduzido questionamentos a respeito da comprovação de sua segurança. Essa discussão tem focado principalmente na avaliação de variedades de planta GM, mas geralmente é também aplicável a outras espécies de novas plantas. A abordagem básica, internacionalmente aceita, é a análise comparativa, onde novas variedades são comparadas àquelas com histórico de consumo seguro para humanos (FAO/WHO, 1996; KOK; KUIPER, 2003; KOK; KEIJER; et al., 2008; OECD, 1993a; OECD, 2002). A avaliação deve compreender os conceitos de identificação do perigo, caracterização do perigo e avaliação da exposição que levam à caracterização do risco, incluindo tanto efeitos intencionais como os potenciais efeitos não-intencionais que sejam provenientes de modificação genética ou não (KNUDSEN et al., 2008; RENWICK, 2004). Na Europa, o conceito de análise comparativa de segurança para novas variedades de plantas GM foi detalhado em um documento de orientação elaborado pela Autoridade Europeia de Segurança de Alimentos (*European Food Safety Authority*, EFSA) (EFSA, 2011). Grande parte deste documento foi incluída na legislação europeia (EUROPEAN-COMISSION, 2013).

Uma parte importante da identificação do perigo é a comparação da análise composicional de plantas GM com a de variedades convencionais, como foi formulado pela Organização das Nações Unidas para Alimentação e Agricultura (*Food and Agricultural Organisation*, FAO), Organização para Cooperação e Desenvolvimento Econômico (*Organisation for Economic Co-operation and Development,* OECD) e pela EFSA (EFSA, 2011; FAO/WHO, 1996; OECD, 1993a). A análise composicional deve incluir todos os compostos-chave (nutrientes, antinutrientes e toxinas) específicos para a variedade investigada (EL SANHOTY; ABD EL-RAHMAN; BÖGL, 2004). Tais compostos foram descritos pela OECD em documentos consenso para diferentes variedades (OECD, 2002). Por um lado, testes estatísticos são realizados para identificar diferenças para certos compostos com um comparador direto, por exemplo, o genótipo parental. Por outro lado, uma comparação mais abrangente é feita com a variação natural destes compostos sob diferentes condições ambientais e efetivamente em diferentes variedades consideradas seguras como, por exemplo, numa abordagem com teste de equivalência (VAN DER VOET et al., 2011).

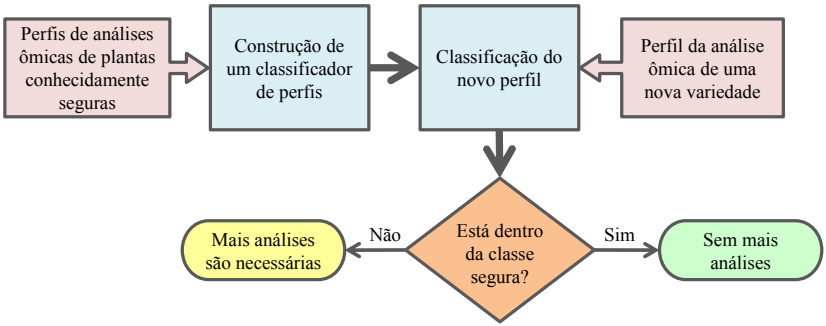
Em 2013, o Comitê Permanente da União Europeia da Cadeia Alimentar e da Saúde Animal (*EU Standing Committee on the Food Chain and Animal Health*) atualizou a regulamentação sobre aplicações de

organismos GM para alimentação humana e animal na União Europeia (EUROPEAN-COMISSION, 2013). Um novo item nessa regulamentação é a exigência de estudos de alimentação em roedores com o alimento inteiro durante 90 dias para cada evento de transformação, e em casos específicos, o mesmo ensaio para plantas que contêm eventos de transformação com genes combinados por cruzamento convencional. No entanto, o guia da EFSA de 2011 recomenda esse tipo de experimento somente em determinadas condições, como por exemplo, quando houver baixa qualidade dos dados disponíveis para a avaliação de risco (EFSA, 2011). Comentários em revistas científicas também têm questionado vários aspectos das regulações atuais com relação a variedades GM (DEFRANCESCO, 2013; HERMAN; PRICE, 2013; KUIPER, HARRY A.; KOK; DAVIES, 2013). Neste trabalho levanta-se a hipótese de que, mudando de análises-alvo para análises amplas de perfil (não-alvo), maior valor será agregado para a identificação de perigos do que realizar ensaios de alimentação em animais. Um estudo de prova de princípio é apresentado neste artigo.

Nos últimos anos, diferentes "abordagens ômicas" desenvolveram-se de maneira que podem ser adequadas para análise ampla de perfil (análise não-alvo) na análise composicional comparativa. Dentre as diversas abordagens ômicas, a transcriptômica apresenta maior cobertura do sistema biológico quando comparada, por exemplo, à metabolômica. Assim, este acaba sendo o método escolhido quando a comparação deve ser feita de maneira abrangente. Diversos estudos mostraram perfis transcriptômicos diferenciais reprodutíveis de produtos derivados de plantas em diferentes situações relacionados com GM (BARROS et al., 2010; BAUDO et al., 2006; CHENG et al., 2008; COLL et al., 2010), e também outros fatores como insumos agrícolas, ano de colheita e local de cultivo (VAN DIJK et al., 2009; VAN DIJK et al., 2012; ZORB et al., 2009). A interpretação da relevância toxicológica das diferenças encontradas tem sido dificultada pela carência de conhecimento em relação ao impacto toxicológico de diversos genes e rotas metabólicas subjacentes. Outra limitação refere-se ao tipo de análise de dados. Geralmente, métodos multivariados são usados para exploração dos dados seguidos por análise univariada de um único gene ou rota. Nesta fração univariada existe a possibilidade de falsos-positivos devido a testes múltiplos (BENJAMINI; HOCHBERG, 1995; STOREY; TIBSHIRANI, 2003). Outra dificuldade é a quantidade muito maior de variáveis do que de amostras (KOSOROK; MA, 2007). Consequentemente, um resultado típico de comparação por análise transcriptômica é uma lista com níveis de transcritos com determinados valores p, que representam uma estimativa da taxa de falsos-positivos, sendo que muitos destes dados

têm funções desconhecidas. Este resultado fornece um ponto de partida para, por exemplo, o desenvolvimento de medicamentos ou diagnóstico de doenças, quando o objetivo é encontrar os genes responsáveis numa comparação de situações sabidamente distintas. Porém, é menos adequado para avaliação de segurança de alimentos, onde a existência de uma diferença, na verdade, deve ser estabelecida em primeiro lugar. Para esse propósito específico, uma alternativa melhor para usar os dados de análises ômicas seria incorporar conhecimento biológico das plantas que têm sido cultivadas por muitas gerações como uma base comparativa para obtenção de cultivares seguras. O primeiro passo seria estimar se uma nova planta é classificada como "segura" baseada no perfil de expressão gênica como um todo por meio da classificação multivariada. Essa estimativa poderia ser calibrada contra amostras conhecidamente inseguras ou com características indesejadas. Indivíduos cujos perfis sejam classificados fora desta classe segura deverão ser submetidos a mais análises, similares às praticadas atualmente, baseadas em alvos únicos. A fase seguinte da caracterização de risco deveria determinar se perfis atípicos (mais afastados da classe segura) têm relevância (Figura 2.1). Este seria o primeiro passo para a identificação de variáveis que fazem com que a amostra seja classificada fora da classe segura.

Figura 2.1: Resumo esquemático da abordagem proposta para a avaliação de segurança de alimentos. O princípio é uma expansão da atual análise-alvo composicional comparativa na identificação de perigos em segurança de alimentos para uma comparação não-alvo baseada em perfis de análises ômicas. É proposta, ainda, a aplicação de um sistema de classificação multivariada de uma classe para identificação de perigos, que depende se os perfis de novas variedades de planta cairão dentro ou fora de uma classe de perfis conhecidamente seguros. A caracterização dos perigos somente seria realizada para classificações fora da classe segura.



A classificação multivariada considera simultaneamente perfis de muitas variáveis como os valores de expressão gênica do genoma inteiro de uma planta assim como as potenciais relações biológicas entre elas. Essa abordagem conduz à tradução do perfil a uma associação de classe por intermédio de um chamado "classificador". Esse classificador pode ser usado para designar uma nova amostra em uma ou mais classes predefinidas. Para a identificação de perigos de novas variedades de plantas, a abordagem mais adequada à sua finalidade é a classificação de uma classe. Ela já foi aplicada em diversas situações em que fora da classe de base existe pouca disponibilidade de amostras ou uma variedade muito ampla (TAX, 2001). Na avaliação de segurança de alimentos, as variedades não seguras são escassas e, ao mesmo tempo, diversificadas, o que leva ao uso mais apropriado de um classificador uma classe para determinar um patamar de segurança para esses alimentos. Esse patamar e, por conseguinte, essa classificação, deve necessariamente incluir diversos parâmetros como diferentes cultivares, anos de coleta, tipos de solo e localização geográfica (BERRUETA et al., 2007).

Neste trabalho o objetivo foi explorar a classificação de uma classe para perfis gerados por análise transcriptômica usando o método de modelagem independente flexível por analogia de classe (SIMCA) (DE MAESSCHALCK et al., 1999; WOLD; SJÖSTRÖM, 1977). Perfis de

amostras de batatas com fontes de variação diferentes e bem definidas foram usados. A aplicabilidade dessa classificação de uma classe para o aperfeiçoamento da atual avaliação de segurança de novas variedades de plantas é avaliada e discutida.

2.2 MATERIAL E MÉTODOS

2.2.1 Amostras

Cinco variedades de batatas (Biogold, Fontane, Innovator, Lady Rosetta e Maris Piper) foram cultivadas e colhidas em Wageningen, Países Baixos (amostras NL), em 2010. Cada variedade foi cultivada em dois lotes contendo diferentes substratos (solo argiloso ou arenoso), e cada lote continha dois grupos de cada cultivar, totalizando quatro plantas individuais. A exceção foi a variedade Maris Piper, que foi cultivada em somente um lote de cada substrato, com três plantas individuais em cada lote. Os tubérculos foram armazenados por sete dias no escuro à temperatura ambiente antes das análises. Uma variedade adicional (Sante) foi utilizada; que foi cultivada em 2005 no Reino Unido (amostras UK) como parte do estudo QLIF (*Quality Low Input Food*) como descrito por van Dijk e colaboradores (2012). Amostras identificadas pelos números 1058 e 1059 foram selecionadas. Uma replicata técnica para essas duas "amostras UK" foi realizada. Foram incluídos perfis analisados no estudo original em 2008 e novamente em 2011 junto com as "amostras NL" do presente estudo, resultando em quatro perfis para essas duas amostras. Vale salientar que um dos perfis originais era similar à maioria das outras amostras naquele estudo, enquanto que a outra amostra era mais distante, baseada na PCA do estudo QLIF original (VAN DIJK et al., 2012). Esta amostra mais distante foi chamada *outlier* nesse artigo. As replicatas técnicas das duas amostras UK foram realizadas a partir de um segundo isolamento de RNA das mesmas batatas liofilizadas armazenadas a -80°C. Um resumo das amostras utilizadas neste estudo é apresentado na tabela 2.1.

Para explorar as diferenças entre as variedades, incluindo a variação biológica entre os lotes nas replicatas, os tubérculos foram agrupados e processados de acordo com Lehesranta (LEHESRANTA et al., 2007). Foram selecionadas quatro batatas de tamanho médio de cada lote (totalizando aproximadamente 800 g) e cortadas em cubos. Dois oitavos opostos de cada batata foram selecionados para minimizar os efeitos de variação das batatas. O material foi liofilizado por 18 h, moído até formar

um pó fino com o auxílio de graal e pistilo e armazenado a -80°C até o isolamento de RNA.

Tabela 2.1: Descrição resumida das variedades de batatas utilizadas neste estudo

	Variedade	Característica	Número de amostras	Identificação
Amostras NL	Biogold	Solo argiloso	2	BK
		Solo arenoso	2	BZ
	Fontane	Solo argiloso	2	FK
		Solo arenoso	2	FZ
	Innovator	Solo argiloso	2	IK
		Solo arenoso	2	IZ
	Lady Rosetta	Solo argiloso	2	LK
		Solo arenoso	2	LZ
	Maris Piper	Solo argiloso	1	MK
		Solo arenoso	1	MZ
Amostras UK	Sante	Cultivada em 2005 e analisada em 2008 (*outlier*)	1	S08-1158
		Cultivada em 2005 e analisada em 2008	1	S08-1159
		Cultivada em 2005 e analisada em 2011 (*outlier*)	1	S11-1158
		Cultivada em 2008 e analisada em 2011	1	S11-1159
Total de amostras			22	

NL, amostras cultivadas nos Países Baixos; UK, amostras cultivadas no Reino Unido. Todas as amostras NL foram cultivadas em 2010 e analisadas em 2011.
Fonte: autor.

2.2.2 Isolamento e qualidade dos RNA

O RNA das batatas foi isolado a partir de 0,5 g de cada amostra liofilizada e triturada segundo metodologia descrita anteriormente (VAN DIJK et al., 2009), baseada no método CTAB com consecutiva extração clorofórmio/álcool isoamílico seguida por precipitação com LiCl. As seguintes modificações no método foram feitas: o tampão de extração foi preaquecido a 60°C antes de ser utilizado, a extração clorofórmio/álcool isoamílico foi repetida três vezes antes da precipitação com LiCl e a precipitação final com etanol 96% foi feita com os tubos mantidos em gelo e centrifugação a 4°C por 15 min a 14.000 g. O RNA foi dissolvido em 100 µL de 10 mM Tris (pH 7) e aquecido a 65°C por 10 min.

A concentração e pureza dos RNAs isolados foram avaliadas pela medição dos picos de absorbância a 230, 260 e 280 nm em Nanodrop 1000 Instrument (Thermo Fisher Scientific, NanoDrop Technologies Wilmington, DE, EUA). Para avaliação da integridade, 1 µg de RNA foi migrado por eletroforese (60 min, 80 V) em gel de agarose desnaturante (1% agarose, 5% formamida, TBE 1X) corado com brometo de etídeo. Os géis foram visualizados em fotodocumentador Gel Doc XR+ Systems (Bio-Rad Laboratories, Life Technologies Corporation, Carlsbad, CA, EUA) e analisados com o auxílio do software Quantity One 1-D (Bio-Rad Laboratories). Amostras com a relação quantidade de rRNA 18S e rRNA total acima de 40% foram consideradas adequadas para as análises de microarranjo. A avaliação da pureza e qualidade do RNA foi confirmada para todas as amostras.

2.2.3 Marcação fluorescente e hibridização

Foram utilizados 2 µg de RNA para a marcação fluorescente por meio da incorporação de Cy3-dCTP durante a reação de síntese de cDNA utilizando-se Quick Amp Labeling Kit (Agilent Technologies, Inc., Santa Clara, CA, EUA). A partir dos cDNAs marcados, foi feita a síntese de cRNA e, após a purificação, foi verificada sua qualidade e rendimento. Todos os cRNAs marcados apresentaram rendimento acima de 1,65 µg e atividade específica do fluoróforo acima 9,0 pmol Cy3/µg RNA, sendo então considerados próprios para hibridização nos *slides* de microarranjo.

Os cRNAs foram distribuídos aleatoriamente nos *slides* de microarranjo POCI 4x44 (KLOOSTERMAN et al., 2008) e hibridizados

durante 17 h em câmara de hibridização com temperatura controlada e mantida a 65°C. Os *slides* foram então lavados com tampões de lavagem e de fixação, solução de acetonitrila e por fim tampão de estabilização/secagem (Agilent; conforme orientações do fabricante) em temperatura ambiente e imediatamente digitalizados.

2.2.4 Digitalização, análise das imagens e análise dos dados do microarranjo

Os slides foram digitalizados após excitação do fluoróforo Cy3 com laser a 543 nm em Scanner C (Agilent) usando as configurações-padrão do equipamento. Os arquivos de imagens obtidos foram inspecionados visualmente quanto à presença de arranhões ou manchas que pudessem interferir posteriormente na interpretação dos dados. Os dados das imagens foram extraídos pelo Feature Extraction Software v 8.5; as intensidades de fluorescência e de ruído (*background*) foram determinadas para cada *spot* e exportados para o software Excel (Microsoft Office 2007).

Os dados utilizados para análise foram coletados a partir de colunas contendo "*feature number*", número de acesso, sinal do *spot* e sinal de ruído; os *spots* controle (aqueles sem número de acesso) foram removidos. Das 44.000 variáveis (genes) de cada perfil, somente foram incluídas as que apresentaram intensidade de fluorescência maior que duas vezes o sinal de ruído correspondente (MASSART et al., 1997); também os dados originais do estudo QLIF (VAN DIJK, JEROEN P. et al., 2012) foram considerados para a seleção dos *spots*, resultando em 20.370 variáveis das 44.000 leituras originais dos *chips*. Os sinais selecionados foram tratados por transformação logarítmica de base 2, normalizados por arranjo (subtraindo de cada sinal de *spot* individual o valor da mediana dos 20.370 *spots* daquele arranjo) e por *spot* (subtraindo de cada *spot* individual o valor da mediana de todos os 22 arranjos utilizados). Os dados brutos e processados foram depositados sob o número E-MTAB-1707 no banco de dados *Array Express* (http://www.ebi.ac.uk/arrayexpress) do Instituto Europeu de Bioinformática (*European Bioinformatics Institute*, EBI) (BRAZMA et al., 2000; RUSTICI et al., 2013). PCA e SIMCA foram realizados utilizando Pirouette Software v. 4 (Infometrix, Inc., Bothell, WA, USA).

2.3 TEORIA

2.3.1 Classificação multivariada

A construção de um classificador supervisionado multivariado robusto requer três conjuntos independentes de amostras reconhecidamente seguras (ou inseguras). (MASSART et al., 1997; VANDEGINSTE et al., 1998). O *conjunto de treinamento* é usado para construir uma primeira versão do classificador, ajustando os parâmetros do modelo para otimização de sua classificação. Um *conjunto de validação* separado é necessário para adaptar outros parâmetros do classificador, tais como o número de variáveis latentes de um componente ou as variáveis selecionadas para classificação. Este passo de validação pode ser feito por validação cruzada interna, normalmente realizada quando somente um conjunto limitado de amostras está disponível (WOLD, 1978). Na validação cruzada, um grupo de amostras é removido por vez e um classificador é construído usando as amostras restantes, sendo que as amostras deixadas de fora são usadas como conjunto de validação. Esse procedimento é repetido até que todos os grupos de amostras tenham sido removidos uma vez e validados nos diferentes classificadores. No entanto, eles não são mais independentes do modelo final porque o conjunto de validação é usado para escolha de parâmetros para os classificadores. Sendo assim, um *conjunto de teste* é usado para determinar a precisão da classificação do modelo final. Essa abordagem com os três conjuntos é essencial para evitar *overfitting* do classificador descrevendo não especificamente as amostras dentro do conjunto de dados, mas sim a população da qual foram retiradas. Classificadores SIMCA determinam uma distância de classe para cada amostra, baseada no perfil da amostra. Consequentemente, um intervalo de confiança (IC) de 95% pode ser determinado como sendo um limite para a associação de classe das amostras do conjunto de teste. Uma amostra teste é classificada como não pertencente à mesma classe quando uma distância maior que o equivalente aos 95% IC é observado. Para o presente estudo, um classificador de uma classe é utilizado, portanto a distância de classe em todos os casos está relacionada a essa uma classe de perfis considerados seguros. Neste trabalho a ferramenta de classificação SIMCA foi escolhida porque utiliza PCA para reduzir a complexidade na variabilidade dos dados, o que se espera que beneficie a construção de um novo classificador com um pequeno, porém representativo, conjunto de amostras (TAX, 2001). Também foi escolhido utilizar todos os classificadores distintos resultantes da validação cruzada em vez de fundi-los em um classificador geral,

análogo ao trabalho de Westerhuis (WESTERHUIS et al., 2008). Estes autores argumentaram contra o uso de um único classificador final; como alternativa, propuseram o uso de vários classificadores diferentes para a obtenção de uma série de predições de associações de classe pois, até o momento, não há critérios aceitos para o modo de escolha de um modelo geral e único.

2.3.2 Configuração do estudo

Dois cenários foram explorados para avaliar a influência das diferentes fontes de variação no transcriptoma e consequente influência das diferentes fontes de variação nos resultados de classificação: cenário "diferente" e "similar". Variedade de batata, localização de cultivo, ano de colheita, replicata biológica e replicata técnica (ano de análise) foram incluídas como fontes de variação de um total de 22 perfis a partir de 20 amostras de tubérculos de batata (Figura 2.2). Todos os perfis foram usados em ambos cenários; a diferença foi sua distribuição entre os conjuntos de teste e de treinamento. Por conveniência de nomenclatura neste capítulo, o cenário "diferente" a partir deste ponto será chamado "análise A" e o cenário "similar", "análise B". Ambas as análises foram usadas para simular novas variedades que fossem diferentes ou similares às variedades de base (treinamento) em avaliação comparativa de segurança.

A princípio, esperava-se que os quatro perfis Sante (dois S08 e dois S11) fossem os mais diferentes devido às diferenças de variedade, ano e local de cultivo, quando comparados às outras amostras. Ainda dentro deste grupo, dois perfis provinham de análise realizada dois anos antes (S08), prevendo-se serem ainda mais diferentes. Além disso, um dos perfis era um *outlier* no estudo original (S08-1158), esperando-se, portanto, que apresentasse a maior diferença de todos os outros perfis do presente estudo. Como todas as variedades de batatas utilizadas são seguras para consumo humano, esperava-se que todas as amostras fossem classificadas como pertencentes à classe segura, exceto o perfil da amostra *outlier*. Na análise A, os perfis das quatro amostras UK (variedade Sante) foram usados como conjunto de teste. Neste caso, todas as amostras do conjunto de treinamento (amostras B, F, I, L e M) compartilharam o mesmo ano e local de cultivo e momento de análise, sendo todas estas características diferentes do conjunto de teste. Por isso, certa quantidade de erros de classificação é esperada, uma vez que o conjunto de treinamento não representa completamente a variação das amostras teste.

