versão impressa ISSN 0101-7438 / versão online ISSN 1678-5142

OTIMIZAÇÃO DE MÉTODOS DE CONTROLE DE QUALIDADE DE FÁRMACOS USANDO ALGORITMO GENÉTICO E BUSCA TABU

Pedro Henrique de Almeida Konzen
João Carlos Furtado *
Claudia Wollmann Carvalho
Marco Flôres Ferrão
Rolf Fredi Molz
Irionson Antonio Bassani
Silvio Luiz Hüning
Universidade de Santa Cruz do Sul – UNISC
Santa Cruz do Sul – RS
jcarlos@dinf.unisc.br

* *Corresponding author*/autor para quem as correspondências devem ser encaminhadas

Recebido em 06/2002, aceito em 11/2002 após 1 revisão

Resumo

Neste trabalho novas metodologias de análise para a quantificação de princípios ativos em fármacos são estudadas, empregando técnicas de otimização meta-heurísticas em modelos de regressão multivariada e dados de espectroscopia por reflexão difusa no infravermelho com transformada de Fourier (DRIFTS). Inicialmente foram preparadas amostras do princípio ativo hidroclorotiazida, a partir de um planejamento experimental. Os espectros foram coletados empregando um espectrofotômetro, sendo aplicada a correção multiplicativa de sinal (MSC). Foram construídos modelos de calibração por mínimos quadrados parciais (PLS), sendo implementados métodos de otimização dos Algoritmos Genético (AG) e Busca Tabu (BT), objetivando minimizar o erro das amostras de validação (SEV). Os melhores modelos foram selecionados com base no coeficiente de regressão (R^2) e comparados. Ambas as técnicas de otimização resultaram em modelos de calibração com excelente capacidade de quantificação do princípio ativo, sendo adequadas para a utilização no controle de qualidade do processo produtivo.

Palavras-chave: algoritmo genético; busca tabu; quimiometria; meta-heurísticas.

Abstract

Two new methodologies of analysis for the antihipertensive drugs quantification are provided in this work, using meta-heuristic algorithms optimization with infrared spectroscopy by diffuse reflectance Fourier transform (DRIFTS) and partial-least squares regression method (PLS). A diffuse reflectance joined to a spectrophotometry in the mid-infrared region was used. The calibration was through PLS in order to use it as a pre-processing. The data were used in multiplicative signal correction (MSC). After the best wavelength selection, which results in models with better coefficient regression (R^2), these were improved using routines based on genetic algorithm heuristic and tabu search methods. For the evaluation of the optimization a standard error of validation (SEV) was used. The new technique using DRIFTS/PLS/AG or DRIFTS/PLS/TS shows excellent choices for the quality control of the productive process of factories and drugstore.

Keywords: genetic algorithm; tabu search; chemometrics; meta-heuristics.

1. Introdução

Recentemente a aplicação da técnica de transformada de Fourier na espectroscopia por infravermelho (FT–IR) [Birth & Hecht (1987); Osborne, Fearn & Hindle (1993)] e de técnicas quimiométricas de regressão multivariada [Sjöström, Wold, Lindberg, Persson, & Martens (1983); Geladi (2002)], possibilitaram a análise de misturas complexas, como fármacos, sem a necessidade de qualquer separação prévia de seus componentes.

Este avanço da espectroscopia no infravermelho médio, como técnica para análise quantitativa, deve-se à combinação da transformada de Fourier, da nova geometria dos espectrofotômetros com a utilização do interferômetro de Michelson [Kalasinsky (1990), Durig & Sullivan (1990), Eikrem (1990), Coates (1998)] e principalmente da utilização dos algoritmos de calibração multivariada que permitiram a modelagem de muitos sinais do espectro simultaneamente.

Entretanto, o elevado número de variáveis, geradas na espectroscopia no infravermelho, tem resultado em modelos com excelente desempenho na calibração, porém com pouca habilidade de previsão de amostras externas a modelagem. Uma alternativa para melhorar o condicionamento dos modelos está em selecionar um conjunto de variáveis que intensifiquem o comportamento sinérgico na modelagem, de forma a resultar numa habilidade equivalente na calibração e na previsão. A grande dificuldade existente é realizar essa seleção, pois, o número de variáveis existentes é extremamente elevado, bem como as combinações resultantes. Com este propósito, métodos de otimização têm sido propostos, sendo o Algoritmo Genético (AG) o mais usado [Leardi (2001); Costa Filho & Poppi (1999)]. Entre outras razões, a flexibilidade do AG em permitir o seu uso na otimização de diferentes problemas combinatoriais e em suas formulações matemáticas, tem sido apontado como característica importante do AG (Holland, 1975).

Neste trabalho, foi implementado o AG e a Busca Tabu (BT) (Glover, 1989a; 1989b) em processos de otimização de modelos de regressão multivariada via dados de espectroscopia por reflexão difusa no infravermelho com transformada de Fourier (DRIFTS). Esta abordagem traz pela primeira vez o uso da BT para a seleção de comprimentos de onda de espectros no infravermelho visando a calibração multivariada.

No processo foi aplicado à regressão por mínimos quadrados parciais (PLS) na determinação de princípios ativos de ação anti-hipertensiva presentes em fármacos; visando a análise quantitativa da hidroclorotiazida, pertencente à classe dos diuréticos tiazídicos.

2. Fármacos Anti-hipertensivos

Alguns fármacos têm a propriedade de reduzir a pressão atuando na resistência periférica e/ou no débito cardíaco, conhecido como anti-hipertensivos. Entre os anti-hipertensivos mais empregados encontramos os da classe das Tiazidas, também chamadas derivados da benzotiazina ou benzotiazidas, que são quimicamentes dióxidos de benzotiadiazina-7-sulfonamida (Harvey & Champe, 1998).

O princípio ativo 6-cloro-3,4-diidro-2*H*-1,2,4-benzotiodiazina-7-sulfonamida-1,1-dióxido, conhecido como hidroclorotiazida, apresenta fórmula estrutural conforme **Figura 1**.

A hidroclorotiazida apresenta-se como um pó cristalino branco ou quase branco, inodoro e com peso molecular de 297,7 g/mol. É muito pouco solúvel em água, pouco solúvel em álcool, solúvel em acetona e em soluções diluídas de hidróxidos alcalinos, apresentando faixa de fusão de 266-270^{o}C, com decomposição.

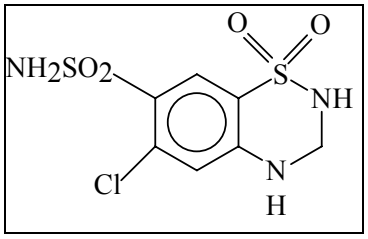

Figura 1 – Estrutura da hidroclorotiazida

Os comprimidos de hidroclorotiazida devem conter, no mínimo, 93% e, no máximo, 107% da quantidade de hidroclorotiazida declarada. São realizados testes de desintegração, dureza, friabilidade e uniformidade de conteúdo para a caracterização dos mesmos. A quantificação do princípio ativo é realizada através da técnica de espectrofotometria no ultravioleta e por cromatografia líquida de alta eficiência (Farmacopéia Brasileira, 1988).

3. Reflexão Difusa

A técnica de reflexão difusa está bem fundamentada na literatura [Wilson (1990); Van de Voort (1992); Hart, Norris & Golumbic (1962); Abdullah & Sherman (1997); Birth & Hecht (1987); Culler (1993); Osborne, Fearn & Hindle (1993)], sendo amplamente associada aos equipamentos que operam no infravermelho próximo, com os quais geralmente é designada de espectroscopia de reflexão no infravermelho próximo (NIRS ou NIRR); ou associada àqueles que operam na região do infravermelho médio, sendo conhecida por espectroscopia por reflexão difusa no infravermelho médio com transformada de Fourier (DRIFTS).

A reflexão difusa ocorre em superfícies não totalmente planas, podendo o substrato ser contínuo ou fragmentado (na forma de pó). Neste processo de reflexão o feixe incidente penetra a superfície da amostra interagindo com a matriz, retornando à superfície da mesma, após absorção parcial e múltiplos espalhamentos, conforme ilustrado pela **Figura 2**.