Na análise B, os perfis S08-1158 e S08-1159 foram agrupados com os perfis de treinamento. Em troca, a variedade M, que contém dois perfis, foi agrupada no conjunto de teste, mantendo o mesmo número de perfis nos conjuntos de treinamento e de teste nas duas análises. Os dois perfis S08 foram mantidos no conjunto de teste. Deste modo, os dois locais e anos de cultivo ficaram presentes em ambos os conjuntos de teste e de treinamento. As diferenças entre estes conjuntos foram: a variedade M estar presente no conjunto de teste na análise B, e a presença de replicata técnica para amostra Sante. Além disso, o *outlier* S08-1158 ainda seria a amostra mais diferente. Assim, a classificação esperada deste conjunto de teste seria similar ao patamar de base em todos os casos exceto pelo perfil *outlier*. Nesta análise, menos erros de classificação eram esperados.

Figura 2.2: Distribuição dos perfis entre as análises A e B. Cada linha representa um classificador e contém o mesmo número de perfis. Os perfis foram distribuídos em duas análises, representando um conjunto de teste diferente (A) ou conjunto de teste similar (B). Dentro de cada análise, cinco classificadores foram construídos a partir de um patamar estabelecido com quatro variedades usadas como calibradores (conjunto de treinamento) e a variedade remanescente usada para validação (conjunto de validação, sombreada em cinza). Números entre parênteses indicam o número de perfis de cada variedade. Fonte regular indica local e ano de colheita: Países Baixos (NL), 2010. Fonte em negrito indica local e ano de colheita: Reino Unido (UK), 2005. B: Biogold, F: Fontane, L: Lady Rosetta, I: Innovator, M: Maris Piper, S08: Sante analisada em 2008, S11: Sante analisada em 2011.

A — **SEM** sobreposição de local, ano de colheita nem variedade

Classificador	Conjunto de treinamento					Conjunto de teste "diferente"	
1	B (4)	L (4)	I (4)	F (4)	M (2)	**S11** (2)	**S08** (2)
2	B (4)	L (4)	I (4)	F (4)	M (2)	**S11** (2)	**S08** (2)
3	B (4)	L (4)	I (4)	F (4)	M (2)	**S11** (2)	**S08** (2)
4	B (4)	L (4)	I (4)	F (4)	M (2)	**S11** (2)	**S08** (2)
5	B (4)	L (4)	I (4)	F (4)	M (2)	**S11** (2)	**S08** (2)

B — **COM** sobreposição de local, ano de colheita e variedade

Classificador	Conjunto de treinamento					Conjunto de teste "similar"	
1	B (4)	L (4)	I (4)	F (4)	**S11** (2)	M (2)	**S08** (2)
2	B (4)	L (4)	I (4)	F (4)	**S11** (2)	M (2)	**S08** (2)
3	B (4)	L (4)	I (4)	F (4)	**S11** (2)	M (2)	**S08** (2)
4	B (4)	L (4)	I (4)	F (4)	**S11** (2)	M (2)	**S08** (2)
5	B (4)	L (4)	I (4)	F (4)	**S11** (2)	M (2)	**S08** (2)

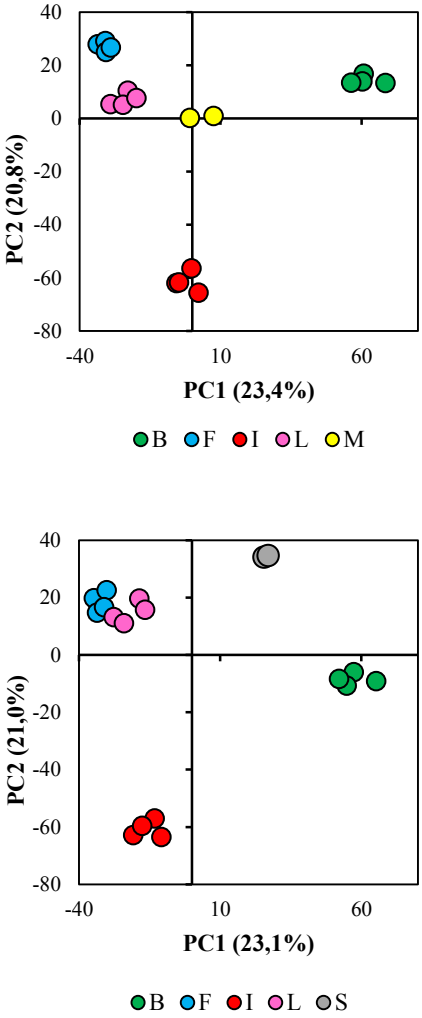
Fonte: autor.

2.4 RESULTADOS

2.4.1 Construção do classificador e validação cruzada

Os primeiros dois componentes dos gráficos de PCA das análises A e B apresentaram nítido agrupamento de acordo com a variedade (Figura 2.3). Em ambas as análises, combinações com os cinco primeiros componentes mostraram diferentes agrupamentos, todos relacionados às variedades, explicando um total de 84,3% de variação para análise A e 85,8% para análise B (ver Apêndice A). Nenhum agrupamento relacionado com outra fonte de variação foi observado para os componentes remanescentes até o décimo componente, que correspondeu a 94,9% e 95,3% de variação respectivamente para as análises A e B. Consequentemente, a validação cruzada interna baseou-se nas diferentes variedades e por isso foi realizada cinco vezes, deixando, em cada vez, uma variedade de fora. Este procedimento resultou em cinco classificadores, cada um com quatro variedades como calibradores (amostras de treinamento) e uma como amostra de validação.

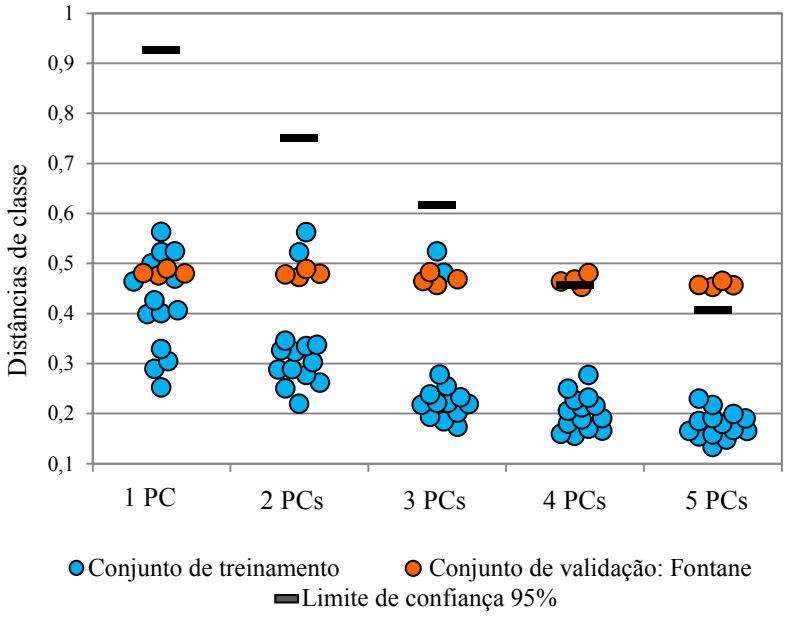
Figura 2.3: Gráficos de PCA mostrando agrupamento de acordo com variedade de batata. (A) análise A (cenário diferente), (B), análise B (cenário similar). PC: componente principal. Entre parênteses estão as porcentagens correspondentes ao total de variação explicada por cada PC. B: Biogold, F: Fontane, I: Innovator, L: Lady Rosetta, M: Maris Piper, S: Sante S11.



Fonte: autor.

Para cada um dos cinco classificadores em cada análise (A e B), foi escolhido o número ótimo de componentes a serem incluídos no classificador SIMCA. Este número foi definido com o maior número que fizesse todos os perfis do conjunto de classificação pertencerem ao patamar de base com confiança de 95%, que foi usado como limiar (*threshold*). Um exemplo de como foi feita essa escolha, na análise A para a variedade Fontane como conjunto de validação, está ilustrado na figura 2.4. Ao invés de harmonizar os cinco classificadores em um único, todos os cinco foram utilizados para analisar os conjuntos teste separadamente, resultando em cinco classificações para os conjuntos teste da análise A e cinco para a análise B.

Figura 2.4: Validação cruzada de um classificador SIMCA. São apresentadas as distâncias de classe (eixo vertical) para cada número de componentes principais (eixo horizontal) incluídas no classificador SIMCA. Como exemplo, mostra-se a validação cruzada da variedade Fontane, na análise A, cujo número ótimo de componentes foi três.



Fonte: autor.

2.4.2 Resultados da classificação

Os resultados dos classificadores SIMCA são apresentados na tabela 2.2. Na análise A, todos os perfis de teste apresentaram distâncias de classe de fato maiores do que aquelas dos perfis de treinamento, para todos os classificadores. Os perfis analisados em 2008 (S08-1158 e S08-1159), que tinham o agravante da replicata técnica como fonte de variação, mostraram maiores distâncias de classe do que os perfis analisados em 2011. Além disso, o perfil *outlier* apresentou maiores distâncias de classe em todos os classificadores. Esta amostra foi classificada como diferente da classe de base em todos os cinco classificadores, considerando 95% como limite de confiança. No entanto, para as outras três amostras teste, nove classificações foram diferentes da classe segura, podendo assim ser consideradas falso-positivas neste estudo.

Na análise B, as distâncias de classe dos conjuntos de teste ficaram mais próximas do conjunto de treinamento quando comparadas com a análise A. De fato, houve sobreposição entre as distâncias de classe dos conjuntos de treinamento e de teste em quatro dos cinco classificadores. O perfil S08-1158 (*outlier*) mostrou a maior distância de classe em todos os casos, conforme esperado. Utilizando 95% de confiança, os perfis das análises NL neste conjunto de teste foram classificados como dentro do patamar em sete de dez casos; o perfil S08-1159 como pertencente em quatro dos cinco casos; e o perfil S08-1158 como *não* pertencente também em quatro casos. Todas as distâncias de classe são apresentadas na figura 2.5, incluindo aquelas dos diferentes conjuntos de treinamento, de validação e de teste para ambas as análises.

Tabela 2.2A: Distâncias de classe dos conjuntos de teste e resultados da validação cruzada para análise A.

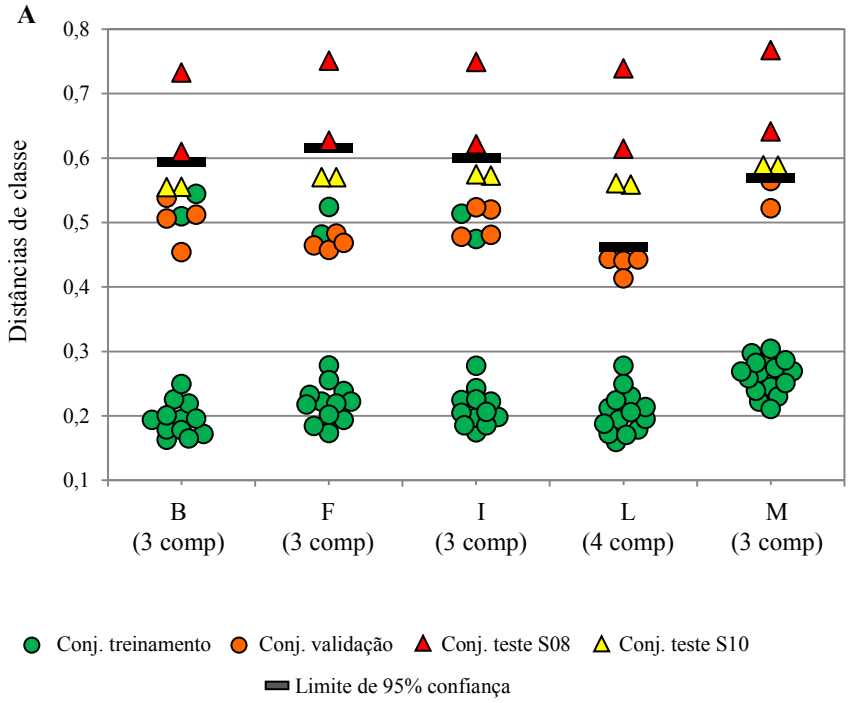
Variedade de validação cruzada	B	F	I	L	M
# componentes	3	3	3	4	3
Limite de confiança de 95%	0,594	0,616	0,599	0,462	0,569
S11-1158	0,555	0,570	0,573	**0,559**	**0,588**
S11-1159	0,555	0,570	0,575	**0,561**	**0,589**
S08-1158*	**0,733**	**0,751**	**0,749**	**0,739**	**0,767**
S08-1159	**0,610**	**0,627**	**0,621**	**0,614**	**0.641**

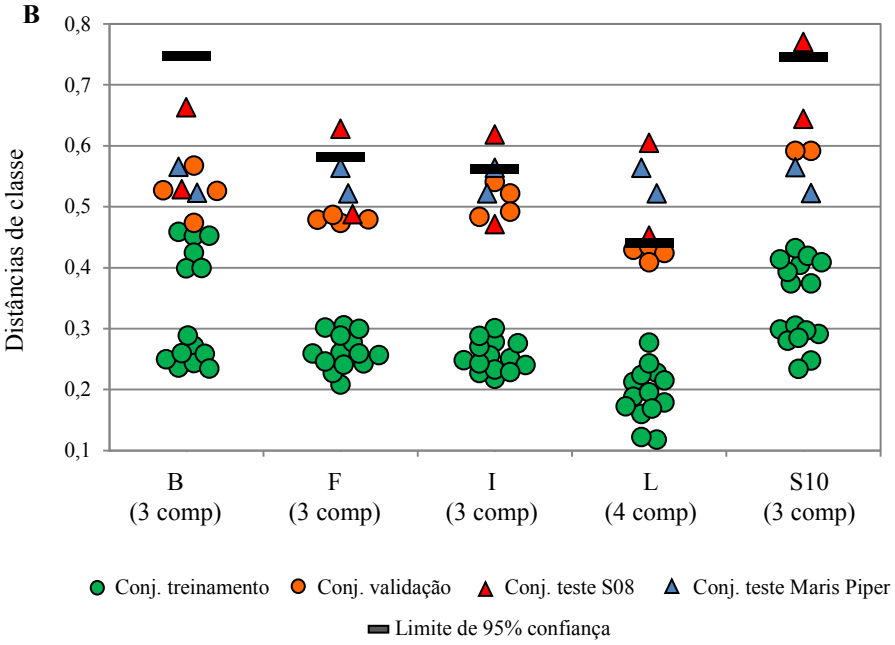
Tabela 2.2B: Distâncias de classe dos conjuntos de teste e resultados da validação cruzada para análise B.

Variedade de validação cruzada	B	F	I	L	**S11**
# componentes	2	3	3	4	2
Limite de confiança de 95%	0,747	0,581	0,562	0,441	0,745
M	0,523	0,522	0,522	**0,522**	0,523
M	0,565	0,564	**0,564**	**0,564**	0,565
S08-1158*	0,663	**0,628**	**0,618**	**0,605**	**0,770**
S08-1159	0,529	0,488	0,471	**0,452**	0.644

B: Biogold, F: Fontane, I: Innovator, L: Lady Rosetta, M: Maris Piper, S11: Sante analisada em 2011, S08: Sante analisada em 2008. * indica amostra *outlier* no estudo original. Números em negrito indicam distância de classe maior que o limite de 95%. Sombreamento em cinza indica falta de correspondência entre classificação esperada e observada.
Fonte: autor.

Figura 2.5: Distâncias de classe após classificação SIMCA para dois conjuntos teste. No quadro A (análise A), a variação extra, causada pelos diferentes anos de colheita, local de cultivo e replicata técnica não foram incluídos nos conjuntos de treinamento e validação. No quadro B (análise B) somente a replicata técnica foi deixada de fora do conjunto de treinamento, resultando em mais amostras sendo classificadas como pertencentes à classe. Para ambos os conjuntos teste, todos os cinco sub-modelos resultantes da validação cruzada são mostrados com as letras indicando o conjunto de treinamento. Entre parênteses o número ótimo de componentes principais em cada validação cruzada. B: Biogold, F: Fontane, I: Innovator, L: Lady Rosetta, M: Maris Piper, S08: Sante analisada em 2008, S11: Sante analisada em 2011.





Fonte: autor.

2.5 DISCUSSÃO

Um conjunto de perfis de transcriptomas de amostras de batatas foi usado para explorar as possibilidades de classificação por análise multivariada supervisionada como ferramenta para avaliação de segurança de alimentos para novas variedades. Um dos objetivos foi elucidar o modo como diferentes fontes de variação influenciariam na classificação. Como fontes de variação, foram incluídos: variedade de batata, local de cultivo, ano de colheita e replicata biológica. Uma amostra "extrema" foi incluída como representante que se esperaria ser classificada fora do patamar na maioria dos casos, permitindo assim erros de classificação devido ao limite de confiança de 95%. Para maior clareza, esta amostra (S08-1158) foi considerada um *outlier* neste trabalho, apesar de não ter sido considerada *outlier* no estudo original (VAN DIJK et al., 2012), já que esta amostra

ainda fazia parte do agrupamento no gráfico de PCA de acordo com os grupos de tratamento naquele estudo.

As análises A e B combinadas permitiram salientar a importância de incluir suficiente variação representativa para a construção de um classificador para evitar a presença de falsos-positivos. Na análise A, nove das 20 classificações individuais foram falso-positivas à luz da avaliação de segurança. Classificação fora da classe segura foi baseada nas diferenças combinadas de ano e local de colheita e variedade, e não em alguma amostra realmente insegura. Na análise B, distâncias de classe para amostras de teste e de treinamento ficaram mais próximas entre si e, em alguns casos, observou-se até sobreposição dos valores, exceto para a amostra *outlier* S08-1158. A razão para isso é que as amostras de treinamento utilizadas para construir esse classificador eram mais representativas para todas as amostras deste conjunto particular de dados.

Os conjuntos de dados então disponíveis não eram representativos da verdadeira população de variedades de batatas disponível no mercado. Consequentemente, esses resultados não são ainda suficientes para inferir probabilidades de falsos positivos e negativos. No entanto, é informativo sobre como as taxas de erro calculadas são dependentes da metodologia escolhida. Por exemplo, a variedade Maris Piper foi classificada como não estando dentro do patamar em três dos 10 classificadores, apesar de ser uma batata conhecidamente segura. Consequentemente, poderia se dizer que existiram três observações falso-positivas para este cultivar. Por outro lado, se uma classificação média por amostra fosse considerada, a maioria dos classificadores ajustaram ambas as amostras desta variedade como pertencente ao patamar, indicando dois verdadeiros negativos e nenhum falso positivo ou falso negativo. Especialmente para o perfil de amostra mais diferente neste trabalho, S08-2011, a classificação ficou fora do patamar para todos os classificadores na análise A e em quatro (de cinco) da análise B. Tal resultado levaria a maiores investigações dos dados em uma avaliação de segurança de alimentos. Ou seja, nesta configuração, a variedade Sante claramente não seria considerada como não pertencente à classe segura baseada somente nesta amostra, uma vez que as outras amostras Sante foram menos diferentes. Ao contrário, isso prova que uma amostra *outlier* vai ser identificada como tal. Transportando para uma situação real, isso significa que se: 1) todas as replicatas de uma nova variedade caíssem fora da classe segura e 2) as amostras na classe segura fossem representativas de uma variação normal, essa nova variedade exigiria de fato investigação mais aprofundada.

Este estudo mostra que, a princípio, perfis de transcriptoma podem ser usados para classificar tubérculos de batatas como pertencentes ou não

pertencentes a um grupo conhecido de batatas. Essa classificação pode formar a base para identificar potenciais perigos em novas variedades de batatas, por exemplo, no caso de efeitos não intencionais de técnicas de melhoramento, incluindo GM. Sobre esse aspecto, é importante notar que a identificação de perigo de novas variedades conforme apresentado aqui está no contexto do cumprimento dos requerimentos regulatórios mundiais. Quando esta abordagem for mais desenvolvida e validada, será provavelmente mais informativa e eficaz em termos de custos do que os testes de alimentação feitos em animais atualmente obrigatórios na União Europeia para identificar efeitos não intencionais em novas plantas.

Os classificadores apresentados aqui servem como prova de princípio. Para aplicações práticas e validação da classificação multivariada de uma classe, é preciso determinar o correto *threshold* para associação de classe segura. Na avaliação de segurança, falsos negativos são mais preocupantes do que os falsos positivos. Falsos negativos podem levar ao aumento do risco, enquanto falsos positivos apenas aumentarão desnecessariamente a necessidade de análises toxicológicas pós-classificação para os perfis classificados fora da classe segura. Um estudo bem definido, contendo amostras não seguras conhecidas no conjunto de teste, assim como maior quantidade de amostras seguras conhecidas, auxiliará na determinação deste valor *threshold* e na metodologia proposta como um todo. Essa abordagem com o classificador ainda é basicamente a mesma que a atualmente utilizada (análise-alvo) para análise composicional de novas variedades de plantas. No entanto, a utilização de perfis provenientes de análises ômicas aumentará significativamente o conteúdo de informação subjacente à identificação de perigos.

O presente estudo propõe uma expansão da atual análise-alvo para uma abordagem não-alvo composicional comparativa, baseada em perfis ômicos, para identificação de perigos na avaliação de segurança de alimentos. Propõe, ainda, a aplicação de classificação multivariada de uma classe para identificação de perigo, dependendo se os perfis de novas variedades de plantas caírem dentro ou fora de uma classe de perfis geralmente reconhecidos como seguros. A correspondência entre os resultados esperados e os observados do conjunto de classificadores deste trabalho abre as portas para desenvolvimento de mais estudos e validação desta abordagem.

CAPÍTULO 3

ANÁLISE DE COMPONENTES PRINCIPAIS COM DADOS DE PROTEÔMICA DE BATATAS VISANDO AVALIAÇÃO COMPLEMENTAR DE ALIMENTOS

Artigo submetido para publicação:
MELLO, C. S., VAN DIJK, J.P., VOORHUIJZEN, KOK, E. J., ARISI, A. C. M. Principal component analysis of potato proteomic data aiming complementary food assessment. ***New Biotechnology*** (2013).

3 ANÁLISE DE COMPONENTES PRINCIPAIS COM DADOS DE PROTEÔMICA DE BATATAS VISANDO ANÁLISE COMPLEMENTAR DE SEGURANÇA DE ALIMENTOS

PRINCIPAL COMPONENT ANALYSIS OF POTATO PROTEOMIC DATA AIMING COMPLEMENTARY FOOD ASSESSMENT

Carla Souza de Mello[1], Jeroen van Dijk[2], Marleen Voorhuijzen[2], Esther Kok[2], Ana Carolina Maisonnave Arisi[1]*

[1] Departamento de Ciência e Tecnologia de Alimentos, Centro de Ciências Agrárias, Universidade Federal de Santa Catarina, Rod. Admar Gonzaga, 1346, 88034-001, Florianópolis, SC, Brazil.
[2] RIKILT – Institute of Food Safety, Wageningen University and Research centre, P.O. Box 230, 6700 AE Wageningen, The Netherlands.