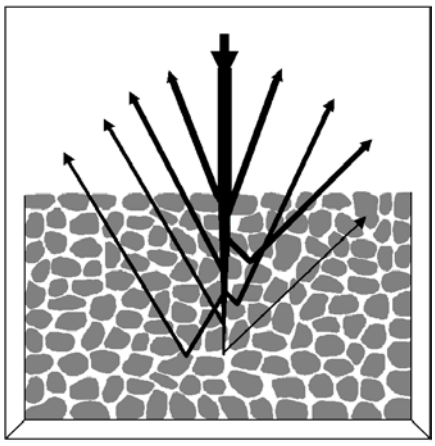

Figura 2 – Representação da reflexão especular e difusa de uma onda eletromagnética em uma amostra particulada (Wetzel, 1983).

Na reflexão difusa, a radiação incidente entra em contato diversas vezes com as partículas da amostra sendo conseqüentemente atenuada. Com isto a radiação que sofre a reflexão difusa fornece informações qualitativas e quantitativas, sobre a natureza química da amostra.

As informações qualitativas dizem respeito a capacidade de que cada funcionalidade da molécula orgânica tem de absorver energia, em determinada região do infravermelho, para promoção de vibração entre dois centros (átomos) que compõem determinada ligação.

Já a quantificação destas informações químicas seguem a função de Kubelka Munk (expressão 1) que relaciona os espectros de reflectância difusa com a concentração de cada molécula presente na amostra, transformando o espectro de reflectância em formato que assemelha-se a um espectro de absorbância. Por esta razão, a análise de Kubelka Munk é conhecida como a Lei de Beer da espectroscopia de reflectância.

$$f(R_\infty) = (1 - R_\infty)^2 / 2R_\infty \qquad (1)$$

onde R_∞ é a reflectância difusa.

De acordo com a teoria, $f(R_\infty)$ está relacionado com o coeficiente de absorção e o coeficiente de dispersão da superfície S de uma amostra espessa, conforme expressão 2.

$$f(R_\infty) = K/S \qquad (2)$$

O efeito do tamanho da partícula causa deslocamento da linha de base, efeito que torna-se muito pronunciado em comprimentos de onda de grande absorção pela amostra. Por exemplo, em duas amostras com mesma composição, porém diferente granulometria, há maior reflexão das partículas menores (mais finas), e os efeitos da reflexão especular são minimizados em uma superfície composta de menor tamanho.

A correta quantificação de analitos em amostras em pó é dependente do efetivo poder de penetração, sendo grande o suficiente para providenciar um espectro representativo da amostra por inteiro. A equação de Kubelka Munk é mais simples e a forma geralmente mais usada para definir uma afinidade linear entre intensidade da banda e concentração da amostra na espectroscopia por reflectância difusa.

Um maior detalhamento desta técnica, bem como de outras técnicas de reflexão empregadas em análises que potencialmente aplicam informações químicas obtidas via espectroscopia no infravermelho com transformada de Fourier, pode ser encontrado em recente revisão publicada por Ferrão (2001).

4. Regressão Multivariada

A técnica de regressão por mínimos quadrados parciais (PLS) foi desenvolvida na década de 70 por Herman Wold (1978) e foi usada, inicialmente, na região do infravermelho próximo.

No modelo PLS, há algumas suposições feitas em torno dos dados que serão analisados. As informações espectrais e as informações das concentrações são usadas ao mesmo tempo na fase de calibração (Sjostrom *et al.*, 1983).

O alicerce do método dos mínimos quadrados parciais, reside na decomposição de uma matriz de dados **X**, em termos da soma das várias matrizes **M**, que apresentem dimensionalidade um, e que são acrescidas de uma matriz de resíduos (que corresponde a parte não modelada de **X**), conforme equação 3.

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + ... + \mathbf{M}_a + \mathbf{E} \quad (3)$$

Onde *a* corresponde ao número de variáveis latentes (componentes principais ou fatores) selecionados para truncar a igualdade, e **E** corresponde a matriz de resíduos, relacionada ao número de variáveis latentes escolhido.

As matrizes **M** constituem os chamados componentes principais (ou variáveis latentes), e são formadas pelo produto de dois vetores, **t** (os escores) e **p** (os pesos) conforme as expressões 4 e 5.

$$\mathbf{X} = \mathbf{t}_1.\mathbf{p}_1^t + \mathbf{t}_2.\mathbf{p}_2^t + ... + \mathbf{t}_v.\mathbf{p}_a^t + \mathbf{E} \quad (4)$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E} \quad (5)$$

A dimensionalidade do espaço original é igual ao número de colunas em **X**, ou seja, o número de variáveis originais, expresso por *m*. No novo modelo, a dimensionalidade é descrita pelo número de matrizes $\mathbf{M}_i$ necessárias para descrever **X**. Assim, se for possível descrever uma matriz **X** que tenha muitas variáveis, por um número pequeno dessas matrizes $\mathbf{M}_i$ haverá um decréscimo na dimensionalidade, sem perda significativa de informação.

Na modelagem por mínimos quadrados parciais, tanto a matriz das variáveis independentes **X**, como a das variáveis dependentes **Y** são representadas pelos escores e pelos pesos, conforme as expressões 5 e 6.

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \quad (6)$$

A relação entre as duas matrizes de dados **X** e **Y** pode ser obtida, correlacionando-se os escores de cada bloco, afim de obter uma relação linear descrita pela expressão 7, ou de forma equivalente por 8.

$$\mathbf{U}_a = \mathbf{b}_a.\mathbf{t}_a \quad (7)$$

$$\mathbf{U} = \mathbf{b}\mathbf{T} + \mathbf{e} \quad (8)$$

Na equação linear representada em 8, **T** é uma matriz de resposta (assim como um conjunto de espectros) para uma série de amostras de calibração; **U** é uma matriz contendo as propriedades (variáveis dependentes) de todas estas amostras; **b** é um vetor contendo os parâmetros do modelo, e **e** é um vetor que representa o ruído do espectro e os erros do modelo.

Entretanto, pode ocorrer que o modelo assim concebido não seja totalmente satisfatório. Isto porque cada matriz (**X** e **Y**) é decomposta separadamente, podendo resultar numa relação não muito adequada (não linear) entre os escores dos dois blocos.

Em virtude disto, deve-se buscar um modelo onde as matrizes de resíduos **E** e **F** sejam as menores possível e, ao mesmo tempo, conseguir uma relação linear entre **T** e **U**.

No método de regressão PLS, este ajuste é concretizado através de uma leve mudança nos valores dos escores, de forma a produzir a melhor correlação possível.

Existe um compromisso entre a capacidade em descrever as amostras nos espaços individuais (modelagem dos blocos **X** e **Y**), e o aumento na correlação **T** e **U** (Adams, 1995).

O método de regressão por mínimos quadrados parciais foi amplamente utilizado na última década, tendo se tornado um dos mais populares em função do seu desempenho (Pimentel & Neto, 1996).

Usualmente os modelos de calibração multivariados são avaliados a partir da correlação entre a técnica alternativa proposta (neste caso a DRIFTS) e o valor de referência (geralmente o método oficial) com base nos valores do coeficiente de correlação (R^2) e dos erros padrões de calibração (SEC) e de validação (SEV), sendo estes últimos calculados conforme equação 9, onde y_i e $\hat{y}_i$ são respectivamente os valores de referência e estimado para a i-ésima amostra e *n* o número de amostras.

$$\text{SEC (ou SEV)} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \qquad (9)$$

Neste processo um outro parâmetro crítico é o número de variáveis latentes (VL's) empregados na modelagem. De uma forma geral quanto maior o número de VL's utilizadas, melhor é o coeficiente de regressão (R^2) e conseqüentemente menor é o erro de calibração, visto que mais informações do conjunto de espectros dos padrões são modeladas. Entretanto, o SEV geralmente não acompanha este comportamento, tendendo atingir um mínimo e depois aumentar conforme são utilizadas mais VL's.

Não existe um método exato para determinar o número de VL's a ser empregado. A melhor forma, apresentada pelos pesquisadores da área, tem sido experimentar de forma crescente e verificar o comportamento da modelagem, principalmente os valores de SEV.