RESUMO

Análise de avaliação de segurança são necessárias para atestar o consumo seguro de novos produtos alimentares. Análises-alvo são realizadas rotineiramente com esse fim, no entanto elas são limitadas e podem deixar de fora possíveis efeitos não-intencionais ou não-esperados. Técnicas amplas de perfil conhecidas como "ômicas" tem sido sugeridas como abordagens não-alvo para detecção de efeitos não-intencionais para avaliação de novos alimentos. Uma dessas técnicas é a análise proteômica. A análise dos dados gerados pelas ômicas deve ser feita por análise multivariada que, ao contrário da univariada, permite obtenção de mais informações a partir do conjunto de dados. Neste trabalho, perfis de proteômica de cinco variedades de batatas foram avaliados por eletroforese bidimensional (2-DE) utilizado dois comprimentos de tiras de gradiente de pH imobilizado (IPG): 13 e 24 cm, ambos com abrangência de pH de 4 a 7. Para cada tamanho de tira, dois géis foram feitos para cada variedade; no total, foram 10 géis por análise. O propósito foi esclarecer o comportamento de perfis proteômicos sobre a separação entre as amostras quando submetidas à análise de componentes principais (PCA). Eventuais valores omissos de cada perfil foram eliminados do conjunto de dados ou substituídos pelo mínimo valor detectado. Para as tiras de 13 cm, 29 ou 740 spots em comum a todos os géis foram detectados quando os valores omissos foram eliminados ou substituídos, respectivamente. Para as tiras de 24 cm, esses valores foram 43 e 1756. Os quatro diagramas de PCA feitos com esses conjuntos de dados mostraram claro agrupamento de amostras de acordo com as variedades. Os dados apresentados aqui indicam que PCA é aplicável para análise proteômica de batatas e é capaz de separar as amostras por variedade. Mais variação e amostras devem, no entanto, ser incluídas para maiores investigações de análise de segurança de alimentos.

Palavras-chave: PCA. Eletroforese bi-dimensional. 2-DE. Valores omissos. Proteoma. Análise de alimentos

ABSTRACT

Safety assessment analyses are necessary to attest harmless consumption of new food products. Targeted analyses are routinely performed for that purpose, however they are limited and may leave out relevant unintended or unexpected effects. Profiling techniques known as “omics” have been suggested as non-targeted approaches for detection of unintended effects for novel food assessment. One of these techniques is the proteomics analysis. Data analysis of omics data should be performed by multivariate analysis, instead of univariate statistics, in order to get most information from the datasets. On the present work, proteomic profiles of five potato varieties were evaluated by 2-DE performed on two IPG strip lengths: 13 and 24 cm, both under pH range of 4-7. For each strip length, two gels were made from each variety; in total there were 10 gels per analysis. The purpose was to provide insight into the behavior of proteomic profiles concerning the separation of samples when data is analyzed by PCA. Possible missing values from each profile were eliminated from the dataset or substituted by the minimum value detected. Using strips of 13 cm, 29 or 740 spots in common to all gels were detected when the missing values were eliminated or substituted, respectively. For 24 cm strips, these amounts corresponded to 43 and 1756. The four PCAs performed with these datasets presented clear grouping of samples according to the varieties. The data presented here show that PCA is applicable for proteomic analysis of potato and is able to separate the samples by varieties. More variation and samples should be included for further investigation of food safety assessment.

Keywords: PCA. Two-dimensional electrophoresis. 2-DE. Missing values. Proteome. Food assessment.

3.1 INTRODUÇÃO

A avaliação da segurança de alimentos é um assunto amplamente discutido e que envolve todos os tipos de novos alimentos antes que eles cheguem até os consumidores. Quando se trata de novas variedades de plantas, sejam elas culturas convencionais ou geneticamente modificadas (GM), podem ocorrer efeitos indesejados ou inesperados (KUIPER et al., 2013). Análises amplas de perfil como a transcriptômica, a proteômica e a metabolômica têm sido recomendados como abordagens não-alvo para a detecção de efeitos não intencionais das plantas (ANTTONEN et al., 2010; CELLINI et al., 2004; KOK; KEIJER; et al., 2008; METZDORFF et al., 2006; VAN DIJK et al., 2010). Geralmente, essas análises são aplicadas como ferramentas complementares às avaliações de segurança já existentes, tais como análises-alvo (D'ALESSANDRO; ZOLLA, 2012). Esses métodos permitem um enfoque mais completo sobre possíveis mudanças não previstas no metabolismo da planta, que não podem ser detectadas em abordagens específicas, como análises de compostos isolados (KUIPER; KOK; ENGEL, 2003).

Em geral existem diversos métodos quantitativos já estabelecidos na análise proteômica de plantas (BINDSCHEDLER; CRAMER, 2011; VAN WIJK, 2001), e a eletroforese bidimensional (2-DE) é uma das técnicas mais utilizadas (RUEBELT et al., 2006a; RUEBELT et al., 2006b; RUEBELT et al., 2006c; ZOLLA et al., 2008). A análise proteômica utilizando o método 2-DE já demonstrou capacidade de caracterizar e diferenciar variedades de acordo com seus níveis e padrões de acúmulo de proteínas em diversas plantas como soja (BRANDAO; BARBOSA; ARRUDA, 2010), milho (ANTTONEN et al., 2010; BALSAMO et al., 2011; COLL et al., 2011; ZOLLA et al., 2008), feijão (DE LA FUENTE et al., 2011), arroz (TIAN et al., 2009) e trigo (GE et al., 2012), entre outros. A técnica também é capaz de fornecer informações moleculares para classificação de diversas culturas agronômicas (AGRAWAL; RAKWAL, 2006). Na análise de batatas ela já foi descrita para caracterização de proteínas envolvidas no estresse hídrico (ZERZUCHA et al., 2012) e abiótico (FOLGADO et al., 2013) ou mesmo para avaliar estádios de desenvolvimento dos tubérculos (LEHESRANTA et al., 2006).

A técnica de 2-DE requer um protocolo de extração de proteínas bem estabelecido seguido pela separação do extrato proteico na primeira e na segunda dimensão. Primeiramente, os polipeptídeos são separados pelo seu ponto isoelétrico (pI) e, em seguida, é feita a separação de acordo com suas massas moleculares (RUEBELT; 2006a). A digitalização dos géis gera mapas tridimensionais contendo proteínas (*spots*) que são analisados por

softwares específicos e através destes são atribuídos valores de intensidade de volume a cada *spot*. Quando se analisa mais de uma amostra biológica, existe a necessidade de se combinar os mesmos *spots* de proteínas entre os diferentes géis (ALBRECHT et al., 2010); no entanto, frequentemente são detectados valores omissos, que geralmente são devidos a limitações experimentais (CHICH et al., 2007). Valores omissos são problemáticos tanto porque representam perda de informação sobre os padrões proteicos como também podem gerar dificuldades para as análises de dados subsequentes (GROVE et al., 2006).

A análise de perfil por proteômica gera uma grande quantidade de dados que correspondem ao grupo de proteínas simultaneamente encontradas em determinada amostra biológica. Os dados de proteômica podem ser avaliados por estatística univariada (t-teste, Kruskal-Wallis, ANOVA). No entanto, a interpretação dos resultados gerados por estes métodos aumentam a chance de detecção de falsos-positivos, além de serem afetados pela estrutura dos dados brutos e não detectarem tendências e nem relações entre as proteínas (CHICH et al., 2007). Abordagens de estatística multivariada são mais efetivas para este tipo de análise, pois permitem a redução da complexidade dos dados, prevenção de tendências e podem ser menos afetados pela estrutura dos dados (GROVE et al., 2008). A estatística multivariada exerce um papel importante na análise de dados ômicos porque muito mais informações podem ser extraídas quando comparado aos testes univariados. Muitos pesquisadores recomendam que se tratem os dados de proteômica simultaneamente por meios de estatística multivariada (GOTTLIEB et al., 2004; GROVE et al., 2006; VALLEDOR; JORRIN, 2011). A análise de componentes principais (PCA) é uma ferramenta de análise multivariada não-supervisionada utilizada para transformar um conjunto de variáveis observadas em um novo conjunto de novas variáveis não-correlacionadas, levando à redução da dimensionalidade pelo desenvolvimento de novos eixos, chamados de componentes principais (PC) (JANES; YAFFE, 2006). O uso de estatística multivariada como a PCA em dados de proteômica pode, no entanto, ser comprometida pela presença de valores omissos no conjunto de dados. Não existem regras específicas para tratar dados omissos (PEDRESCHI et al., 2008), mas alguns pesquisadores sugerem alternativas para a exclusão de todas as variáveis contendo valores omissos ou a substituição destes dados pela inserção de outros valores (ZELLNER et al., 2012). Contudo, simplesmente remover as variáveis reduziria o número total de *spots* consideravelmente, uma vez que dados de 2-DE chegam a apresentar cerca de 50% de dados omissos (KROGH et al., 2007). A inserção de valores pode ser uma opção mais adequada uma vez que algoritmos específicos

podem ser aplicados para estimar os valores omissos, baseados nos *spots* já existentes.

No presente trabalho perfis proteômicos de cinco variedades de batatas foram avaliados por 2-DE realizada em duas condições diferentes de isoeletrofocalização (tiras de IPG de 13 ou 24 cm, ambas na faixa de pH 4-7). O objetivo foi elucidar como as amostras comportam-se quando os dados fossem visualizados em diagramas de PCA, tendo seus valores omissos substituídos ou simplesmente removidos do conjunto de dados. As implicações dos resultados são discutidas com vistas à análise de segurança de alimentos.

3.2 MATERIAL E METODOS

3.2.1 Material vegetal

Cinco variedades de batatas (Biogold, Fontane, Innovator, Lady Rosetta and Maris Piper) foram analisadas por 2-DE. As amostras e o pré-tratamento dos tubérculos foram os mesmos utilizados na análise transcriptômica descrita no Capítulo 2.

3.2.2 Extração de proteína total solúvel

Proteína total solúvel foi extraída com base no procedimento descrito por Koistinen e colaboradores (2002) com modificações. Um total de 260 mg de batata moída liofilizada foi usada para cada amostra. O pó foi vigorosamente misturado a 1 mL de tampão de lise (ureia a 7 M, tioureia a 2 M, CHAPS a 4% (p/v), Triton-X a 1% (v/v) e DTT a 14 mM) contendo inibidores de protease (PMSF a 1 mM e Mini Complete Tablets 1X (Roche, Mannheim, Alemanha)), DNAse I a 1 U/mL (Sigma Aldrich, St. Louis, MO, EUA) e RNAse a 20 mg/mL (Invitrogen, St. Louis, MO, EUA). A mistura foi mantida por 10 min à temperatura ambiente e então centrifugada (13.000 × g por 10 min a 4°C). O sobrenadante foi transferido para um novo tubo e a mesma centrifugação repetida. O extrato foi então purificado com o kit comercial 2-D Clean-Up (GE Healthcare, Uppsala, Suécia) de acordo com as instruções do fabricante. O precipitado final foi ressuspendido em 150 µL de tampão de reidratação (7 M ureia, 2 M tioureia, 2% (w/v) CHAPS, 2% (v/v) IPG Buffer, 2,8 mg/µL DTT, 0,002% (p/v) azul de bromofenol, 12 µL/mL DeStreak Reagent (GE Healthcare)). As extrações foram realizadas em duplicata de 10 amostras em paralelo

(cinco variedades cultivadas em solo arenoso e cinco em argiloso) em dois dias diferentes (total de 20 extratos). A proteína total solúvel foi quantificada com o 2-D Quant Kit (GE Healthcare), seguindo o manual do fabricante.

3.2.3 Eletroforese bidimensional

A 2-DE foi feita com um gel para cada extrato proteico. A focalização isoelétrica (IEF) foi conduzida em duas condições (20 tiras em cada condição): primeiro usando tiras de IPG de 13 cm e depois de 24 cm, ambas com intervalo de pH de 4-7. (GE Healthcare), que serão referenciadas neste trabalho como "análise I" e "análise II", respectivamente. Para a análise I, a eletroforese foi realizada com quatro géis das amostras da mesma variedade numa mesma corrida (4 géis/corrida). Na análise II, cada corrida eletroforética continha um gel de uma amostra de cada variedade, num total de cinco géis por corrida.

As tiras de 13 cm (análise I) foram reidratadas durante 18 h em uma solução contendo 350 μg de proteína total diluída no tampão de reidratação (mesmo descrito anteriormente) contendo 2% (v/v) IPG Buffer pH 4-7 (GE Healthcare) num volume total de 250 μL. Após a rehidratação, as tiras foram focalizadas em equipamento Ettan IPGphor 3 Isoelectric Focusing System (GE Healthcare) sob as seguintes condições: passo de 500 V até 500 Vh, gradientes de voltagem de 1.000 V e 8.000 V até 14.500 Vh, e um passo final de 8.000 V até 17.800 Vh, acumulando um total de 34.000 Vh, mantendo-se o limite de 50 mA/tira. Em seguida à focalização, as tiras foram mantidas a -70°C por pelo menos 18 h. As proteínas imobilizadas nas tiras passaram por uma etapa de redução com 10 mg/mL DTT em tampão de equilíbrio (ureia a 6 M, Tris-HCl a 50 mM (pH 8,8), glicerol a 30% (v/v), SDS a 2% (w/v) e azul de bromofenol a 0,002% (w/v)), seguida de alcalinização com iodoacetamida a 25 mg/mL em tampão de mesma composição. As tiras foram então colocadas no topo de géis de poliacrilamida 12,5% (p/v) de medidas 18 cm × 16 cm × 1,5 mm; foi utilizado marcador de M_r (Precision Plus Protein Dual Color Standards, BioRad). A segunda dimensão foi realizada no equipamento SE 600 Ruby System (GE Healthcare) com corrente de 5 mA/gel durante a primeira hora e 30 mA/gel até que o corante atingisse aproximadamente 2 mm da base do gel. A temperatura foi mantida a 10 °C utilizando MultiTemp III Thermostatic Circulator (GE Healthcare).

As tiras de 24 cm (análise II) foram reidratadas com 500 μg de proteína total diluída em tampão de reidratação (450 μL), também contendo

2% (v/v) IPG Buffer pH 4-7 (GE Healthcare). Grande parte do procedimento foi idêntico ao descrito para a análise I, com as seguintes modificações: focalização isoelétrica foi realizada até um acúmulo total de 55.000 Vh; a segunda dimensão da eletroforese foi feita em géis com dimensões 26 cm × 16 cm × 2 mm no equipamento Ettan DALTsix Large Vertical System (GE Healthcare) a 25 °C, sob corrente de 20 mA/gel na primeira hora e 24 mA/gel até que o corante chegasse ao final do gel.

3.2.4 Captura de imagem e análise dos dados

Proteínas (*spots*) foram visualizadas pela coloração dos géis com 0,1% (p/v) de Coomassie Brilliant Blue G-250 (Bio-Rad, Herculer, CA, EUA) como já descrito por Balsamo e colaboradores (2011). Os géis foram digitalizados usando o Image Scanner System II e analisados com o programa ImageMaster 2-D Platinum Software v. 7.0 (ambos GE Healthcare). Para a detecção dos *spots* de proteínas no gel foram utilizados os seguintes parâmetros no programa: rugosidade ≥ 5, área mínima ≥ 4 e saliência ≥ 100. A combinação automática entre os *spots* nos géis foi complementada por correção manual quando necessário. Volumes relativos dos *spots* (%Vol) foram comparados e analisados. Géis pertencentes à mesma análise (I ou II) foram comparados por solo (dois géis) e também por variedade (quatro géis). Finalmente, todos os géis foram comparados simultaneamente (20 géis). Todos os *spots* correspondentes em pelo menos dois géis (do total de 20) foram selecionados e transformados em log2. Estes dados foram utilizados para a análise de PCA na linguagem R (R CORE TEAM, 2013) empregando a função “prcomp” do pacote “gdata”.

3.3 RESULTADOS

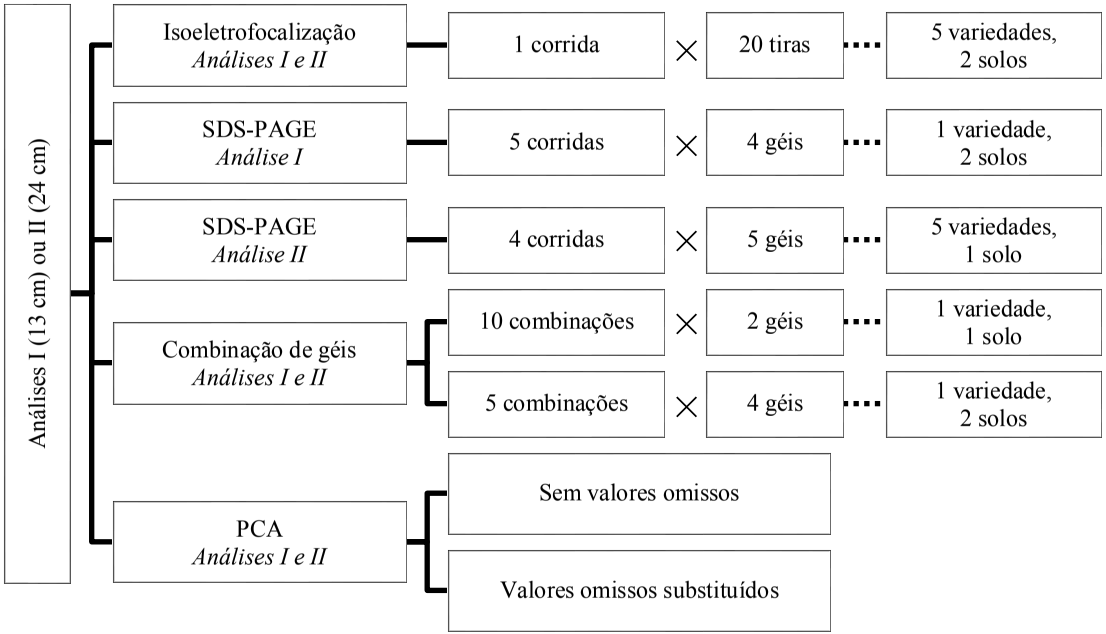
Extratos de proteína total solúvel de cinco variedades de batatas cultivadas em dois diferentes substratos (solo argiloso e arenoso) foram separados por 2-DE. A separação por ponto isoelétrico nas tiras de IPG de 13 cm (análise I) e 24 cm (análise II) foi realizada para comparação do número de *spots* detectados e da habilidade de demonstrar separação das amostras em cada análise quando observados em gráficos de PCA. O procedimento realizado para ambas as análises, desde a focalização isoelétrica até a PCA, está resumida na figura 3.1. Houve diferença entre os procedimentos tomados para cada análise devido às diferentes condições dos equipamentos: para os géis menores (análise I), o sistema Ruby permite

que até quatro géis corram simultaneamente, então quatro géis da mesma variedade (duas replicatas das amostras cultivadas em solo argiloso e duas em arenoso) foram usados em cada corrida; para os géis maiores (análise II), o sistema DALTsix permite até seis géis por vez, então um total de cinco géis foram usados em cada corrida, um gel de cada variedade, todas amostras cultivadas no mesmo solo.

Dois géis da variedade Biogold cultivada em areia, representativos das corridas, são apresentados na figura 3.2 (mapas de todas as variedades deste estudo são apresentados no Apêndice B). As corridas realizadas em géis de 24 cm têm maior quantidade e melhor separação dos *spots*. O número de *spots* de proteínas detectadas e *spots* correspondentes (*matched spots*) são apresentados nas tabelas 3.1 e 3.2. Para análise I, um total de 267 ± 28 (média ± DP) *spots* foram detectados. Na análise II, foi observado um aumento de 54% no número de *spots* (em média, 497 ± 84).

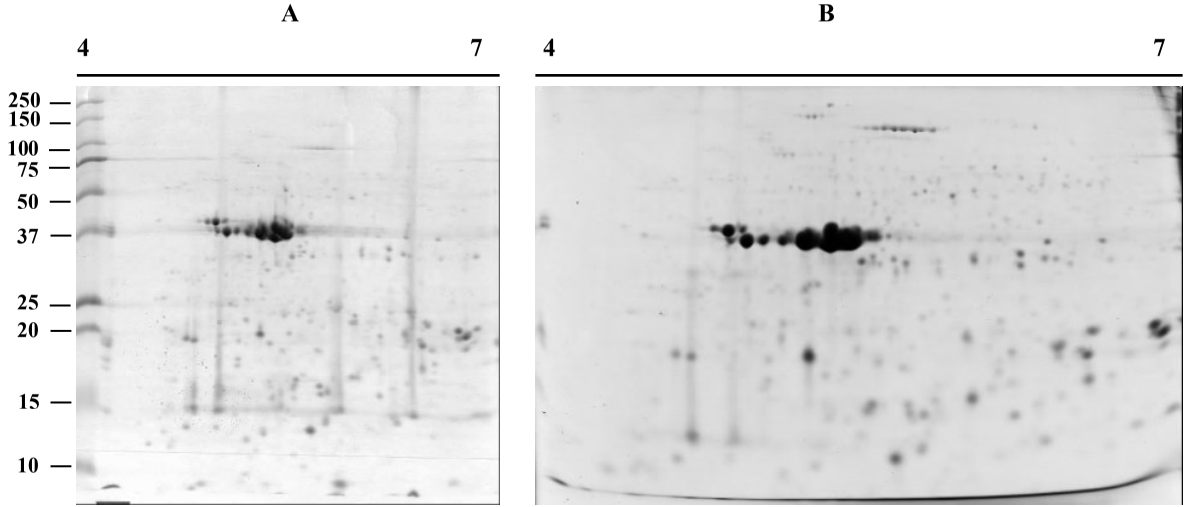
Para cada análise, foram realizadas três configurações diferentes nas correspondências entre os géis. A primeira correspondência foi feita entre dois géis da mesma variedade cultivada no mesmo substrato (Correspondência por Solo); esta análise representa a similaridade entre as replicatas. O número médio de *spots* correspondentes para géis de 13 cm e 24 cm foi de 209 ± 28 e 392 ± 62, respectivamente (tabelas 3.1 e 3.2). Os coeficientes de correlação (r) apresentaram-se acima de 0,94, exceto na combinação dos géis FZ1 x FZ2 de 13 cm, cujo r foi 0,88.