Um bom modelo é alcançado com o menor número de VL's que apresenta um SEV mínimo e os valores de SEC e R^2 satisfatórios.

5. Técnicas Heurísticas de Otimização

Dada a dificuldade de métodos matemáticos exatos em obter boas soluções, em tempo computacional aceitável, para problemas combinatórias que envolvem variáveis discretas, freqüentemente são propostos algoritmos heurísticos para obter soluções de qualidade.

Hibbert (1993) chama a atenção para as dificuldades encontradas na escolha de um método de otimização:

- Tamanho da região experimental: sistemas com muitas variáveis precisam ser eficientes na busca e determinação dos melhores parâmetros, pois muitas vezes o tempo é um fator limitante para a realização de uma otimização;
- Continuidade na superfície de respostas: o método utilizado deve possuir continuidade na superfície de resposta, isto é, os resultados não devem sofrer influências de informações que estão nos limites da região experimental (bordas da matriz);
- Detecção de ótimos locais: o método utilizado tem de ser sensível à presença de ótimos locais, e precisos na detecção de ótimos globais;
- Sensibilidade na detecção dos parâmetros: os métodos podem ser prejudicados se as funções forem demasiado sensíveis para um determinado parâmetro. O processo de otimização gastaria muito tempo para aquele parâmetro enquanto ignoraria o resto.

Nos problemas encontrados no dia-a-dia, combinações desses fatores, alguns dos quais desconhecidos, fazem da escolha do método de otimização uma tarefa bastante difícil. Neste sentido, os métodos heurísticos estão sendo cada vez mais aplicados.

5.1 Algoritmo Genético

Algoritmo Genético (AG) é uma técnica de busca aleatória direcionada, desenvolvida por Holland (1975), capaz de obter a solução ótima global num espaço de busca complexo multi-dimensional. O AG é baseado na evolução das espécies, usando operadores inspirados no processo de evolução natural. Estes operadores, conhecidos como operadores genéticos de *crossover* e mutação, manipulam indivíduos de uma população, através de gerações, para melhorar (aperfeiçoar) a adaptação (*fitness*) gradativamente. Os indivíduos numa população, também denominados de cromossomos, são representados por cadeias (*strings*) de números binários. A função de avaliação (*fitness*) estabelece a relação entre o AG e o problema de otimização.

O AG revela-se útil na resolução de problemas complexos (Lorena & Furtado, 2001), que podem ser divididos em três classes, segundo Lucasius & Kateman (1993) na primeira classe de problemas em que os AG's são passíveis de aplicação, encontra-se a otimização de modelos que descrevam relações entre informações coletadas de um sistema (espectros de substâncias, por exemplo) a valores de propriedades (respostas do sistema) que se queira estudar. Um segundo tipo de aplicação para os AG's seria na resolução de problemas envolvendo relações, capacidade de planejamento e tomada de decisões. A seleção de subconjuntos de dados de um sistema ilustra a terceira classe de aplicações a que os AG's se destinam.

Como métodos de otimização, os AG's encontram-se em uma posição intermediária entre os métodos de otimização classificados como "fracos" e "fortes" – são considerados moderados. Os métodos de otimização "fracos" fazem poucas suposições quanto às hipóteses do problema. Eles podem ser amplamente empregados, porém levam um tempo relativamente grande para encontrar uma solução aceitável. De outro lado, métodos de otimização "fortes", se utilizam de várias suposições a respeito do problema, convergindo rapidamente para um resultado, sendo limitados a uma pequena região de problemas relativamente simples. Quando aplicados a um problema que excede os limites dessa região, estes métodos tendem a encontrar resultados errôneos ou entrar em colapso (Lucasius & Kateman, 1993).

Os métodos moderados podem ser competitivos sob circunstância onde tanto métodos fracos e fortes costumam falhar. No problema em questão, métodos fortes iriam aumentar o erro nas suposições que faria do modelo, até chegar ao colapso. De outro lado, como o espaço dimensional é um tanto grande, métodos fracos despenderiam muito tempo, sendo praticamente inviáveis. Normalmente, em circunstâncias do dia-a-dia, os métodos moderados são os mais atrativos para a solução de problemas de otimização.