Figura 3.1: Esquema descritivo da separação e análise de proteínas utilizando tiras de IPG pH 4-7 de 13 cm (análise I) e 24 cm (análise II). As corridas em SDS-PAGE apresentam diferentes desenhos devido às diferenças entre os equipamentos.



Fonte: autor.

Figura 3.2: Mapas representativos do perfil de proteínas da variedade de batata Biogold cultivada em solo arenoso (BZ). Proteínas foram separadas em tiras de focalização linear IPG pH 4-7 de comprimento 13 cm (A) e 24 cm (B) na primeira dimensão e em géis 12,5% SDS-PAGE na segunda dimensão. Géis foram corados com Coomassie Brilliant Blue G-250. A faixa de pH está indicada horizontalmente no topo dos géis e a massa molecular (kDA) correspondente está especificada verticalmente ao lado dos géis.



Fonte: autor.

Tabela 3.1: Dados de correspondência de *spots* de batatas (número de *spots*) detectados em géis de 2-DE corados com Coomassie Brilliant Blue G-250 após separação das proteínas em tiras de IPG pH 4-7 de 13 cm seguida de gel 12,5% SDS-PAGE (análise I). Em negrito são os géis referência (aqueles com maior número de *spots*) em cada categoria de correspondência.

Variedade	Solo	Amostra	*Spots* detectados	Correspondência por solo			Correspondência por variedade		
				Spots correspondentes (% dos detectados)	Inclinação	r	*Spots* correspondentes (% dos detectados)	Inclinação	r
Biogold	Argiloso	BK1	276	**219 (79%)**	0,923	0,971	223 (81%)	1,040	0,934
		BK2	256	219 (86%)			237 (93%)	0,961	0,972
	Arenoso	BZ1	323	**252 (78%)**	0,880	0,985	**258 (80%)**	-----	-----
		BZ2	294	252 (86%)			262 (89%)	0,880	0,985
Fontane	Argiloso	FK1	320	**256 (80%)**	0,985	0,938	**257 (80%)**	-----	-----
		FK2	292	256 (88%)			273 (93%)	0,985	0,938
	Arenoso	FZ1	264	215 (81%)	0,826	0,879	241 (91%)	0,968	0,938
		FZ2	267	**215 (81%)**			216 (81%)	1,180	0,927
Innovator	Argiloso	IK1	278	**211 (76%)**	0,827	0,961	**227 (82%)**	-----	-----
		IK2	249	211 (85%)			223 (90%)	0,827	0,961
	Arenoso	IZ1	254	200 (79%)	1,040	0,980	212 (83%)	0,933	0,985
		IZ2	267	**200 (75%)**			215 (81%)	0,892	0,976

Continua

Continuação Tabela 3.1

Lady Rosetta	Argiloso	LK1	261	**169 (65%)**	0,823	0,957	**187 (72%)**	-----	-----
		LK2	199	169 (85%)			181 (91%)	0,823	0,957
	Arenoso	LZ1	230	169 (73%)	0,941	0,974	173 (75%)	0,962	0,971
		LZ2	247	**169 (68%)**			204 (83%)	0,898	0,948
Maris Piper	Argiloso	MK1	250	196 (78%)	1,120	0,986	211 (84%)	1,120	0,986
		MK2	288	**196 (68%)**			**215 (75%)**	-----	-----
	Arenoso	MZ1	261	201 (77%)	1,100	0,975	210 (80%)	1,150	0,976
		MZ2	261	201 (77%)			219 (84%)	1,040	0,986

Fonte: autor.

Na segunda correspondência (Correspondência por Variedade), quatro géis da mesma variedade cultivada nos dois substratos foram comparados para verificar se a diferença entre os solos de cultivo influenciaria no número total de *spots*. A análise I apresentou uma média de 222 ± 27 *spots* correspondentes e a análise II, 434 ± 78. À exceção de somente uma combinação de géis, todos os géis correlacionados apresentaram coeficientes de correlação maiores que 0,93. A exceção foi FK2 x FZ2, com r = 0,85.

Tabela 3.2: Dados de correspondência de *spots* de batatas (número de *spots*) detectados em géis de 2-DE corados com Coomassie Brilliant Blue G-250 após separação das proteínas em tiras de IPG pH 4-7 de 24 cm seguida de gel 12,5% SDS-PAGE (análise II). Em negrito são os géis referência (aqueles com maior número de *spots*) em cada categoria de correspondência.

Variedade	Solo	Amostra	*Spots* detectados	Correspondência por solo			Correspondência por variedade		
				Spots correspondentes (% dos detectados)	Inclinação	r	*Spots* correspondentes (% dos detectados)	Inclinação	r
Biogold	Argiloso	BK1	439	378 (86%)	1,030	0,971	378 (86%)	1,030	0,971
		BK2	542	**378 (70%)**			**473 (87%)**	---	---
	Arenoso	BZ1	381	341 (90%)	0,898	0,968	341 (90%)	0,993	0,980
		BZ2	479	**341 (71%)**			434 (91%)	1,100	0,970
Fontane	Argiloso	FK1	471	425 (90%)	1,000	0,948	425 (90%)	1,000	0,948
		FK2	601	**425 (71%)**			**509 (85%)**	---	---
	Arenoso	FZ1	418	358 (86%)	1,220	0,944	358 (86%)	0,940	0,941
		FZ2	491	**358 (73%)**			437 (89%)	0,789	0,848
Innovator	Argiloso	IK1	365	305 (84%)	0,956	0,964	305 (84%)	0,847	0,956
		IK2	411	**305 (74%)**			369 (90%)	0,894	0,974
	Arenoso	IZ1	524	422 (81%)	1,130	0,955	422 (81%)	1,130	0,955
		IZ2	563	**422 (75%)**			**458 (81%)**	---	---

Continua

Continuação Tabela 3.2

Lady Rosetta	Argiloso	LK1	451	388 (86%)	0,830	0,974	388 (86%)	0,830	0,974
		LK2	538	**388 (72%)**			**468 (87%)**	---	---
	Arenoso	LZ1	422	342 (81%)	1,040	0,969	342 (81%)	1,130	0,962
		LZ2	483	**342 (71%)**			426 (88%)	1,080	0,982
Maris Piper	Argiloso	MK1	580	526 (91%)	0,896	0,978	526 (91%)	0,896	0,978
		MK2	684	**526 (77%)**			**615 (90%)**	---	---
	Arenoso	MZ1	477	437 (92%)	0,962	0,981	437 (92%)	1,050	0,978
		MZ2	621	**437 (70%)**			561 (90%)	1,080	0,986

Fonte: autor.

A terceira combinação de géis foi feita pela verificação dos pontos correspondentes existentes em todos os 20 géis de cada análise. Esse procedimento confere quantos *spots* detectados em todos os géis são correlacionados. Respectivamente para as variedades Biogold, Fontane, Innovator, Lady Rosetta e Maris Piper, os números de *spots* correlacionados foram: na análise I, 303, 294, 256, 225 e 260 (tabela 3.3); e na análise II 551, 579, 563, 547 e 709 *spots* (tabela 3.4).

As porcentagens de *spots* correspondentes detectados nos géis da mesma variedade quando todos os 20 géis são comparados simultaneamente são mostrados nos gráficos da figura 3.3. Os gráficos mostram que cerca de 60% dos *spots* correspondentes estão presentes em todos os géis da mesma variedade para análise I (figura 3.3A) e 45% para análise II (figura 3.3B). Essas porcentagens decrescem quando menor número de géis é considerado. No entanto, quando dois géis são comparados, existe um número ligeiramente maior de *spots* correspondentes para análise II em comparação à análise I. Por isso, foi feito o teste t com o intuito de verificar se este número maior de *spots* deve-se à diferença entre os solos (uma vez que são dois géis para cada solo) (tabelas 3.3 e 3.4).

Tabela 3.3: Número (e porcentagem) de *spots* correspondentes (corresp.) detectados em géis da mesma variedade quando combinados todos os 20 géis da análise I (13 cm). Números de *spots* significativamente relacionados e diferentes de acordo com solo de cultivo são indicados na coluna direita.

Variedade	*Spots* corresp. entre todas as variedades[1] (%)	*Spots* corresp. presentes em todos os géis da mesma variedade (%)	*Spots* corresp. presentes em 3 géis (%)	*Spots* corresp. presentes em 2 géis (%)	*Spots* corresp. presentes em 1 gel (%)	t-teste Areia/Argila*	
						Spots relacionados	*Spots* diferentes
Biogold	303 (100)	194 (64)	26 (9)	63 (21)	20 (7)	21	173
Fontane	294 (100)	192 (65)	31 (11)	63 (21)	8 (3)	5	187
Innovator	256 (100)	168 (66)	43 (17)	38 (15)	7 (3)	11	157
Lady Rosetta	225 (100)	127 (56)	53 (24)	39 (17)	6 (3)	4	123
Maris Piper	260 (100)	157 (60)	45 (17)	46 (18)	12 (5)	8	149

[1] presente em pelo menos uma das replicatas.
[2] t-teste, $p < 0.05$

Fonte: autor.

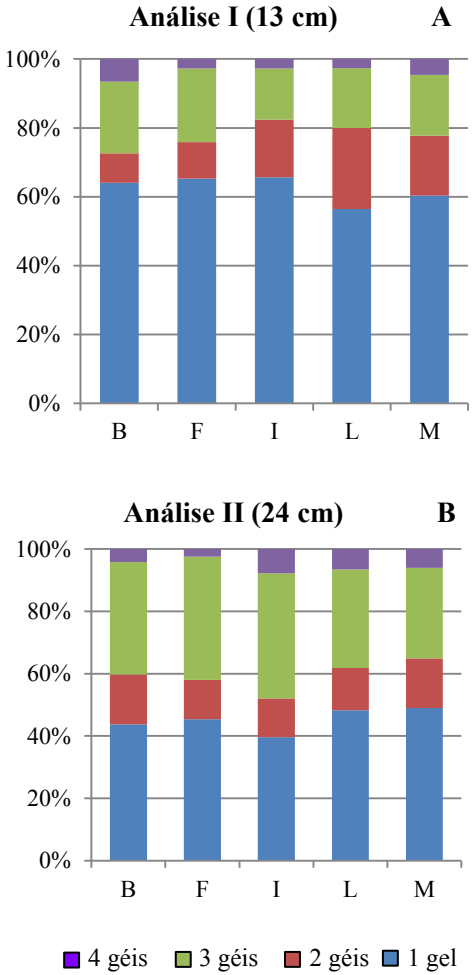
Tabela 3.4: Número (e porcentagem) de *spots* correspondentes (corresp.) detectados em géis da mesma amostra biológica quando combinados todos os 20 géis da análise II (24 cm). Números de *spots* significativamente relacionados e diferentes de acordo com solo de cultivo são indicados na coluna direita.

Variedade	*Spots* corresp. entre todas as variedades[1] (%)	*Spots* corresp. presentes em todos os géis da mesma variedade (%)	*Spots* corresp. presentes em 3 géis (%)	*Spots* corresp. presentes em 2 géis (%)	*Spots* corresp. presentes em 2 géis (%)	t-teste Areia/Argila*	
						Spots relacionados	*Spots* diferentes
Biogold	551 (100)	241 (44)	88 (16)	199 (36)	23 (4)	9	232
Fontane	579 (100)	263 (45)	73 (13)	229 (40)	14 (2)	10	253
Innovator	563 (100)	223 (40)	70 (12)	226 (40)	44 (8)	13	210
Lady Rosetta	547 (100)	264 (48)	74 (14)	173 (32)	36 (7)	11	253
Maris Piper	709 (100)	347 (49)	113 (16)	206 (29)	43 (6)	14	333

[1] presente em pelo menos uma das replicatas.
[2] t-teste, $p < 0.05$

Fonte: autor.

Figura 3.3: Porcentagem de *spots* correlacionados quando comparados os 20 géis das corridas utilizando tiras de IPG pH 4-7 e gel 12,5% SDS-PAGE para (A) análise I, géis de 13 cm e (B) análise II, géis de 24 cm. Colunas indicam a quantidade de *spots* correspondentes em quatro (azul), três (vermelho), dois (verde) e um (roxo) géis para cada variedade. Variedades: B - Biogold, F - Fontane, I - Innovator, L - Lady Rosetta, M - Maris Piper.



Fonte: autor.

Na figura 3.3B verifica-se que, para géis maiores (24 cm), mais *spots* correspondentes estão simultaneamente presentes em dois géis do que em três ou um só gel da mesma variedade. Isso não necessariamente significa que o solo de cultivo, que é comum para cada dois géis, influencia na correspondência. Para confirmar a influência do tipo de solo, o teste t foi aplicado entre os géis da mesma variedade para detectar diferença significativa de acordo com o solo de cultivo (tabelas 3.3 e 3.4). Observa-se que somente entre 4 e 21 *spots* são significativamente relacionados na análise I; para análise II, o número de *spots* relacionados variou entre 9 e 14. Isso pode significar que para cada variedade, mais de 90% das proteínas correspondentes são diferentes quando comparadas aquelas detectadas em géis de plantas cultivadas em solo arenoso com as de solo argiloso. Assim, não existe uma discriminação clara entre as proteínas de acordo com solo em relação ao perfil proteômico das amostras testadas. Portanto, os quatro géis de cada variedade podem assim ser considerados replicatas, uma vez que seus dados comportam-se similarmente.

Análise de componentes principais (PCA) foi realizada com conjunto de dados contendo *spots* correspondentes de todos os géis analisados. Primeiramente, as variáveis com valores omissos não foram incluídas, assim, somente *spots* comuns a todos os géis foram considerados. Nesta abordagem, análise I apresentou 29 *spots* em comum e na análise II, 43 *spots* correspondentes foram detectados. Estes dados foram submetidos à transformação logarítmica de base 2 e submetidos à PCA (figura 3.4).

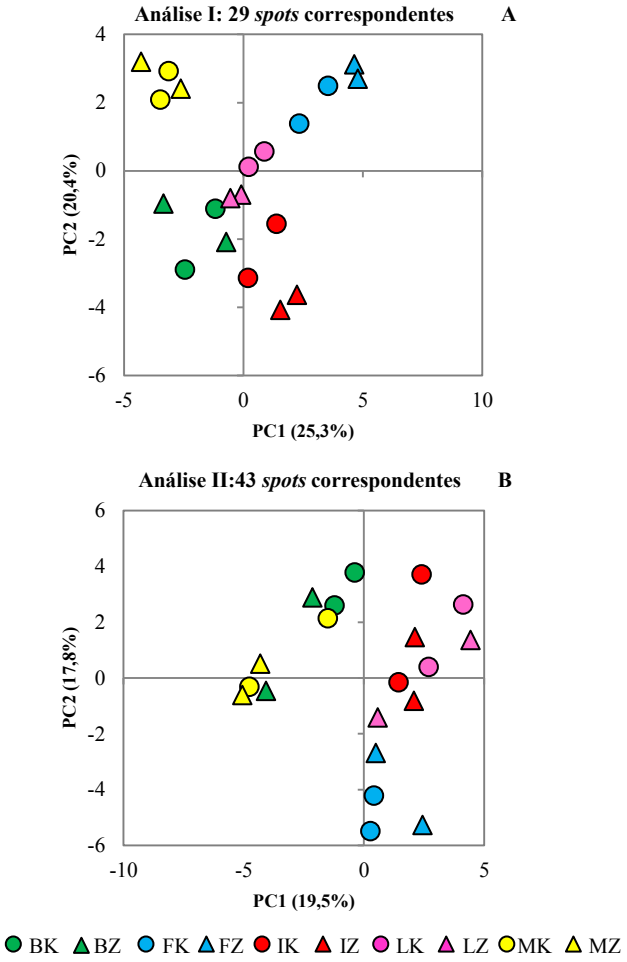
Não foi observada separação de acordo com solo em nenhum dos gráficos, em concordância com o teste t. No gráfico de PCA da análise I foi possível observar clara separação de acordo com a variedade (figura 3.4A). O primeiro componente (PC1) representou 25,3% da variação total do conjunto de dados e o segundo componente correspondeu a 20.4%. A PCA realizada com os dados da análise II também apresentou separação por variedades, apesar de uma das amostras Biogold crescida em solo arenoso estar localizada distante das outras da mesma variedade (figura 3.4B). O mesmo comportamento foi observado com a variedade Maris Piper cultivada em solo argiloso. Os primeiros dois componentes, PC1 e PC2, explicam 19,5% e 17,8% do total de variação no conjunto de dados, respectivamente. Mais de 90% da variação dos dados é explicada pelos primeiros nove componentes na análise I e por 10 componentes na análise II.

Na segunda abordagem, todos os dados omissos foram substituídos pelo valor mínimo considerado como ponto de corte para %Vol (0,005) e os gráficos correspondentes são mostrados nas figuras 3.5A e 3.5B. Este procedimento resultou em um conjunto de dados muito mais representativo,

composto por 740 *spots* correspondentes na análise I e 1756 na análise II. Ambos os gráficos de PCA desta abordagem apresentaram nítida separação de acordo com variedade.

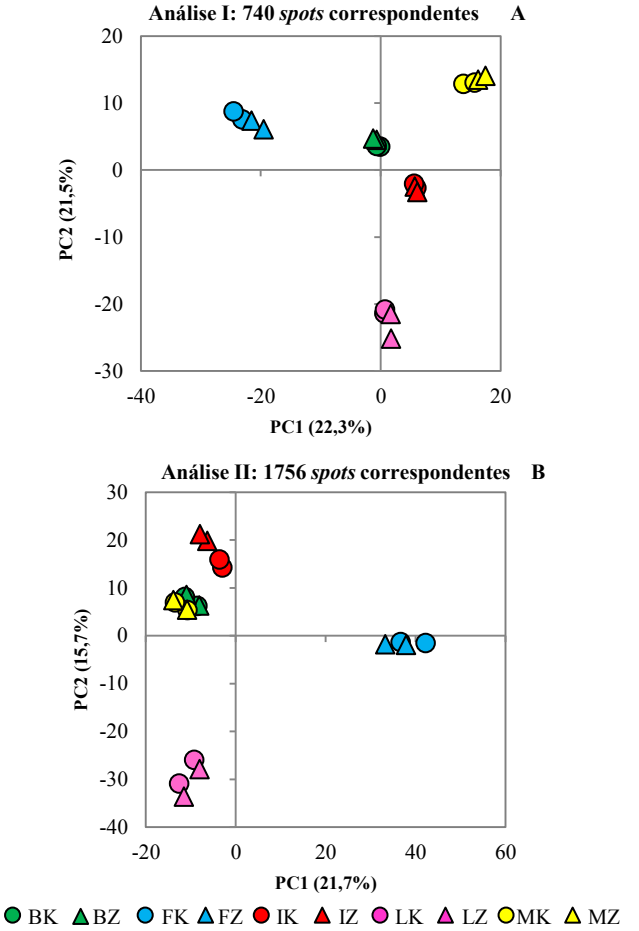
O primeiro componente principal, que representou 22,3% e 21,7% do total de variação para análise I e II respectivamente, claramente separou as amostras Fontane das outras quatro variedades. Lady Rosetta também apresentou nítida separação das outras variedades no segundo componente principal, que representou 21,5% da variação total na análise I e 15,7% na análise II. Aproximadamente 91% da variação total dos dados são explicados por nove componentes na análise I e 10 na análise II.

Figura 3.4: Gráficos de PCA dos dados de intensidade de *spots* de proteínas (%Vol) sem valores omissos obtidos de cinco variedades de batatas cultivadas em dois solos (marcador circular, arenoso; marcador triangular, argiloso). Proteínas foram separadas por 2-DE SDS-PAGE usando tiras de IPG pH 4-7 de comprimento (A) 13 cm ou (B) 24 cm. Variedades: B - Biogold, F - Fontane, I - Innovator, L - Lady Rosetta, M - Maris Piper; solo de cultivo: K - argiloso, Z - arenoso.



Fonte: autor.

Figura 3.5: Gráficos de PCA dos dados de intensidade de *spots* de proteínas (%Vol) de cinco variedades de batatas cultivadas em dois solos (marcador circular, arenoso; marcador triangular, argiloso). Todos os valores de %Vol omissos do conjunto de dados foram substituídos pelo valor mínimo (0,005) considerado na detecção dos *spots*. Proteínas foram separadas por 2-DE SDS-PAGE usando tiras de IPG pH 4-7 de comprimento (A) 13 cm ou (B) 24 cm. Variedades: B - Biogold, F - Fontane, I - Innovator, L - Lady Rosetta, M - Maris Piper; solo de cultivo: K - argiloso, Z – arenoso.



Fonte: autor.

3.4 DISCUSSÃO

O objetivo do presente estudo é ampliar o entendimento sobre como os perfis proteômicos de batatas comportam-se quando analisados por ferramenta estatística multivariada como PCA. Duas condições de isoeletrofocalização (tiras de IPG de 13 e 24 cm) foram testadas utilizando o mesmo material de cinco variedades de batatas. Adicionalmente, dois conjuntos de dados foram comparados, um sem valores omissos e outro com valores omissos substituídos.

A escolha do comprimento das tiras de IPG, assim como a faixa de pH para separação de proteínas, depende do propósito da investigação. O perfil proteômico de batatas pode ser acessado em tiras pequenas (como as de 7, 11 ou 13 cm) quando as proteínas de interesse são abundantes ou pré-fracionadas (por exemplo, quando o objetivo é estudar um grupo específico de proteínas como as isoformas de patatina) (BARTA et al., 2012). Tiras maiores (17, 18 ou 24 cm) são recomendadas quando o objetivo é a obtenção de mapas com maior resolução, maior número de *spots* e uma maior variedade de proteínas (CHAVES et al., 2009). Além disso, o uso dessas tiras permite maior facilidade para selecionar e identificar as proteínas nos géis. Corridas realizadas em tiras maiores admitem que maior quantidade de proteína total seja aplicada nos géis (até 1300 µg de proteína, comparado com um máximo de 450 µg para tiras menores, em pH 4-7). Neste trabalho, dois tamanhos de tiras de IPG foram usados, 13 e 24 cm. Em média, os géis maiores (24 cm) apresentaram quase o dobro do número de *spots* que os géis menores (13 cm) (tabela 3.5), o que provavelmente deve-se à maior área de gel disponível em géis de 24 cm . Além disso, maior quantidade de proteína foi colocada nas tiras de 24 cm: 500 µg, enquanto 350 µg foram colocadas nas tiras menores. Lee e co-autores (2012) isolaram proteína de raízes de batatas doces aplicando 400 µg em géis de 17 cm (pI 4–7) e 370 *spots* foram reprodutíveis. No entanto, maior quantidade de *spots* foi detectada quando géis maiores (24 cm) foram corados com fluoróforo, mesmo utilizando menor quantidade de proteína. Na investigação do ciclo de vida de tubérculos de batata, 150 µg de proteína total foi aplicada em tiras de IPG de 24 cm. Entre 900 e 1500 *spots* foram detectados (LEHESRANTA, SATU J. et al., 2006). Em outro estudo, dois métodos de extração de proteína foram comparados aplicando 450 µg de proteína em tiras de 24 cm. Um total de 1572 e 1482 *spots* foram detectados em cada método (DELAPLACE et al., 2006). Somente alguns pesquisadores investigam o proteoma de batata utilizando uma faixa maior de pH como 3-10. A menos que o extrato proteico de batatas seja conhecidamente básico em seu ponto isoelétrico (pI maior que 7), a faixa de

pH 4-7 é geralmente preferida. Esta garante uma boa resolução horizontal uma vez que a maioria das proteínas extraídas de batatas é neutra ou ácida (DELAPLACE et al., 2006; KOISTINEN et al., 2002; LEHESRANTA, S. J. et al., 2005).