No campo da instrumentação, os AG's têm aplicação em diversas áreas, tais como no treinamento de redes neurais para interpretação de espectros de fluorescência de raios-x, seleção de comprimentos de onda em análises multicomponentes, otimização de dados de calibração por seleção de comprimentos de onda, entre outros, sendo esta última aplicação o foco deste trabalho.

O princípio e funcionamento dos AG's é baseado no processo evolutivo de seres vivos, seguindo o princípio básico de que as gerações derivadas serão mais "evoluídas" do que os seus precursores. Gerações melhores continuariam existindo, enquanto que gerações mais "frágeis", tenderiam a sucumbir.

Nos AG's, as condições em que o experimento é conduzido são consideradas como sendo o material genético dos seres vivos, e a resposta experimental é considerada como medida de adaptação ao ambiente. O material genético é assumido como sendo um simples cromossomo, onde cada gene corresponde a uma variável do processo. A **Figura 3a** representa um desses cromossomos.

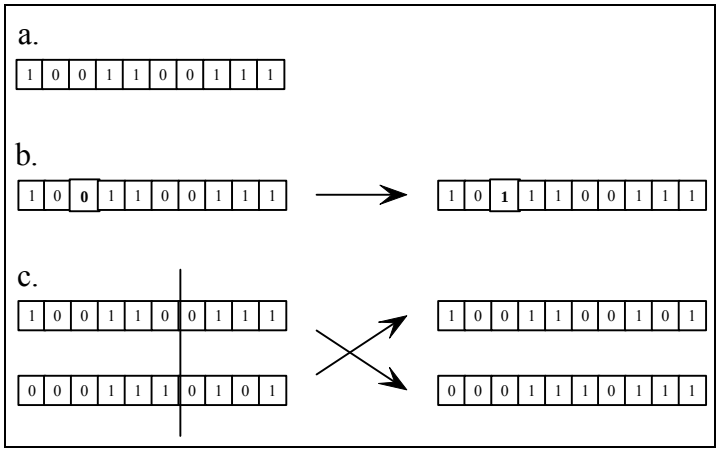

Figura 3 – Representação de um cromossomo (a) e de suas propriedades de mutação (b) e recombinação (c).

A avaliação da população é feita através de uma função que calcula a resposta de determinado cromossomo em relação ao modelo. Esta função é também chamada de *fitness*. Shaffer *et al.* (1996) demonstram em seu trabalho que a função *fitness* baseada no erro quadrático do conjunto de calibração, no erro quadrático do conjunto de validação e no número de fatores utilizados no PLS apresentam-se como uma das melhores.

A evolução da população é realizada através de operadores genéticos: *crossover* e mutação. No *crossover*, a partir de um determinado número de cruzamentos entre os cromossomos, selecionados aleatoriamente a partir das melhores respostas experimentais, estes combinam-se formando novos indivíduos (**Figura 3c**). Na mutação, os componentes de uma população podem sofrer perturbações em seus genes (**Figura 3b**).

Basicamente, pode-se descrever o funcionamento dos AG's nos passos mostrados na **Figura 4**:

Passo 1. Gerar uma população inicial de soluções;
Passo 2. Avaliar a *fitness* dos indivíduos da população
Passo 3. Repetir
Início
(a) Selecionar um conjunto de pais na população e realizar a mutação
(b) Cruzar os pais de modo que se reproduzam
(c) Avaliar a *fitness* dos filhos gerados
(d) Substituir os filhos julgados inadequados
Fim
(e) Até que o critério de parada seja atendido
Passo 4. Fim.

Figura 4 – Algoritmo simplificado do AG implementado.

5.1.1 Aplicação do AG na Seleção de Comprimentos de Onda de Infravermelho

O funcionamento do AG, aplicado ao problema, obedece aos passos descritos na **Figura 4**:

Passo 1: Gerar uma população inicial de soluções

Inicialmente é criada aleatoriamente uma população de α cromossomos. Cada cromossomo é formado por uma cadeia (string) binária, onde 1 (um) representa um comprimento de onda selecionado e 0 (zero), caso contrário. O parâmetro β determina o percentual de comprimentos de onda selecionados no cromossomo (percentual de bits com valor 1 no cromossomo). A posição do bit na cadeia corresponde ao comprimento de onda. Assim, o primeiro bit representa o primeiro comprimento de onda, o segundo bit da cadeia representa o segundo comprimento de onda e assim por diante. O tamanho do cromossomo (parâmetro χ) corresponde a todos os comprimentos de onda do espectro selecionado.