A 2-DE é largamente usada para avaliar a quantidade de proteínas em diversos organismos pela comparação de diferentes grupos e suas replicatas. Após a separação e coloração do gel, é feita a detecção e quantificação dos *spots* nos géis e a comparação entre todos os géis. Este procedimento gera vários valores omissos para intensidades dos *spots* dentre as replicatas, e estes podem estar relacionados aos procedimentos de coloração (GROVE et al., 2006). A ocorrência de valores omissos pode dever-se a inúmeras razões, incluindo: intensidade de *spot* abaixo do limiar detectável ou limite de detecção (LOD); erros na comparação entre *spots* causados por distorções nos géis; ausência dos *spots* como resultado de erros experimentais durante a transferência da primeira para a segunda dimensão; abundância muito baixa ou real ausência do *spot* nas amostras (CHICH et al., 2007; PEDRESCHI et al., 2008). O tratamento de valores omissos em dados proteômicos é altamente recomendado por muitos pesquisadores (ALBRECHT et al., 2010; PEDRESCHI et al., 2008; VALLEDOR; JORRIN, 2011; ZELLNER et al., 2012). O principal motivo é essencialmente técnico: a maior parte dos métodos de análise multivariada, sejam ferramentas de classificação supervisionadas ou não-supervisionadas, não podem lidar com dados omissos (GROVE et al., 2006; KROGH et al., 2007).

Para contornar esta situação, algumas opções para lidar com valores omissos já foram discutidas (ALBRECHT et al., 2010). Substituir esses valores é uma alternativa para evitar perda de informação, contanto que não sejam substituídos por zero (ALBRECHT et al., 2010; GROVE et al., 2006). Meleth e colaboradores (2005) testaram três métodos de substituição utilizando valores mínimos. Primeiro substituíram os valores omissos com o valor mais baixo das intensidades log transformadas; depois selecionaram um conjunto de dados aleatórios contendo os 15 valores mais baixos e, finalmente, os valores omissos foram substituídos por um valor aleatoriamente selecionado a partir dos 30 valores mais baixos de cada gel. O estudo concluiu que os diferentes métodos de atribuição de valores não causam diferença na detecção de diferentes níveis de proteínas. Alguns autores provaram que o algoritmo baseado na análise de componentes principais Bayesiana (BPCA) é um método consistente para estimar os valores omissos em dados de proteômica em gel (PEDRESCHI et al., 2008). Além dessas alternativas, a atribuição de valores pode ser também

feita utilizando o valor mínimo global de todos os *spots* em todos os géis do experimento (ALMEIDA et al., 2005).

No presente trabalho, PCA foi realizada com dois diferentes conjuntos de dados: o primeiro contendo somente *spots* presentes simultaneamente em todos os 20 géis e o segundo com dados de proteínas presentes em pelo menos dois géis quando todos os géis são comparados, situação que requer substituição dos valores omissos. Em ambos os casos, os dados foram pretratados por transformação logarítmica de base 2, procedimento que não altera a distribuição dos dados, mas garante sua capacidade de ser utilizados para análise estatística. Substituição pelo valor mínimo foi escolhida como alternativa para lidar com os valores faltantes. Os gráficos de PCA a partir dos conjuntos de dados contendo somente *spots* presentes em todos os géis (figura 3.4) mostram sutil separação de acordo com as variedades. Comparando-se PCAs dos conjuntos de dados maiores, que tiveram os valores omissos substituídos (figura 3.5), as variedades apresentam-se claramente separadas. A nítida separação entre variedades, como observada após a substituição dos valores omissos, também foi detectada em batatas submetidas a estresse abiótico (FOLGADO et al., 2013). Nesse trabalho foi observado que as principais diferenças nos gráficos de PCA foram encontradas entre as espécies testadas, independentemente do tratamento.

Os primeiros conjuntos de dados, tanto para análise I quanto para análise II, continham cerca de 3% do número total de *spots* correlacionados detectados em cada análise (tabela 3.5).

Tabela 3.5: Resumo dos números de *spots* detectados em cada análise, com e sem substituição de valores. Dados obtidos pela isoeletrofocalização de proteína total de cinco variedades de batatas realizada em tiras de IPG de 13 cm (análise I) e 24 cm (análise II) pH 4-7 seguida eletroforese em géis 12,5% SDS-PAGE, corados com Coomassie Brilliant Blue G-250.

	Análise I (13 cm)	**Análise II (24 cm)**
Número médio de *spots* detectados (média[1] ± DP)	267 ± 28	497 ± 84
Número de *spots* correspondentes em comum a todos os 20 géis	29	43
Número total de *spots* correspondentes presentes em pelo menos 2 géis [2]	740	1756

DP, desvio padrão.
[1] média dos 20 géis de cada análise.
[2] valores omissos foram substituídos pelo mínimo detectado (0,005).
Fonte: autor.

Esta pequena proporção em relação ao total de *spots* correspondentes representa um conjunto de dados relativamente pequeno. Por outro lado, deve-se levar em consideração que mais de 50% da informação do conjunto de dados maior consiste de valores substituídos e, apesar de a distribuição ser normalizada, pode haver uma tendência de agrupamento das variedades porque a maior parte dos valores omissos são os mesmos dentro da mesma variedade.

A ferramenta PCA reduz a complexidade dos dados e fornece uma melhor visualização das possíveis conexões entre as variáveis. A ideia básica é reduzir a dimensionalidade e revelar estruturas ocultas em um conjunto de dados para que estas estruturas possam ser descritas (GOTTLIEB et al., 2004). Este é o princípio de inúmeras ferramentas de classificação como a Modelagem Independente Flexível por Analogia de Classe (SIMCA). No capítulo anterior, foi mostrado que a modelagem SIMCA uma-classe foi eficaz na classificação de batatas como pertencendo ou não a um conjunto conhecido de batatas por dados de transcriptômica, obtidos a partir de análise de microarranjo. Esta classificação revela uma descoberta interessante, que pode levar à sua aplicação na avaliação de segurança de alimentos. Assim, a identificação de perigos seria aperfeiçoada com base na similaridade de novas variedades com uma base

de referência (classe) construída a partir de variedades conhecidamente seguras.

Os dados apresentados aqui mostram que PCA aplica-se para análise proteômica de batatas e por conta disso sugere-se que sua utilização em sistemas de classificação podem ser considerados. Todavia, aspectos essenciais como a substituição de valores omissos seguramente devem ser tratados antecipadamente. Substituição pelo menor valor de intensidade de *spot* permite que se lide com conjuntos de dados mais extensos, enriquecendo a quantidade de dados considerada na análise. Mostrou-se que as amostras são claramente discriminadas de acordo com suas variedades, o que pode ser desejável no caso de se ter uma grande quantidade de diferentes plantas. O tamanho de tiras de IPG não apresentou diferença na classificação por PCA quando comparados os tamanhos de 13 cm e 24 cm. Ambos apresentaram número de *spots* proporcionalmente similar ao total de *spots* correlacionados em cada análise. Consequentemente, talvez fosse preferível trabalhar com os géis menores por conveniência técnica. No entanto, para se continuar o trabalho com os sistemas de classificação como SIMCA, seria definitivamente necessário ter-se maior quantidade de amostras analisadas. Outros métodos de coloração de gel (como a coloração por prata) possivelmente gerariam um maior conjunto de dados composto pelos *spots* correspondentes, aumentando a fidelidade do comportamento das amostras. A oportunidade de um maior número de amostras para a análise permitiria um aprimoramento do estudo e até mesmo o desenvolvimento de um classificador SIMCA uma classe, de preferência com a possibilidade de um conjunto independente de amostras de teste.

É notável que a análise proteômica por 2-DE seja uma ferramenta bastante completa e elaborada capaz de representar o perfil de proteínas presentes em alimentos. Essa metodologia tem se mostrado adequada para complementação da avaliação de segurança de novos alimentos, no entanto muitos aspectos influenciam seu desempenho. Eles devem ser tratados previamente para obtenção de melhores conclusões. O presente estudo mostrou que escolhendo o comprimento adequado da tira de IPG e a faixa de pH, assim como tratamento dos valores omissos são passos importantes a serem tomados, que levarão a um melhor tratamento dos dados e maior compreensão dos resultados.

CONSIDERAÇÕES FINAIS

A extensa caracterização de novas variedades de plantas é uma prática padrão e pode auxiliar na garantia de que nem agricultores nem consumidores sejam indevidamente impactados. No entanto, diversas lacunas precisam ser preenchidas, principalmente porque não se tem conhecimento preciso de todos os possíveis perigos associados aos novos alimentos. E justamente por isso, efeitos desconhecidos podem deixar de ser detectados.

Atualmente a avaliação de alimentos é baseada no histórico de consumo seguro de cultivares similares tradicionais, e por isso geralmente não precisam ser submetidos a testes sistemáticos toxicológicos e nutricionais. No entanto, existem exceções em que alguns compostos apresentaram efeitos tóxicos agudos em humanos, que devem então ser avaliados caso a caso. Na Europa, a EFSA fornece orientações sobre estudos específicos, incluindo estudos em animais, necessários para avaliar níveis de toxicidade e alergenicidade de novas proteínas ou metabólitos. No caso de a nova planta, incluindo variedades GM, apresentar potenciais efeitos não intencionais baseado em análises molecular, composicional, fenotípica ou agronômica, não somente os constituintes potencialmente tóxicos da planta, mas sim a planta inteira deve ser testada. Nestes casos, os testes devem incluir pelo menos um estudo de toxicidade de 90 dias em roedores, enquanto que outros estudos comparativos de crescimento podem ser conduzidos com categorias de alimentos adequadas para produção de animais. A avaliação de estudos desse tipo com dieta em animais mostrou que dificuldades relacionadas à preparação da dieta, que ainda não é padronizada, podem levar a diferenças no desempenho do experimento, na análise dos dados, e no processamento e interpretação dos resultados. Além disso, testes de alimentação em animais são difíceis de serem realizados, tem baixo poder de detecção de feitos adversos, podem ter os resultados influenciados pela complexidade da matriz analítica e contribuem muito pouco para a avaliação de alimentos inteiros. Em princípio, estes tipos de teste deveriam ser evitados e serem mais analíticos, com avaliações precisa dos aspectos molecular e toxicológico. Ferramentas baseadas em tecnologias ômicas podem contribuir para este fim, pois possibilitam a exploração de processos biológicos complexos de maneira integrada por meio da abordagem focada na biologia de sistemas. Esses métodos conferem um enfoque global às análises, e ao mesmo tempo facilitam a identificação de genes responsáveis pela sobrevivência e persistência em

ambientes específicos, relevantes para o controle da segurança dos alimentos.

Levando isso em consideração, neste trabalho propôs-se a exposição de uma prova de princípio para uma nova abordagem ampla, não-alvo, como alternativa para complementar a atual análise comparativa de alimentos. Através das análises amplas de perfil transcriptômico e proteômico, mostrou-se que é possível obter conjuntos de dados adequados para análise multivariada capaz de detectar diferenças relacionadas às condições de cultivo e variedades.

A técnica de microarranjo foi utilizada para obtenção de dados de transcriptômica para a construção de classificadores SIMCA. Foi obtido um conjunto de classificadores capazes de corretamente qualificar amostras com maior variabilidade técnica em relação ao conjunto de amostras de treinamento. A eletroforese bidimensional foi aplicada nas mesmas amostras e através da análise de componentes principais com os dados de proteômica verificou-se clara separação entre variedades. O êxito destes resultados permite a formação de uma base inicial satisfatória para exploração de análises ômicas como complementação da atual avaliação comparativa de novos alimentos.

Mesmo que as tecnologias ômicas estejam se tornando métodos cada vez mais comuns na pesquisa, oferecendo diferentes oportunidades, ainda existem muitos desafios a serem superados, assim como em qualquer metodologia complexa. É necessária a padronização das técnicas de extração de RNA e proteína; estabelecimento de uma associação entre os impactos ambientais (plantio, solo, localização, estresse) com as potenciais alterações nos perfis genômico, proteômico, etc. Além disso, é preciso lidar adequadamente com a grande quantidade de dados brutos complexos gerados, de maneira que se possa comparar os resultados adequadamente, como por exemplo, utilização e validação de métodos de estatística multivariada para lidar com a presença de grande número de variáveis (que são medidas analíticas independentes) contra pequeno número de replicatas. Uma vez que estes desafios estejam contornados, a avaliação de segurança de novos alimentos poderá ser mais eficaz e abrangente, garantindo a comercialização e consumo de alimentos saudáveis e seguros para a saúde da população.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, G. K.; RAKWAL, R. Rice proteomics: A cornerstone for cereal food crop proteomes. **Mass Spectrometry Reviews,** v. 25, n. 1, p. 1-53, 2006.

ALBRECHT, D. et al. Missing values in gel-based proteomics. **Proteomics,** v. 10, n. 6, p. 1202-1211, 2010.

ALMEIDA, J. S. et al. Normalization and analysis of residual variation in two-dimensional gel electrophoresis for quantitative differential proteomics. **Proteomics,** v. 5, n. 5, p. 1242-1249, 2005.

ANDERSON, N. L.; ANDERSON, N. G. Proteome and proteomics: new technologies, new concepts, and new words. **Electrophoresis,** v. 19, n. 11, p. 1853-1861, 1998.

ANTTONEN, M. J. et al. Genetic and Environmental Influence on Maize Kernel Proteome. **Journal of Proteome Research,** v. 9, n. 12, p. 6160-6168, 2010.

BALSAMO, G. et al. Proteomic Analysis of Four Brazilian MON810 Maize Varieties and Their Four Non-Genetically-Modified Isogenic Varieties. **Journal of Agricultural and Food Chemistry,** v. 59, n. 21, p. 11553-11559, 2011.

BARROS, E. et al. Comparison of two GM maize varieties with a near-isogenic non-GM variety using transcriptomics, proteomics and metabolomics. **Plant Biotechnology Journal,** v. 8, n. 4, p. 436-451, 2010.

BARTA, J. et al. Cultivar Variability of Patatin Biochemical Characteristics: Table versus Processing Potatoes (Solanum tuberosum L.). **Journal of Agricultural and Food Chemistry,** v. 60, n. 17, p. 4369-4378, 2012.

BATISTA, R. et al. Microarray analyses reveal that plant mutagenesis may induce more transcriptomic changes than transgene insertion. **Proceedings of the National Academy of Sciences,** v. 105, n. 9, p. 3640-3645, 2008.

BAUDO, M. M. et al. Transgenesis has less impact on the transcriptome of wheat grain than conventional breeding. **Plant Biotechnology Journal,** v. 4, n. 4, p. 369-380, 2006.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. **Journal of the Royal Statistical Society Series B-Methodological,** v. 57, n. 1, p. 289-300, 1995.

BERRUETA, L.; ALONSO-SALCES, R.; HEBERGER, K. Supervised pattern recognition in food analysis. **Journal of Chromatography a,** v. 1158, n. 1-2, p. 196-214, 2007.

BERRY, M. J.; LINOFF, G. S. **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**. Wiley, 2004. ISBN 9780471470649.

BINDSCHEDLER, L. V.; CRAMER, R. Quantitative plant proteomics. **Proteomics,** v. 11, n. 4, p. 756-775, 2011.

BRANDAO, A. R.; BARBOSA, H. S.; ARRUDA, M. A. Z. Image analysis of two-dimensional gel electrophoresis for comparative proteomics of transgenic and non-transgenic soybean seeds. **Journal of Proteomics,** v. 73, n. 8, p. 1433-1440, 2010.

BRAZMA, A. et al. One-stop shop for microarray data. **Nature,** v. 403, n. 6771, p. 699-700, 2000.

BRERETON, R. G. **Chemometrics: Data Analysis for the Laboratory and Chemical Plant**. John Wiley & Sons. 489 ISBN 978-0-471-48978-8. 2003.

CELLINI, F. et al. Unintended effects and their detection in genetically modified crops. **Food and Chemical Toxicology,** v. 42, n. 7, p. 1089-1125, 2004.

CHAVES, I. et al. Proteomic evaluation of wound-healing processes in potato (*Solanum tuberosum* L.) tuber tissue. **Proteomics,** v. 9, n. 17, p. 4154-4175, 2009.

CHENG, K. C. et al. Effect of transgenes on global gene expression in soybean is within the natural range of variation of conventional cultivars. **Journal of Agricultural and Food Chemistry,** v. 56, n. 9, p. 3057-3067, 2008.

CHICH, J.-F. et al. Statistics for proteomics: Experimental design and 2-DE differential analysis. **Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences,** v. 849, n. 1-2, p. 261-272, 2007.

COCKBURN, A. Assuring the safety of genetically modified (GM) foods: the importance of an holistic, integrative approach. **Journal of Biotechnology,** v. 98, n. 1, p. 79-106, 2002.

COLL, A. et al. Natural variation explains most transcriptomic changes among maize plants of MON810 and comparable non-GM varieties subjected to two N-fertilization farming practices. **Plant Molecular Biology,** v. 73, n. 3, p. 349-362, 2010.

COLL, A. et al. Proteomic analysis of MON810 and comparable non-GM maize varieties grown in agricultural fields. **Transgenic Research,** v. 20, n. 4, p. 939-949, 2011.

CONSTABLE, A. et al. History of safe use as applied to the safety assessment of novel foods and foods derived from genetically modified organisms. **Food and Chemical Toxicology,** v. 45, n. 12, p. 2513-2525, 2007.

D'ALESSANDRO, A.; ZOLLA, L. We Are What We Eat: Food Safety and Proteomics. **Journal of Proteome Research,** v. 11, n. 1, p. 26-36, 2012.

DAVIES, H. A role for "omics" technologies in food safety assessment. **Food Control,** v. 21, n. 12, p. 1601-1610, 2010.

DE LA FUENTE, M. et al. 2-DE-based proteomic analysis of common bean (Phaseolus vulgaris L.) seeds. **Journal of Proteomics,** v. 74, n. 2, p. 262-267, 2011.

DE MAESSCHALCK, R. et al. Decision criteria for soft independent modelling of class analogy applied to near infrared data. **Chemometrics and Intelligent Laboratory Systems,** v. 47, n. 1, p. 65-77, 1999.

DEFRANCESCO, L. How safe does transgenic food need to be? **Nat Biotechnol,** v. 31, n. 9, p. 794-802, 2013.

DELAPLACE, P. et al. Potato tuber proteomics: Comparison of two complementary extraction methods designed for 2-DE of acidic proteins. **Proteomics,** v. 6, n. 24, p. 6494-6497, 2006.

DUGGAN, D. J. et al. Expression profiling using cDNA microarrays. **Nature Genetetics,** v. 21, n. 1 Suppl, p. 10-14, 1999.

DYBING, E. et al. Hazard characterisation of chemicals in food and diet: dose response, mechanisms and extrapolation issues. **Food and Chemical Toxicology,** v. 40, n. 2-3, p. 237-282, 2002.

EFSA. European Food Safety Authority. Panel on Genetically Modified Organisms (GMO). Guidance for risk assessment of food and feed from genetically modified plants. **EFSA Journal,** v. 9(5):2150, p. 1-37, 2011.

EL SANHOTY, R.; ABD EL-RAHMAN, A. A.; BÖGL, K. W. Quality and safety evaluation of genetically modified potatoes Spunta with Cry V gene: Compositional analysis, determination of some toxins, antinutrients compounds and feeding study in rats. **Food / Nahrung,** v. 48, n. 1, p. 13-18, 2004.

ENGKILDE, K.; JACOBSEN, S.; SØNDERGAARD, I. Multivariate data analysis of proteome data. **Methods in Molecular Biology,** v. 355, p. 195-210, 2007.

ESBENSEN, K.; SCHOENKOPF, S.; MIDTGAARD, T. **Multivariate analysis in practice**. Trondheim: CAMO, 361 p. ISBN 8299333008, 1994.

EUROPEAN-COMISSION. Commission Implementing Regulation (EU) No 503/2013 of 3 April 2013 on applications for authorisation of genetically modified food and feed in accordance with Regulation (EC) No 1829/2003 of the European Parliament and of the Council and amending Commission Regulations (EC) No 641/2004 and (EC) No 1981/2006. **Official Journal of the European Union**: 8.6.2013, No L 157/1 2013.

FAO/WHO – Food and Agriculture Organization of the United Nations/ World Health Organization. Biotechnology and Food Safety. Report of a Joint FAO/WHO Consultation. **FAO Food and Nutrition Paper 61**, 1996.

_____. **Codex Alimentarius Commission – Procedural Manual,** 20th ed., Rome: Joint FAO/WHO Food Standards Programme, 213 p., 2011.

_____. **FAO Statistical Yearbook 2013, Part 3 – Feeding the world.** Rome, 2013. Disponível em: <http://www.fao.org/docrep/018/i3107e/i3107e03.pdf>. Acesso em 11 Outubro de 2013.

FLATEN, G.; GRUNG, B.; KVALHEIM, O. A method for validation of reference sets in SIMCA modelling. **Chemometrics and Intelligent Laboratory Systems,** v. 72, n. 1, p. 101-109, 2004.

FOLGADO, R. et al. Differential Protein Expression in Response to Abiotic Stress in Two Potato Species: Solanum commersonii Dun and Solanum tuberosum L. **International Journal of Molecular Sciences,** v. 14, n. 3, p. 4912-4933, 2013.

GE, P. et al. Comparative proteomic analysis of grain development in two spring wheat varieties under drought stress. **Analytical and Bioanalytical Chemistry,** v. 402, n. 3, p. 1297-1313, 2012.

GORG, A.; WEISS, W.; DUNN, M. J. Current two-dimensional electrophoresis technology for proteomics. **Proteomics,** v. 4, n. 12, p. 3665-3685, 2004.

GOTTLIEB, D. M. et al. Multivariate approaches in plant science. **Phytochemistry,** v. 65, n. 11, p. 1531-1548, 2004.

GROVE, H. et al. Challenges related to analysis of protein spot volumes from two-dimensional gel electrophoresis as revealed by replicate gels. **Journal of Proteome Research,** v. 5, n. 12, p. 3399-3410, 2006.

GROVE, H. et al. Combination of Statistical Approaches for Analysis of 2-DE Data Gives Complementary Results. **Journal of Proteome Research,** v. 7, n. 12, p. 5119-5124, 2008.

HASHIMOTO, W., MOMMA, K., KATSUBE, T. et al. Safety assessment of genetically engineered potatoes with designed soybean glycinin: compositional analyses of the potato tubers and digestibility of the newly expressed protein in transgenic potatoes. **Journal of the Science of Food and Agriculture,** v. 79, p. 1607–1612, 1999.

HEALTH CANADA. **Amendment (Schedule No. 948) to Division 28 of the Food and Drug Regulations**, Sections B.28.001–003. 2003.

HERMAN, R. A.; PRICE, W. D. Unintended compositional changes in genetically modified (GM) crops: 20 years of research. **Journal of Agricultural and Food Chemistry,** v. 61, n. 48, p. 11695-11701, 2013.

HERRERO LATORRE, C. et al. Chemometric Classification of Potatoes with Protected Designation of Origin According to Their Producing Area and Variety. **Journal of Agricultural and Food Chemistry**, v. 61, n. 35, p. 8444-8451, 2013.

HUERTA-OCAMPO, J. et al. Proteomic analysis of differentially accumulated proteins during ripening and in response to 1-MCP in papaya fruit. **Journal of Proteomics,** v. 75, n. 7, p. 2160-2169, 2012.

JANES, K. A.; YAFFE, M. B. Data-driven modelling of signal-transduction networks. **Nature Reviews Molecular Cell Biology,** v. 7, n. 11, p. 820-828, 2006.

JIAO, Z. et al. Study on the compositional differences between transgenic and non-transgenic papaya (*Carica papaya* L.). **Journal of Food Composition and Analysis,** v. 23, n. 6, p. 640‑647, 2010.

KLETER, G.; KOK, E. Safety assessment of biotechnology used in animal production, including genetically modified (GM) feed and GM animals - a review. **Animal Science Papers and Reports,** v. 28, n. 2, p. 105-114, 2010.

KLOOSTERMAN, B. et al. Genes driving potato tuber initiation and growth: identification based on transcriptional changes using the POCI array. **Functional & Integrative Genomics,** v. 8, n. 4, p. 329-340, 2008.

KNUDSEN, I. et al. Risk management and risk assessment of novel plant foods: Concepts and principles. **Food and Chemical Toxicology,** v. 46, n. 5, p. 1681-1705, 2008.

KOISTINEN, K. M. et al. Birch PR-10c is induced by factors causing oxidative stress but appears not to confer tolerance to these agents. **New Phytologist,** v. 155, n. 3, p. 381-391, 2002.

KOK, E.; KUIPER, H. Comparative safety assessment for biotech crops. **Trends in Biotechnology,** v. 21, n. 10, p. 439-444, 2003.

KOK, E. J. et al. Comparative safety assessment of plant-derived foods. **Regulatory Toxicology and Pharmacology,** v. 50, n. 1, p. 98-113, 2008.

KOK, E. J. et al. Changes in Gene and Protein Expression during Tomato Ripening - Consequences for the Safety Assessment of New Crop Plant Varieties. **Food Science and Technology International,** v. 14, n. 6, p. 503-518, 2008.

KOSOROK, M. R.; MA, S. Marginal Asymptotics for the "Large P, Small N" Paradigm: With Applications to Microarray Data. **The Annals of Statistics,** v. 35, n. 4, p. 1456-1486, 2007.

KROGH, M. et al. A probabilistic treatment of the missing spot problem in 2D gel electrophoresis experimentse. **Journal of Proteome Research,** v. 6, n. 8, p. 3335-3343, 2007.

KUIPER, H. et al. Assessment of the food safety issues related to genetically modified foods. **Plant Journal,** v. 27, n. 6, p. 503-528, 2001.

KUIPER, H. A.; KOK, E. J.; DAVIES, H. V. New EU legislation for risk assessment of GM food: no scientific justification for mandatory animal feeding trials. **Plant Biotechnology Journal,** v. 11, n. 7, p. 781-784, 2013.

KUIPER, H. A.; KOK, E. J.; ENGEL, K. H. Exploitation of molecular profiling techniques for GM food safety assessment. **Current Opinion in Biotechnology,** v. 14, n. 2, p. 238-243, 2003.

LEE, J. J. et al. Comparative proteomic study between tuberous roots of light orange- and purple-fleshed sweetpotato cultivars. **Plant Science,** v. 193, p. 120-129, 2012.

LEHESRANTA, S. et al. Comparison of tuber proteomes of potato varieties, landraces, and genetically modified lines. **Plant Physiology,** v. 138, n. 3, p. 1690-1699, 2005.

LEHESRANTA, S. et al. Proteomic analysis of the potato tuber life cycle. **Proteomics,** v. 6, n. 22, p. 6042-6052, 2006.

LEHESRANTA, S. et al. Effects of agricultural production systems and their components on protein profiles of potato tubers. **Proteomics,** v. 7, n. 4, p. 597-604, 2007.

LÓPEZ, M. I. et al. Validation of multivariate screening methodology. Case study: Detection of food fraud. **Analytica Chimica Acta,** v. 827, p. 28-33, 2014.

MASSART, D. L. et al. **Handbook of Chemometrics and Qualimetrics: Part A**. Amsterdam: Elsevier, 1997.

MELETH, S.; DESHANE, J.; KIM, H. The case for well-conducted experiments to validate statistical protocols for 2D gels: different pre-processing = different lists of significant proteins. **Bmc Biotechnology,** v. 5, n. 7, 2005.

METZDORFF, S. B. et al. Evaluation of a non-targeted "Omic"' approach in the safety assessment of genetically modified plants. **Plant Biology,** v. 8, n. 5, p. 662-672, 2006.

MOMMA, K. et al. Quality and safety evaluation of genetically engineered rice with soybean glycinin: Analyses of the grain composition and digestibility of glycinin in transgenic rice. **Bioscience Biotechnology and Biochemistry,** v. 63, n. 2, p. 314-318, 1999.

MORALES, C. A. et al. Correlation of phenotype with the genotype of egg-contaminating Salmonella enterica serovar Enteritidis. **Applied Environmental Microbiology,** v. 71, n. 8, p. 4388-4399, 2005.

NELSON, D. L.; COX, M. M. **Lehninger principles of biochemistry**. 3rd ed. New York: Worth. XXIX, 1152 p ISBN 1572591536, 2000.

NOGUEIRA, S. B. et al. Proteomic analysis of papaya fruit ripening using 2DE-DIGE. **J Proteomics,** v. 75, n. 4, p. 1428-1439, 2012.

OBERTHUER, A. et al. Comparison of performance of one-color and two-color gene-expression analyses in predicting clinical endpoints of neuroblastoma patients. **Pharmacogenomics Journal,** v. 10, n. 4, p. 258-266, 2010.

OECD. Organisation for Economic Cooperation and Development. **Guidelines for the Testing of Chemicals,** v. 2012, 1993a.

_____. **Safety evaluation of foods derived by modern biotechnology: concepts and principles**. Paris, France: 1993b.

_____. Consensus Document on Compositional Considerations for New Varieties of Potatoes: Key Food and Feed Nutrients, Anti-nutrients and Toxicants. **Series on the Safety of Novel Foods and Feeds,** v. 4, p. 1-26, 2002.

_____. An introduction to the food/feed safety consensus documents of the Task Force. **Series on the Safety of Novel Foods and Feeds,** v. 14, 2006.

OLIVERI, P.; DOWNEY, G. Multivariate class modeling for the verification of food-authenticity claims. **Trends in Analytical Chemistry,** v. 35, p. 74-86, 2012.

PEDRESCHI, R. et al. Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. **Proteomics,** v. 8, n. 7, p. 1371-1383, 2008.

PIRRUNG, M. C. How to make a DNA chip. **Angewandte Chemie International Edition,** v. 41, n. 8, p. 1276-1289, 2002.

R CORE TEAM. R: A language and environment for statistical computing. **R Foundation for Statistical Computing** p. Vienna, Austria, 2013.

RENWICK, A. Risk characterisation of chemicals in food. **Toxicology Letters,** v. 149, n. 1-3, p. 163-176, 2004.

RESENDE, D. D. O. et al. **Guia para Comprovação da Segurança de Alimentos e Ingredientes**. Brasília, DF: ANVISA - Agência Nacional de Vigilância Sanitária, 2013. Disponível em: <http://portal.anvisa.gov.br/wps/wcm/connect/2b84a5004eb5354885fb878a610f4177/Guia+para+Comprovação+da+Segurança+de+Alimentos+e+Ingredientes.pdf?MOD=AJPERES>. Acessado em 12 Julho, 2013.

RICROCH, A. E. Assessment of GE food safety using '-omics' techniques and long-term animal feeding studies. **New Biotechnology,** v. 30, n. 4, p. 349-354, 2013.

ROY, S.; SEN, C. K. cDNA microarray screening in food safety. **Toxicology,** v. 221, n. 1, p. 128-133, 2006.

RUEBELT, M. C. et al. Application of two-dimensional gel electrophoresis to interrogate alterations in the proteome of genetically modified crops. 1. Assessing analytical validation. **Journal of Agricultural and Food Chemistry,** v. 54, n. 6, p. 2154-2161, 2006a.

_____. Application of two-dimensional gel electrophoresis to interrogate alterations in the proteome of genetically modified crops. 2. Assessing natural variability. **Journal of Agricultural and Food Chemistry,** v. 54, n. 6, p. 2162-2168, 2006b.

_____. Application of two-dimensional gel electrophoresis to interrogate alterations in the proteome of gentically modified crops. 3. Assessing unintended effects. **Journal of Agricultural and Food Chemistry,** v. 54, n. 6, p. 2169-2177, 2006c.

RUSTICI, G. et al. ArrayExpress update--trends in database growth and links to data analysis tools. **Nucleic acids research,** v. 41, n. Database issue, p. D987-990, 2013.

SCHOLZ, M. et al. Metabolite fingerprinting: detecting biological features by independent component analysis. **Bioinformatics,** v. 20, n. 15, p. 2447-2454, 2004.

SHAW, M. M.; RIEDERER, B. M. Sample preparation for two-dimensional gel electrophoresis. **Proteomics,** v. 3, n. 8, p. 1408-1417, 2003.

SIEBERT, K. Chemometrics in brewing - A review. **Journal of the American Society of Brewing Chemists,** v. 59, n. 4, p. 147-156, 2001.

SPIELBAUER, B.; STAHL, F. Impact of microarray technology in nutrition and food research. **Molecular Nutrition & Food Research,** v. 49, n. 10, p. 908-917, 2005.

STOREY, J. D.; TIBSHIRANI, R. Statistical significance for genomewide studies. **Proceedings of the National Academy of Sciences,** v. 100, n. 16, p. 9440-9445, 2003.

TAX, D. M. J. **One-class classification: Concept-learning in the absence of counter-examples**. Delft, the Netherlands: Technical University of Delft. Doctoral Dissertation, The Netherlands ISBN 90-75691-05-x. 2001.

TIAN, L. et al. Differential proteomic analysis of soluble extracellular proteins reveals the cysteine protease and cystatin involved in suspension-cultured cell proliferation in rice. **Biochimica Et Biophysica Acta-Proteins and Proteomics,** v. 1794, n. 3, p. 459-467, 2009.

VALLEDOR, L.; JORRIN, J. Back to the basics: Maximizing the information obtained by quantitative two dimensional gel electrophoresis analyses by an appropriate experimental design and statistical analyses. **Journal of Proteomics,** v. 74, n. 1, p. 1-18, 2011.

VAN DER VOET, H. et al. A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. **Bmc Biotechnology,** v. 11, n. 15, 2011.

VAN DIJK, J. P. et al. The Identification and Interpretation of Differences in the Transcriptomes of Organically and Conventionally Grown Potato Tubers. **Journal of Agricultural and Food Chemistry,** v. 60, n. 9, p. 2090-2101, 2012.

_____. Transcriptome Analysis of Potato Tubers-Effects of Different Agricultural Practices. **Journal of Agricultural and Food Chemistry,** v. 57, n. 4, p. 1612-1623, 2009.

_____. Gene expression profiling for food safety assessment: Examples in potato and maize. **Regulatory Toxicology and Pharmacology,** v. 58, n. 3, p. S21-S25, 2010.

VAN WIJK, K. J. Challenges and prospects of plant proteomics. **Plant Physiology,** v. 126, n. 2, p. 501-508, 2001.

VANDEGINSTE, B. G. M. et al. **Handbook of Chemometrics and Qualimetrics: Part B**. Amsterdam: Elsevier Science, 1998.

WESTERHUIS, J. A. et al. Assessment of PLSDA cross validation. **Metabolomics,** v. 4, n. 1, p. 81-89, 2008.

WHO. **World Health Organization. Global strategy for food safety: safer food for better health.** Geneva, Switzerland: WHO Library Cataloguing-in-Publication Data, 2002.

WOLD, S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. **Technometrics,** v. 20, n. 4, p. 397-405, 1978.

WOLD, S.; SJÖSTRÖM, M. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In: (Ed.). **Chemometrics: Theory and Application**: AMERICAN CHEMICAL SOCIETY, v.52, cap. 12, p. 243-282. (ACS Symposium Series). ISBN 0-8412-0379-2, 1977.

XU, Y.; BRERETON, R. Diagnostic pattern recognition on gene-expression profile data by using one-class classification. **Journal of Chemical Information and Modeling,** v. 45, n. 5, p. 1392-1401, 2005.

YE, X. et al. Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. **Science,** v. 287, n. 5451, p. 303-305, 2000.

ZELLNER, M. et al. How many spots with missing values can be tolerated in quantitative two-dimensional gel electrophoresis when applying univariate statistics? **Journal of Proteomics,** v. 75, n. 6, p. 1792-1802, 2012.

ZERZUCHA, P. et al. Non-parametric multivariate analysis of variance in the proteomic response of potato to drought stress. **Analytica Chimica Acta,** v. 719, p. 1-7, 2012.

ZOLLA, L. et al. Proteomics as a complementary tool for identifying unintended side effects occurring in transgenic maize seeds as a result of genetic modifications. **Journal of Proteome Research,** v. 7, n. 5, p. 1850-1861, 2008.

ZORB, C. et al. Levels of Compounds and Metabolites in Wheat Ears and Grains in Organic and Conventional Agriculture. **Journal of Agricultural and Food Chemistry,** v. 57, n. 20, p. 9555-9562, 2009.

APÊNDICE A – ANÁLISE DE COMPONENTES PRINCIPAIS DAS ANÁLISES A (CENÁRIO SIMILAR) E B (CENÁRIO DIFERENTE)

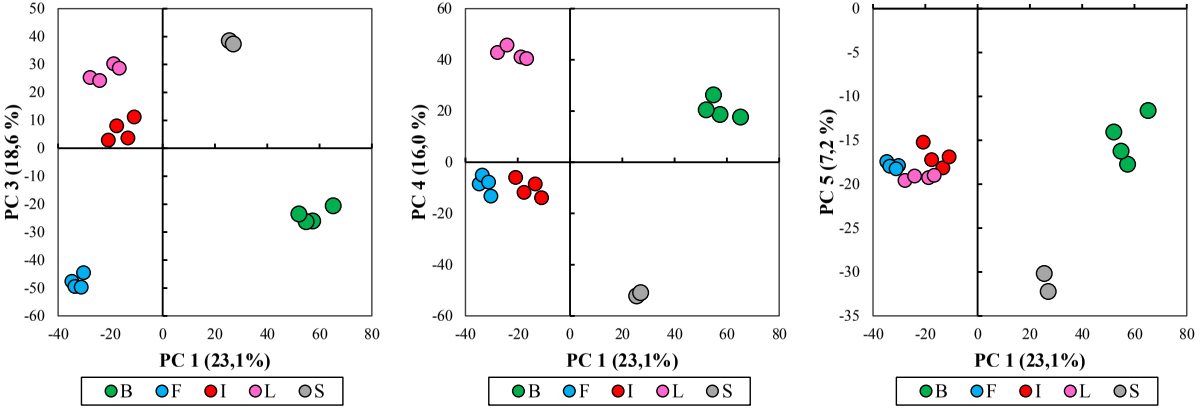
Tabela com a contribuição de cada componente principal no total de variação dos conjuntos de dados e gráficos da análise de componentes principais com combinações dos cinco primeiros componentes (exceto PC1xPC2), para as análises A e B.

PC: componente principal, B: Biogold, F: Fontane, I: Innovator, L: Lady Rosetta, M: Maris Piper, S: Sante S11.

Porcentagem cumulativa (%) de contribuição de cada componente principal (PC) para as análises A (cenário similar) e B (cenário diferente):

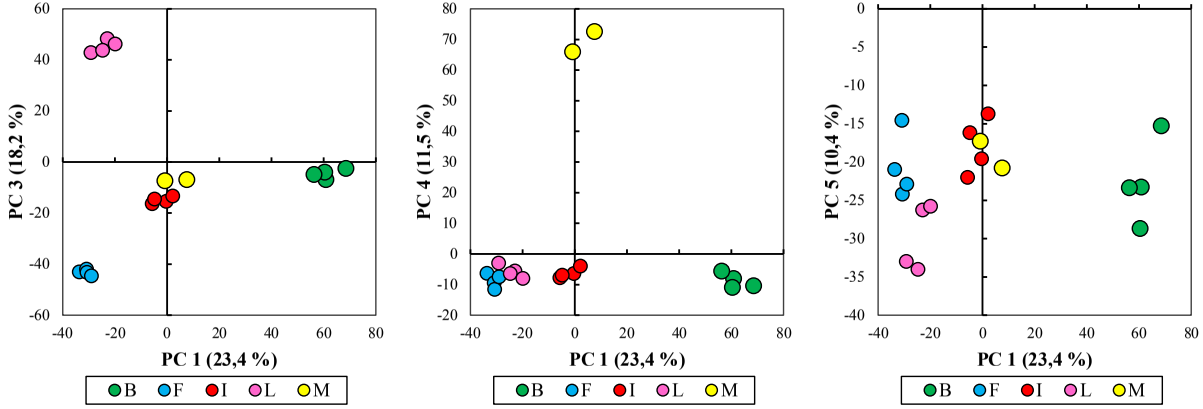
	Análise A	Análise B
PC1	23,41	23,13
PC2	44,22	44,09
PC3	62,44	62,70
PC4	73,92	78,68
PC5	84,29	85,84
PC6	87,54	88,84
PC7	90,05	91,20
PC8	92,03	93,00
PC9	93,76	94,18
PC10	94,90	95,29

Análise A: conjunto de teste similar



PC 2 (21,0%)
PC 3 (18,6 %)
PC 4 (16,0 %)
PC 5 (7,2 %)
PC 3 (18,6%)
PC 4 (16,0%)
B F I L S

Análise B: conjunto de teste diferente

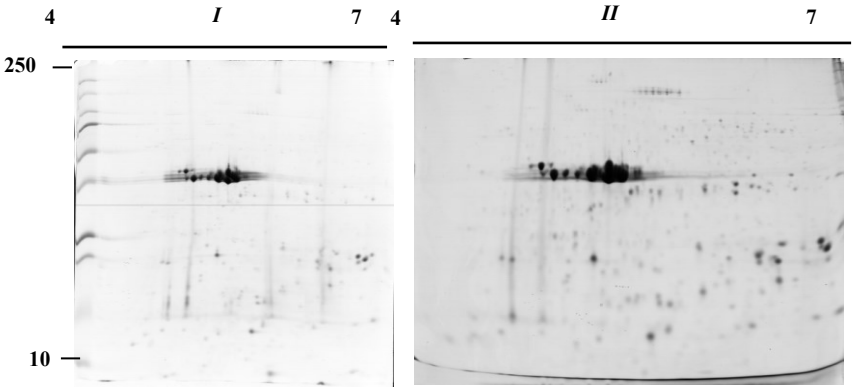


PC 2 (20,8 %)
PC 3 (18,2 %)
PC 4 (11,5 %)
PC 5 (10,4 %)
B
F
I
L
M

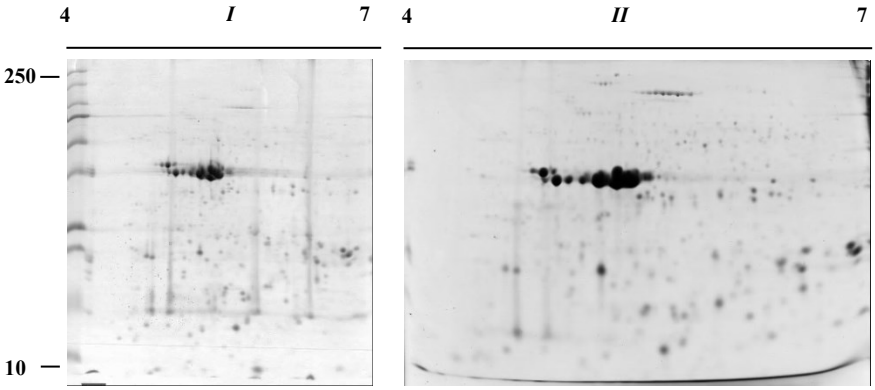
APÊNDICE B – MAPAS DE ELETROFORESE BIDIMENSIONAL

Imagens de géis de poliacrilamida de proteínas das variedades de batata Biogold, Fontane, Innovator, Lady Rosetta e Maris Piper, cultivadas em argila ou areia. Análise foi feita utilizando tiras de focalização linear de 13 cm (Análise *I*) ou de 24 cm (Análise *II*). Linhas horizontais indicam a faixa de pH e a marcação vertical indica o intervalo correspondente à massa molecular (kDa).

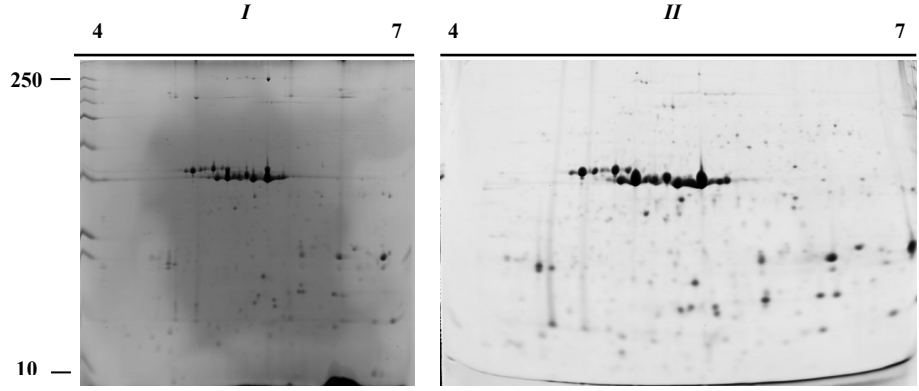
Variedade Biogold – Argila



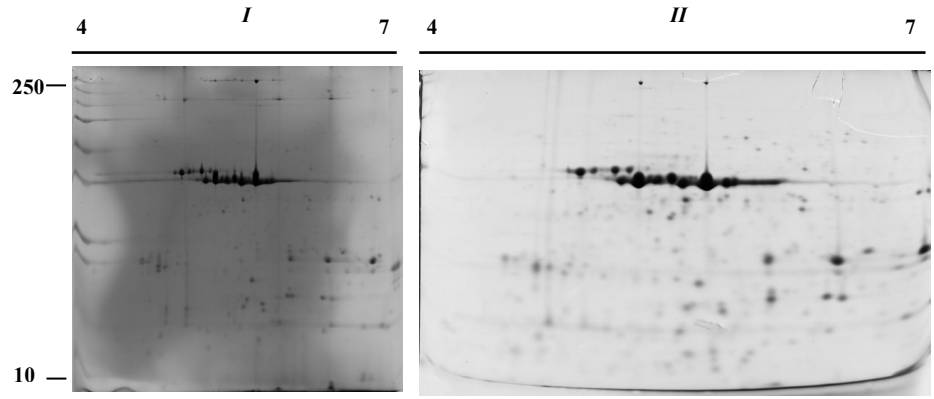
Varieda de Biogold – Areia



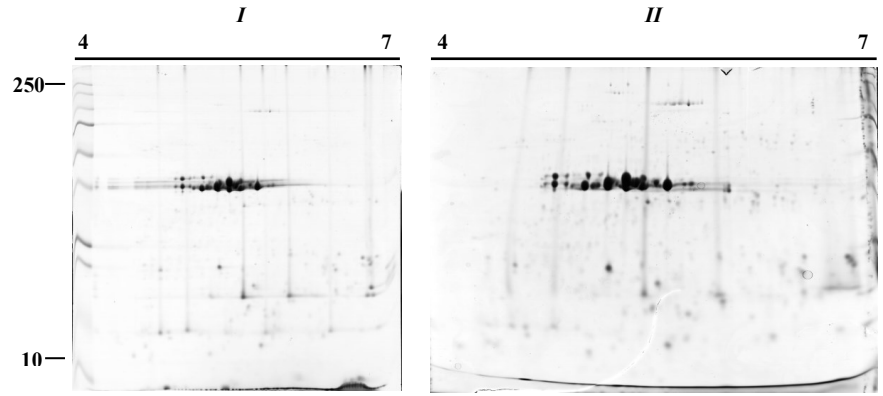
Variedade Fontane – Argila



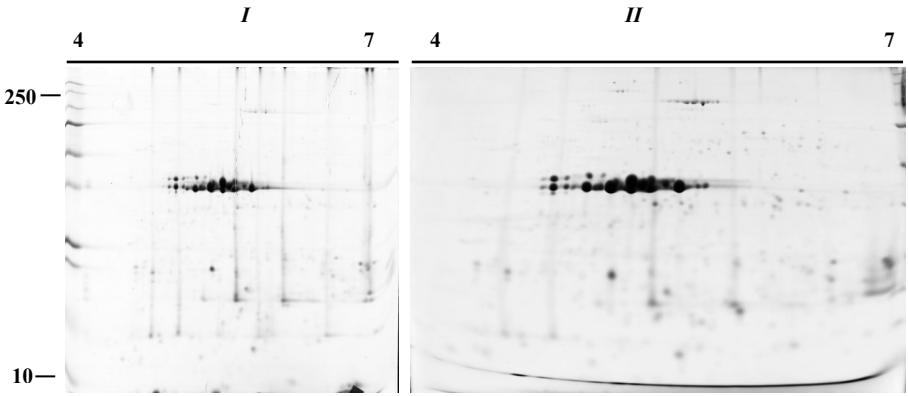
Variedade Fontane – Areia



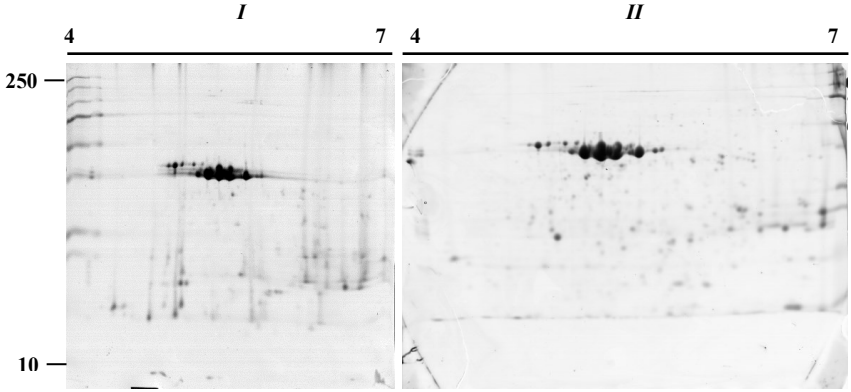
Variedade Innovator – Argila



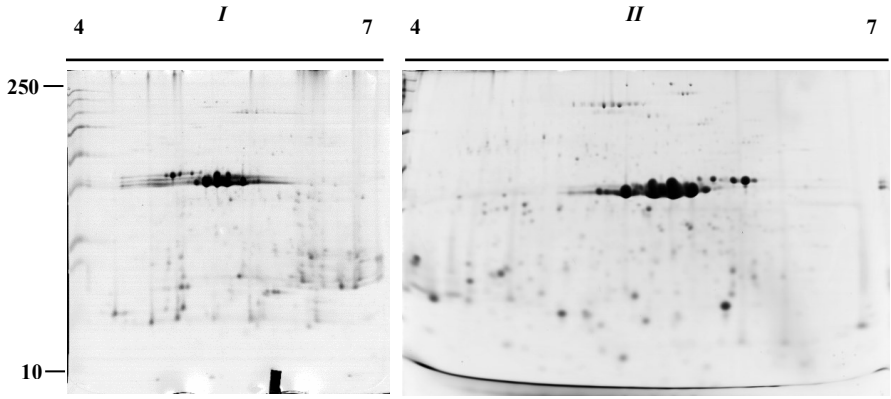
Variedade Innovator – Areia



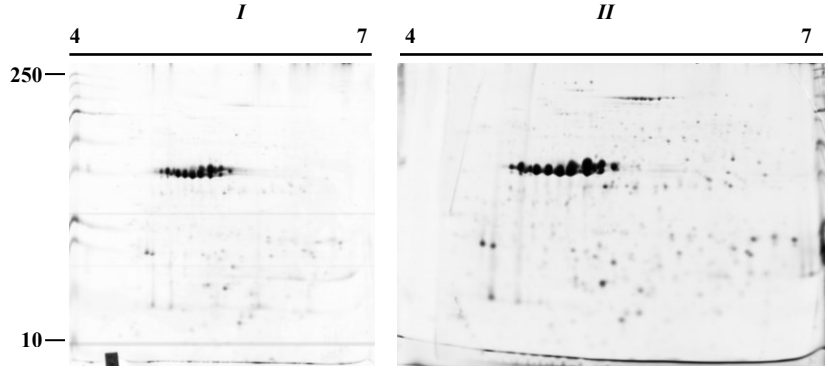
Variedade Lady Rosetta – Argila



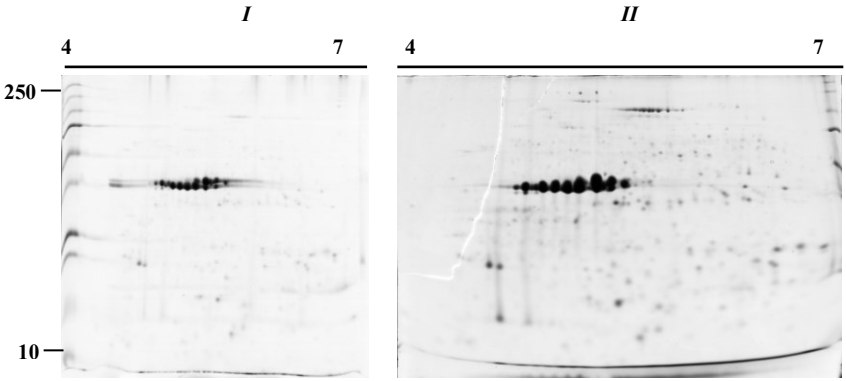
Variedade Lady Rosetta – Areia



Variedade Maris Piper – Argila



Variedade Maris Piper – Areia



APÊNDICE C – ARTIGO PUBLICADO

Regulatory Toxicology and Pharmacology 70 (2014) 297–303

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Safety assessment of plant varieties using transcriptomics profiling and a one-class classifier

Jeroen P. van Dijk [a,*,1], Carla Souza de Mello [a,b,1], Marleen M. Voorhuijzen [a], Ronald C.B. Hutten [c], Ana Carolina Maisonnave Arisi [b], Jeroen J. Jansen [d], Lutgarde M.C. Buydens [d], Hilko van der Voet [e], Esther J. Kok [b]

[a] RIKILT, Wageningen UR, Wageningen, The Netherlands
[b] Federal University of Santa Catarina, Brazil
[c] Plant Breeding, Wageningen UR, Wageningen, The Netherlands
[d] Institute for Molecules and Materials, Radboud University Nijmegen, The Netherlands
[e] Biometris, Wageningen UR, Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 2 April 2014
Available online 18 July 2014

Keywords:
Transcriptomics
Profiling
One-class classifiers
Food safety evaluation
GM plants
Chemometrics

ABSTRACT

An important part of the current hazard identification of novel plant varieties is comparative targeted analysis of the novel and reference varieties. Comparative analysis will become much more informative with unbiased analytical approaches, *e.g.* omics profiling. Data analysis estimating the similarity of new varieties to a reference baseline class of known safe varieties would subsequently greatly facilitate hazard identification. Further biological and eventually toxicological analysis would then only be necessary for varieties that fall outside this reference class. For this purpose, a one-class classifier tool was explored to assess and classify transcriptome profiles of potato (*Solanum tuberosum*) varieties in a model study. Profiles of six different varieties, two locations of growth, two year of harvest and including biological and technical replication were used to build the model. Two scenarios were applied representing evaluation of a 'different' variety and a 'similar' variety. Within the model higher class distances resulted for the 'different' test set compared with the 'similar' test set. The present study may contribute to a more global hazard identification of novel plant varieties.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

The development of novel plant varieties has led to basic questions regarding their safety assessment. This discussion has so far mainly focused on the safety assessment of genetically modified (GM) plant varieties, but is generally applicable to other types of new crop plants. The basic, internationally accepted approach for safety evaluation of novel plant varieties is a comparative safety assessment; new varieties should be compared with those with a history of safe human consumption (FAO, 1996; Kok et al., 2008; Kok and Kuiper, 2003; OECD, 1993, 2002). The assessment should comprise the concepts of hazard identification, hazard characterization, and exposure assessment, leading to a risk characterization that will include both intended as well as potential unintended effects of the breeding program, whether this includes genetic modification or not (Knudsen et al., 2008; Renwick, 2004). Within Europe, the concept of the comparative safety assessment for new GM plant varieties has been detailed by the European Food Safety Authority (EFSA) in a guidance document (EFSA, 2011). This is to a large extent included in EU legislation (European Commission, 2013).

An important part of the hazard identification is a compositional analysis of the GM plant variety compared with one or more conventional comparators, as formulated by the Food and Agriculture Organisation (FAO), the Organisation for Economic Co-operation and Development (OECD) and EFSA (EFSA, 2006, 2011; FAO, 1996; OECD, 1993). Compositional analysis should include all key compounds such as nutrients, anti-nutrients, and natural toxins (El Sanhoty et al., 2004). Key compounds have been described for various crops by the OECD Task Force in consensus documents (OECD, 2002). On the one hand, statistical tests are performed to identify differences for certain compounds with a direct comparator, *e.g.* the parental genotype. On the other hand, a wider comparison is made to the natural variation of these compounds under different environmental conditions, and indeed in different

* Corresponding author. Address: RIKILT Wageningen UR, PO Box 230, 6700AE Wageningen, The Netherlands. Fax: +31 (0) 317 417717.
E-mail address: jeroen.vandijk@wur.nl (J.P. van Dijk).
[1] The first two authors contributed equally to this manuscript.

http://dx.doi.org/10.1016/j.yrtph.2014.07.013
0273-2300/© 2014 Elsevier Inc. All rights reserved.

varieties of the crop that we consider as safe, for instance in an equivalence testing approach (van der Voet et al., 2011).

In 2013, the EU Standing Committee on the Food Chain and Animal Health updated the regulation on applications for authorizing GM food and feed in the European Union (European Commission, 2013). One new item in this regulation is the demand of a 90-day feeding trial with the whole food in rodents for every single transformation event and, in specific cases, the same trial for plants containing transformation events stacked by conventional crossing. However, the EFSA guidance from 2011 recommends this type of experimentation only under certain conditions. (EFSA, 2011). Also in a number of scientific journals commentaries have been published questioning various aspects of current regulations regarding GM crops (DeFrancesco, 2013; Herman and Price, 2013; Kuiper et al., 2013). We hypothesize that moving from targeted analysis to untargeted profiling will be of more added value to hazard identification than performing animal feeding trials. A proof of principle study is presented in this paper.

In recent years, different 'omics' strategies have come of age and might therefore be used for untargeted profiling for comparative compositional analysis. Of the various omics approaches, transcriptomics still by far has the largest coverage of the biological system, as compared with *e.g.* metabolomics and proteomics. Therefore this is the method of choice when the comparison should be as broad as possible. Several studies have shown reproducible differential transcriptome profiles from plant products in different situations, related to GM (Barros et al., 2010; Baudo et al., 2006; Cheng et al., 2008; Coll et al., 2010), but also other factors such as agricultural input, year of harvest, and location of growth (van Dijk et al., 2012, 2009; Zorb et al., 2009). Interpreting toxicological relevance of observed differences has been hampered by the unknown toxicological impact of many of the underlying genes and pathways. A second issue is the type of data analysis. Generally, a multivariate method has been used to explore the data followed by a univariate analysis on single gene or pathway level. This univariate part suffers from a high probability of false discoveries due to multiple testing (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003). Linked to this is the problem of many more variables than samples, often referred to as the 'large *p*, small *n*' paradigm (Kosorok and Ma, 2007). Consequently, a typical outcome of a transcriptomics comparison has been a list of differentially expressed genes, with a certain *p*-value, an estimation of the false discovery rate, with unknown function for many of them. Such an outcome provides a good starting point for mechanistic elucidation or in research fields such as drug development or disease diagnostics, where the goal is to find the acting genes in a comparison of situations already known to be different. It is however much less suited for food safety evaluation, where the existence of a difference actually has to be established first. An improved way to use omics data for this particular purpose is to incorporate the biological knowledge of the crops that are considered as safe as a baseline for the assessment of new crops. The first step would be to estimate whether or not a novel plant falls within this single class of 'safe' plants, based on the gene expression profile as a whole, through multivariate classification, explained in more detail in the next paragraph. This estimation can be calibrated against profiles of known unsafe crop plants or crop plants with otherwise undesirable characteristics. The crop plants whose profiles classify outside the safe baseline class will need further assessment. This approach is similar to current practice for compositional analysis based on single targets. The next phase of the hazard characterization should then determine whether the outlier profiles are of toxicological relevance (see Fig. 1). This will generally be first the identification of the variables that are causing the sample to be classified outside of the safe baseline class. These variables will then form the basis for further assessment.

Multivariate classification takes into account the profiles of (many) variables such as genome-wide gene expression values within a plant as well as potential biological relations between them. This approach leads to a translation of a profile to class membership, via a so-called classifier. This classifier can be used to designate a new sample into one or more predefined classes. For hazard identification of novel plant varieties, the most likely fit-for-purpose approach is 'one-class classification'. This has been applied in various situations when outside the baseline class there is (1) a scarcity of samples, or (2) a too-broad diversity (Tax, 2001). In food safety evaluation, unsafe plant varieties are both scarce and diverse which warrants the use of a one-class classifier to construct a 'food safety baseline'. This baseline, and therefore this classification, should include several parameters such as different cultivars, harvest years and soil types and geographical location (Berrueta et al., 2007), if this variation is likely to be present in the variety to be assessed.

The present study aims to explore one-class classification for transcriptomics profiles using the Soft Independent Modeling of Class Analogy (SIMCA) method (De Maesschalck et al., 1999; Wold and Sjöström, 1977). Potato sample profiles were used with a number of different, well-defined sources of variation. The applicability of this one-class classification for hazard identification as part of improved safety evaluation of novel plant varieties is assessed and discussed.

2. Materials and methods

2.1. Field experimental design

Five potato varieties (Biogold, Fontane, Innovator, Lady Rosetta and Maris Piper) were grown in Wageningen, the Netherlands and

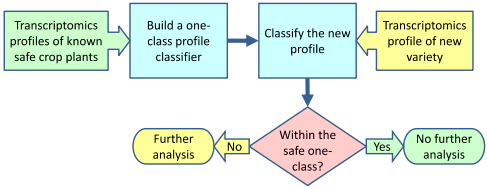


Fig. 1. Schematic overview of the proposed approach for food safety evaluation. The basis is an expansion of the current targeted comparative compositional analysis for hazard identification in food safety evaluation to an untargeted comparison based on transcriptomics profiles. It further proposes the application of multivariate one-class classification for hazard identification, based on whether or not profiles of novel plant varieties fall within or outside of a one-class of known 'safe' profiles. Further hazard characterisation should only be performed for classifications outside the safe one-class.

harvested in September 2010. Each variety was grown in two lots of different fields; one lot was raised in clay soil and the other one in sandy soil, having two individual plots per cultivar, and in each plot, four individual plants. The exception was Maris Piper, which in both fields was grown in a single plot containing three individual plants. The tubers were stored for one week in the dark at room temperature before further procedures. One additional potato variety (Sante) was grown in the UK as part of the QLIF study (quality low input food), as described in (van Dijk et al., 2012), samples identified by the numbers 1058 and 1059 were selected. A technical replication was performed for these two UK samples. Profiles were included that were analyzed in the original study in 2008 and again together with the NL samples in the present study, resulting in four profiles for these two samples. It is important to notice that one of the 'original' profiles was similar to most other samples in that study, while the other was more extreme, based on the principal component analysis (PCA) of the original UK study (van Dijk et al., 2012). The more extreme sample was dubbed 'outlier' in the present study. The technical replication of the two UK samples was performed starting from a second RNA isolation from the same freeze-dried potato tuber powders, stored at −80 °C.

To explore the differences *per* variety, with biological variation between plots included in replicates, the tubers were pooled for each individual plot and pre-processed according to (Lehesranta et al., 2007), to exclude variation between individual plants and/or tubers. In short: four average sized tubers were selected; of these, opposite eights were pooled to minimize variation effects in the tuber. After chopping, the potatoes were immediately submerged in liquid nitrogen, freeze-dried for 18 h, crushed to a fine powder with mortar and pestle and stored at −80 °C until RNA isolation.

2.2. RNA isolation and quality assessment

RNA was isolated from 0.5 g of each freeze-dried sample according to (van Dijk et al., 2009), based on lysis with a hexadecyltrimethylammonium bromide (CTAB) buffer, followed by chloroform/isoamyl alcohol extraction and overnight precipitation with lithium chloride. The modifications were: the extraction buffer was pre-warmed to 60 °C before use, the chloroform/isoamyl alcohol extraction was repeated three times before the lithium chloride precipitation and the final precipitation with 96% ethanol was performed with the tubes kept on ice and then centrifuged at 4 °C for 15 min at 14000 g. Total RNA isolated was dissolved in 100 μL of 10 mM Tris (pH 7) and warmed to 65 °C for 10 min.

The purity and concentration of the RNA was assessed from absorbance measurements using a Nanodrop 1000 instrument (Thermo Fisher Scientific, NanoDrop Technologies Wilmington, DE, USA). In order to evaluate its integrity, 1 μg of RNA was migrated by electrophoresis (60 min, 80 V) in denaturing agarose gel (1% agarose, 5% formamide, 1× TBE) stained with ethidium bromide. Gels were visualized in a Gel Doc XR+ Systems (Bio-Rad Laboratories, Life Technologies Corporation, Carlsbad, CA, USA) and analyzed using Quantity One 1-D (Bio-Rad Laboratories).

2.3. Labeling and hybridization

Purified RNA (2 μg) was labeled by incorporation of Cyanine-3-dCTP (Cy3-dCTP) during cDNA synthesis using the Quick Amp Labeling Kit (Agilent Technologies, Inc., Santa Clara, CA, USA). All cRNA synthesized from the labeled cDNA was purified and its purity and yield were assessed.

The cRNAs were randomly distributed over the arrays of 4×44K POCI slides (Kloosterman et al., 2008) and hybridized for 17 h at 65 °C in an Agilent Hybridization oven. The slides were then washed with Washing Buffers 1 and 2, acetonitrile and finally with stabilization and drying solution (Agilent), according to the manufacturer's instructions. The slides were kept at room temperature and immediately scanned.

2.4. Scanning, image analysis and microarray data analysis

Slides were scanned after excitation of Cy3 at 543 nm in an Agilent C Scanner, using the default settings of the equipment. Tiff images were imported to the Agilent Feature Extraction Software v8.5 and visually inspected prior to subsequent analysis. The fluorescent intensity and background signals were examined for each array to determine array quality.

The data were collected from columns containing feature number, accession number, spot signal, and background signal. Control spots (those without accession numbers) were removed as well as the genes that presented at least one fluorescent intensity signal lower than two times the correspondent background signal. The selected spot signals were log2 transformed. After the transformation data were normalized per array by subtracting from each individual spot signal the median value of the 20,370 spots of that array and also data were normalized per spot by subtracting the median value of the signal of all of the 22 arrays used in this study. Of the 44K variables for each profile, only those were included that showed a signal higher than twice the local background for all microarrays in this study and all the microarrays in the original UK study (van Dijk et al., 2012), resulting in 20,370 variables out of the original 44k readouts on the chip. The raw data and processed data have been deposited under number E-MTAB-1707 in the ArrayExpress database of the European Bioinformatics Institute (Brazma et al., 2000).

PCA and SIMCA were performed using Pirouette Software v.4 (Infometrix, Inc., Bothell, WA, USA).

3. Theory

3.1. Multivariate classification

Building a robust supervised, multivariate classifier requires three independent sets of samples of known (un)safety (Massart et al., 1997; Vandeginste et al., 1998). The *training set* is used to build a rough version of the classifier, fitting the model parameters to optimize their classification. A separate *validation set* is required to adapt the 'meta-parameters' of the classifier, such as the number of components in a component model such as SIMCA or the variables selected for classification. This validation step can be performed by 'internal cross-validation', commonly performed when only a limited set of samples is available (Wold, 1978). In cross-validation, one group of samples is removed at a time and a classifier is constructed using the remaining samples, after which the left-out samples are used as validation set. This procedure is repeated until all the groups of samples are left out once and validated in the different classifiers. However, because the validation set is used for choosing the meta-parameters of the optimal classifier, they are no longer independent of the final model. Therefore, a third *test set* is used to determine the classification accuracy of the resulting classifier. This threefold approach is essential to avoid classifier 'overfit', when classifiers describe the specific samples within the dataset, rather than the population from which they are taken. SIMCA classifiers determine for each sample a class distance based on the sample profile. Based on the class distances of the training set a 95% CI can be set which can be used as a threshold for class membership of the test samples. A test sample is classified as not belonging to the same class when a class distance higher than the 95% CI is observed. For the present study a one-class classifier is used, hence the class distance relates in all cases to this one class

of profiles that are considered as safe. We chose SIMCA as the classification tool as it uses PCA to reduce the complexity in transcription variability, which is expected to benefit the building of a classifier with a small but representative set of samples (Tax, 2001). Also, we chose to use all the separate classifiers resulting from the cross-validation as opposed to merging them into an overall classifier, analogous to (Westerhuis et al., 2008). The authors argue against the use of a single final model and instead promote the use of many slightly different models to obtain a range of class membership predictions, because at the moment there are no accepted criteria for the way to choose the overall model.

3.2. Study set-up in two scenarios; expected classification outcomes

Two scenarios were explored to evaluate the influence of different sources of variation on the transcriptome and subsequently on the classification outcomes. Potato variety, location of growth, year of harvest, biological replication, and technical replication (year of analysis) were included as sources of variation in a total of 22 profiles from 20 samples of potato tubers (Fig. 2). All profiles were used in both scenarios, the difference being the distribution of the profiles over the test and baseline sets.

The four UK sample profiles were *a priori* expected to be most 'different' because of different variety type, harvest year, and growing location compared with the NL sample profiles. Within this group of four, the two profiles from the analysis two years earlier (UK 2008) were expected to be even more different. Furthermore, the outlier profile, which was the most different from the other profiles in the original study was likewise expected to be the most different from all other profiles in the present study. Since all potato varieties in this study are safe for human consumption, all samples were expected to be classified as belonging to the safe baseline class, with the exception of the outlier profile. In the 'different' scenario, the four UK profiles, all of variety Sante, were used as the test set. The baseline varieties in this case (Biogold, Fontane, Innovator, Lady Rosetta, Maris Piper) shared the same harvest year, growing location, and time of analysis, all different from the test set. Misclassifications are expected in this scenario as the baseline set is not expected to sufficiently represent the variation present in the test set.

In the other, 'similar', scenario, the two UK Sante profiles analyzed in 2011 were grouped with the baseline sample profiles. In exchange, the NL variety Maris Piper with two sample profiles was grouped with the test set, keeping the number of profiles in test and baseline sets the same for both scenarios. The two UK Sante profiles analyzed in 2008 were kept in the test set. In this way, the two years of harvest (2005 and 2010) and the two growing locations (NL and UK) were present in both test and baseline sets. The variety Sante, grown in the UK, was also present in test and baseline profiles. The difference between the test and baseline sets was variety for the NL profiles and technical replication for the UK samples. On top of that, the outlier UK 2008 profile was still expected to be the most different sample. Fewer misclassifications were expected in this scenario.

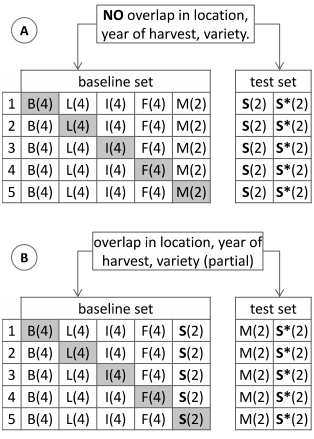


Fig. 2. Distribution of profiles for the different and the similar scenario. Each row contains the same and complete number of profiles. Profiles were distributed in two scenarios representing a different test set (A) or a similar test set (B). Furthermore, within each scenario, five classifiers were built with the baseline set with alternating varieties as cross-validation set (gray shading). Regular font indicates location and year of harvest: the Netherlands 2010. Bold font indicates location and year of harvest: United Kingdom 2005. *Year of analysis 2008 instead of 2011, including the outlier profile. B: Biogold, F: Fontane, L: Lady Rosetta, I: Innovator, M: Maris Piper, S: Sante.

4. Results

4.1. Model building and cross validation

The first two components of the PCAs for the baseline profiles in both scenarios explained in Section 3.2 showed a clear grouping according to variety (Fig. 3). In both PCAs, the first five components showed various groupings, all related to variety, explaining a total of 84.3% of variance for the 'different' scenario and 85.8% for the similar scenario. No grouping according to another source of variation was observed for the remaining components up to 10, explaining 94.9% of the total variation in the 'different' scenario and 95.3% in the 'similar' scenario. Therefore, stratified internal cross-validation of the baseline set was performed five times by leaving out one variety at a time, resulting in five classifiers, each with four varieties as training set and one variety as a validation set. For each of the five classifiers per scenario, the optimal number of components to be included in the SIMCA model was chosen. This optimal number was defined as the highest number still causing all profiles of the validation set to belong to the baseline class, with a confidence of 95%, which was used as threshold. An example of this procedure for the different scenario and variety Fontane as validation set is shown in Fig. 4. Instead of harmonizing the five classifiers into a final one, all five separate classifiers were used to analyze the test sets, leading to five classifications for the 'different' test set and five for the 'similar' test set.

4.2. Observed classification outcomes

The results of all SIMCA classifiers are summarized in Table 1. For the 'different' scenario, all the test profiles showed indeed larger class distances than those of the baseline profiles for all classifiers. The profiles analyzed in 2008, which had a technical replication as added source of variation, showed higher class distances than the profiles analyzed in 2011. Furthermore, the 'outlier' profile showed the highest class distance for all classifiers. The outlier profile was classified as different from the baseline class in all five classifiers, when using the 95% confidence limit

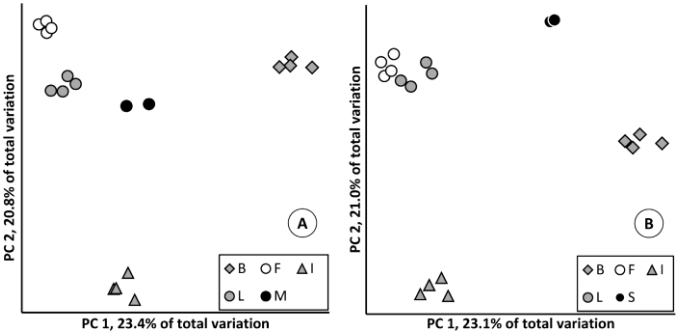


Fig. 3. PCA plots reveal grouping according to potato variety. PCA score plots are shown for the training/validation profiles for the 'different' (A) and the 'similar' scenario (B). B: Biogold, F: Fontane, I: Innovator, L: Lady Rosetta, M: Maris Piper, S: Sante (UK-2011 analysis).

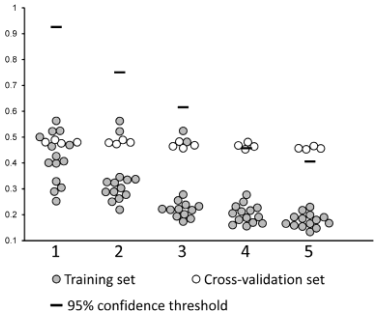


Fig. 4. Cross-validation of a SIMCA classifier. Shown are the class distances (*y*-axis) for different numbers of principal components (*x*-axis) included in the SIMCA classifier. In this graph as an example the cross-validation variety was Fontane, in the different scenario, and the optimal number of components was 3.

(CI) as a threshold for the class membership. However, also for the other three test samples, nine classifications were 'different', which should be considered false positives in this model study.

For the 'similar' scenario, the test profiles' class distances were indeed closer to those of the baseline profiles than for the 'different' scenario. In fact, overlap between test and baseline class distances was observed for four of the five classifiers. Also, the outlier UK 2008 profile showed again the highest class distances in all cases. When using the 95% CI as cut-off, the NL profiles in this test set were classified as belonging to the baseline in 7 of the ten cases, the average UK 2008 profile as belonging to the baseline in 4 of the 5 cases and the outlier UK 2008 profile as not belonging to the baseline in 4 of the 5 cases. All class distances are plotted in Fig. 5, including those of the different training, validation, and test sets for both scenarios.

5. Discussion

A set of transcriptome profiles of potato tuber samples was used to explore the possibilities of supervised multivariate classification as a tool for hazard identification as part of the food safety evaluation of new varieties. One of the objectives was to see how different sources of variation would influence the classification outcomes. Potato variety, location of growth, year of harvest, biological replication, and technical replication were included as sources of variation. One extreme sample was included as a proxy that was expected to be classified as outside of the baseline in most if not all cases, allowing for misclassifications due to a 95% confidence level. This sample was referred to as 'outlier' in the present study for clarity even though it was not considered an outlier in the original study (van Dijk et al., 2012), as the sample was still part of the grouping in the PCA according to the treatment groups in that study.

The scenarios combined underlined the importance of including enough representative variation for the making of a classifier in order to avoid false positives. In the 'different' scenario, 9 out of the 20 individual classifications were false positives in the light of safety evaluation in this model study. The classification as 'outside', was based here on the combined difference in harvest year, location and variety and not on any real unsafety of the samples. In the 'similar' scenario, class distances for test and baseline samples were closer together and even overlapping, except for the outlier sample. The reason is that the baseline samples used to build this classifier were more representative of the samples in this particular test set.

The currently available datasets were not representative for the true population of potato varieties on the market. Consequently, the results cannot be used to infer probabilities of false positive or false negative results. However, it is instructive to see how calculated error rates are dependent on the chosen methodology. For instance, the Maris Piper variety was classified as not belonging to the baseline in 3 of 10 classifier-profile combinations, although it is a perfectly safe potato. Therefore, one could say that there were 3 false positive observations for this cultivar. On the other hand, if an 'average' classification per sample were to be considered, the majority of classifiers classified both sample profiles of this variety

as belonging to the baseline, indicating 2 true negatives. Importantly, for the most different sample profile in this study, classification was 'outside of the baseline' for all classifiers in the 'different' scenario and 4 of 5 in the similar scenario. Such an outcome would lead to further investigation of the data in a food safety evaluation setting. That is, in this setting, Sante would clearly not be regarded as not part of the known safe class based on this sample alone, as the other samples were clearly less different. It rather proves that an outlier sample will be identified as such. Translating this to a real life situation, this means that if (1) replicates of a novel variety would fall outside of the safe class and (2) the samples in the safe class are representative of relevant normal variation, this new variety would indeed require further investigation.

This study shows that in principle transcriptome profiles can be used to classify potato tubers as belonging or not belonging to a known set of tubers. This classification may form the basis to identify potential hazards in novel potato varieties, for instance in case of an unintended effect of breeding strategies, including GM. It is important to note in this respect that hazard identification of novel crops as presented here is in the context of meeting regulatory requirements worldwide. In this respect it would be interesting, if not crucial, to include in the baseline those varieties that are produced worldwide using induced mutation technologies, that do not require the regulatory regimes of GM crops, and are as such regarded to be safe. When further developed and validated, this approach will likely be more informative and more cost-effective than the animal feeding trials with whole foods now mandatory in the EU to identify unintended effects in novel crop plants.

The classifiers presented here serve as a proof of principle. Some items need to be considered when applying this approach in real life. For instance, in different countries different varieties are used due to many factors including environmental factors; this methodology would therefore be best applied for comparative evaluation of GM vs non-GM with varieties specific to each country where the GMOs were introduced. Also, the life span of some varieties is relatively short and new varieties are produced so one would have to construct baseline transcriptomics profiles of new safe varieties as the old ones become redundant. Furthermore, for practical application and validation of multivariate one-class classification, the right threshold for membership of the one class needs to be determined. In safety evaluation, false negatives are of greater concern than false positives. False negatives may lead to increased risk, while false positives will merely lead to more than strictly necessary post-classification assessments of toxicological relevance for outside-of-the-one-class profiles. A well-defined study with known unsafe samples in the test set, besides known safe samples, will aid in the evaluation of such a threshold value and the proposed methodology as a whole. This classifier approach is still basically the same as the current targeted approach in compositional analyses of new plant varieties, but by using omics profiles,

Table 1a
Class distances of the test set and cross-validation outcome (CV) of the baseline set for the 'different' scenario.

CV variety	B	F	I	L	M
# Components	3	3	3	4	3
95% CI limit	0.594	0.616	0.599	0.462	0.569
S	0.555	0.570	0.575	**0.561**	**0.589**
S	0.555	0.570	0.573	**0.559**	**0.588**
S*	**0.610**	**0.627**	**0.621**	**0.614**	**0.641**
S**	**0.733**	**0.751**	**0.749**	**0.739**	**0.767**

Sample codes are identical to the ones in Fig. 1. *Year of analysis 2008 instead of 2011, **Outlier sample. Numbers in bold typeface indicate a class distance larger than that of the 95% CI. Gray shading indicates a mismatch between expected and observed classification.

Table 1b
Class distances of the test set and cross-validation outcome (CV) of the baseline set for the 'similar' scenario.

CV variety	B	F	I	L	S
# Components	2	3	3	4	2
95% CI limit	0.747	0.581	0.562	0.441	0.745
M	0.523	0.522	0.522	**0.522**	0.523
M	0.565	0.564	**0.564**	**0.564**	0.565
S*	0.529	0.488	0.471	**0.452**	0.644
S**	0.663	**0.628**	**0.618**	**0.605**	**0.770**

Sample codes are identical to the ones in Fig. 1. *Year of analysis 2008 instead of 2011, **Outlier sample. Numbers in bold typeface indicate a class distance larger than that of the 95% CI. Gray shading indicates a mismatch between expected and observed classification.

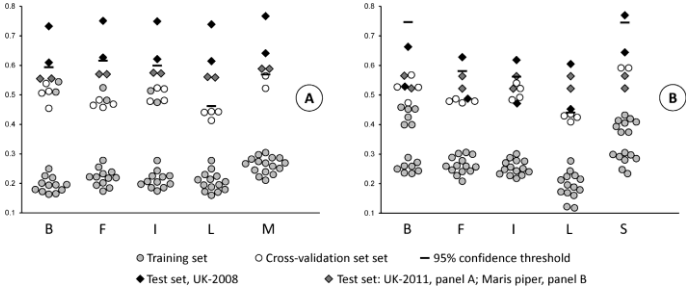


Fig. 5. Inclusion of more variation leads to better classification. Shown are the class distances after SIMCA modeling for two test sets. In panel A, the different scenario, the extra variation caused by a different year of harvest, location and technical replication was not included in the training/validation set. In panel B, the similar scenario, only technical replication was left out of the training plus validation set resulting in more samples being classified as belonging to the class. For both test sets, all five submodels resulting from the cross-validations are shown with the letters indicating the variety that was used as validation set: B: Biogold, F: Fontane, I: Innovator, L: Lady Rosetta, M: Maris Piper, S: Sante (UK-a/b-2).

the information content underlying the hazard identification will increase significantly.

The present study proposes an expansion of the current targeted to an untargeted comparative compositional analysis, based on omics profiles, for hazard identification in food safety evaluation. It further proposes the application of multivariate one-class classification for hazard identification, based on whether or not profiles of novel plant varieties fall within or outside of a class of profiles of varieties that are generally regarded as safe. The current basic approach of hazard identification based on observed differences remains unaltered. The match between expected and observed outcomes of the model dataset in the present study warrants further development and validation of this approach.

Note

CSM was supported by a doctoral fellowship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), Ministry of Education, Brazil, according to cooperation program CAPES-Wageningen (project 005/09). The work described in this paper was financially supported by the Dutch Ministry of Economic Affairs.

Conflict of interest

RIKILT, Wageninge UR received grants from the Ministry of Economic Affairs, the Netherlands, and the Federal University of Santa Catarina, Brazil, received grants from the Ministery of Education, Brazil.

References

Barros, E. et al., 2010. Comparison of two GM maize varieties with a near-isogenic non-GM variety using transcriptomics, proteomics and metabolomics. Plant Biotechnol. J. 8, 436–451.

Baudo, M.M. et al., 2006. Transgenesis has less impact on the transcriptome of wheat grain than conventional breeding. Plant Biotechnol. J. 4, 369–380.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J. R. Stat. Soc. B-Methodol. 57, 289–300.

Berrueta, L.A. et al., 2007. Supervised pattern recognition in food analysis. J. Chromatogr. A 1158, 196–214.

Brazma, A. et al., 2000. One-stop shop for microarray data. Nature 403, 699–700.

Cheng, K.C. et al., 2008. Effect of transgenes on global gene expression in soybean is within the natural range of variation of conventional cultivars. J. Agric. Food Chem. 56, 3057–3067.

Coll, A. et al., 2010. Natural variation explains most transcriptomic changes among maize plants of MON810 and comparable non-GM varieties subjected to two N-fertilization farming practices. Plant Mol. Biol. 73, 349–362.

De Maesschalck, R. et al., 1999. Decision criteria for soft independent modelling of class analogy applied to near infrared data. Chemometrics Intell. Lab. Syst. 47, 65–77.

DeFrancesco, L., 2013. How safe does transgenic food need to be? 31, 794–802.

EFSA, 2006. Guidance document of the scientific panel on genetically modified organisms for the risk assessment of genetically modified plants and derived food and feed. EFSA J. 99, 1–100.

EFSA, 2011. Guidance for risk assessment of food and feed from genetically modified plants. EFSA J. 9 (5), 2150, 1–37.

El Sanhoty, R. et al., 2004. Quality and safety evaluation of genetically modified potatoes Spunta with Cry V gene: compositional analysis, determination of some toxins, antinutrients compounds and feeding study in rats. Food/Nahrung 48, 13–18.

European Commission, 2013. Commission Implementing Regulation (EU) No 503/2013 of 3 April 2013 on applications for authorisation of genetically modified food and feed in accordance with Regulation (EC) No 1829/2003 of the European Parliament and of the Council and amending Commission Regulations (EC) No 641/2004 and (EC) No 1981/2006. Official J. European Union. 8.6.2013, No. L 157/1.

FAO, 1996. Biotechnology and Food Safety. Report of a Joint FAO/WHO Consultation. FAO Food and Nutrition Paper 61.

Herman, R.A., Price, W.D., 2013. Unintended compositional changes in genetically modified (GM) crops: 20 years of research. J. Agric. Food Chem. 61, 11695–11701.

Kloosterman, B. et al., 2008. Genes driving potato tuber initiation and growth: identification based on transcriptional changes using the POCI array. Funct. Integr. Genomics 8, 329–340.

Knudsen, I. et al., 2008. Risk management and risk assessment of novel plant foods: concepts and principles. Food Chem. Toxicol. 46, 1681–1705.

Kok, E.J. et al., 2008. Comparative safety assessment of plant-derived foods. Regul. Toxicol. Pharmacol. 50, 98–113.

Kok, E.J., Kuiper, H.A., 2003. Comparative safety assessment for biotech crops. Trends Biotechnol. 21, 439–444.

Kosorok, M.R., Ma, S., 2007. Marginal asymptotics for the "Large P, Small N" paradigm: with applications to microarray data. Ann. Stat. 35, 1456–1486.

Kuiper, H.A. et al., 2013. New EU legislation for risk assessment of GM food: no scientific justification for mandatory animal feeding trials. Plant Biotechnol. J. 11, 781–784.

Lehesranta, S.J. et al., 2007. Effects of agricultural production systems and their components on protein profiles of potato tubers. Proteomics 7, 597–604.

Massart, D.L. et al., 1997. Handbook of Chemometrics and Qualimetrics: Part A. Elsevier, Amsterdam.

OECD, 1993. Guidelines for the Testing of Chemicals. Organisation for Economic Cooperation and Development, Paris.

OECD, 2002. Consensus document on compositional considerations for new varieties of potatoes: key food and feed nutrients, anti-nutrients and toxicants. Series on the Safety of Novel Foods and Feeds, 4.

Renwick, A.G., 2004. Risk characterisation of chemicals in food. Toxicol. Lett. 149, 163–176.

Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. PNAS 100, 9440–9445.

Tax, D.M.J., 2001. One-Class Classification: Concept-Learning in the Absence of Counter-Examples. Technical University of Delft, Delft, the Netherlands.

van der Voet, H. et al., 2011. A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. BMC Biotechnol. 11 (15), 1–20.

van Dijk, J.P. et al., 2012. The identification and interpretation of differences in the transcriptomes of organically and conventionally grown potato tubers. J. Agric. Food Chem. 60, 2090–2101.

van Dijk, J.P. et al., 2009. Transcriptome analysis of potato tubers–effects of different agricultural practices. J. Agric. Food Chem. 57, 1612–1623.

Vandeginste, B.G.M. et al., 1998. Handbook of Chemometrics and Qualimetrics: Part B. Elsevier Science, Amsterdam.

Westerhuis, J.A. et al., 2008. Assessment of PLSDA cross validation. Metabolomics 4, 81–89.

Wold, S., 1978. Cross-validatory estimation of the number of components in factor and principal components models. Technometrics 20, 397–405.

Wold, S., Sjöström, M., 1977. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. Chemometrics: Theory and Application, vol. 52. American Chemical Society, pp. 243–282.

Zorb, C. et al., 2009. Levels of compounds and metabolites in wheat ears and grains in organic and conventional agriculture. J. Agric. Food Chem. 57, 9555–9562.