Passo 2: Avaliar a *fitness* dos indivíduos da população

Através da técnica de regressão por mínimos quadrados parciais (PLS) a população é avaliada. A função objetivo é:

$$\text{Min } ((SEV + SEC) + |(SEV - SEC)|) \qquad (10)$$

Passo 3a: Selecionar um conjunto de pais na população e realizar a mutação

Em seguida, os indivíduos da população são ordenados de acordo com a *fitness* produzida no passo anterior;

A população é dividida em dois grupos: Grupo A, formado pelos 50% dos indivíduos da população com melhores *fitness* e Grupo B, formado pelos 50% restantes;

Os indivíduos do Grupo A são, por sua vez, divididos em dois grupos, denominados C e D. O Grupo C corresponde aos 50% dos indivíduos melhor avaliados e o grupo D os 50% restantes.

Todos os cromossomos do grupo D são selecionados e sofrem mutação. A mutação consiste em alterar λ bits, conforme **Figura 3b**.

Passo 3b: Cruzar os pais de modo que se reproduzam

Pares de indivíduos do grupo D e do grupo C são selecionados, aleatoriamente, para sofrer a operação de *crossover*.

A operação de *crossover* é realizada produzindo um corte (em posição aleatória) num cromossomo e combinando com outro cromossomo, conforme **Figura 3c**. A operação de *crossover* é realizada repetidamente até gerar indivíduos suficientes para substituir todos os indivíduos do grupo B (passo 3d);

Passo 3c: Avaliar a *fitness* dos filhos gerados

Os indivíduos são avaliados da mesma forma realizada no passo 2, ou seja, através da equação 10.

Passo 3d: Substituir os filhos julgados inadequados

Os indivíduos gerados no passo 3b substituem os indivíduos do grupo B.

Passo 3e: Até que o critério de parada seja atendido

O procedimento retorna ao passo 3a. Quando um número máximo de iterações, δ, é atingido o programa pára.

5.2 Busca Tabu

O método heurístico Busca Tabu (BT), vem sendo aplicado com êxito a um grande número de problemas combinatoriais. A BT foi proposta por Glover (1989a, 1989b).

De modo geral, a BT foi projetada para encontrar boas aproximações para a solução ótima global de qualquer problema de otimização, tendo três princípios básicos: 1) uso de uma estrutura de dados para armazenar o histórico do processo de busca. Esta estrutura é usada para classificar algumas soluções (ou movimentos) como "tabu". Uma solução (ou movimento) "tabu" não pode ser selecionada no procedimento de busca; 2) uso de um mecanismo de controle para fazer um balanceamento entre a aceitação, ou não, de uma nova solução com base nas informações registradas no histórico de busca. O mecanismo que leva de uma solução à outra é denominado "movimento". Uma solução (ou movimento) classificada tabu, somente pode ser selecionada na busca, quando satisfazer o denominado "critério de aspiração". 3) incorporação de procedimentos que alternam as estratégias de diversificação e intensificação na busca.

As estratégias de diversificação e intensificação são cruciais para uma boa convergência da BT. A diversificação tem o objetivo de direcionar a busca para novas regiões, de forma a atingir o maior espaço de soluções possíveis, evitando que o processo estabilize num ótimo local. Já no processo de intensificação há um reforço a busca na vizinhança de uma solução historicamente considerada boa.

Basicamente, a BT parte de uma solução inicial (produzida aleatoriamente ou baseado em algum conhecimento a respeito do problema) e através de "movimentos" (perturbações aleatórias nesta solução) são obtidas soluções "vizinhas". A avaliação e comparação dos "vizinhos" são feitas e determina-se a melhor solução dentre estas. Caso a melhor solução não seja "tabu", esta solução é selecionada para reiniciar o processo, sendo que os processos de intensificação e diversificação podem controlar as perturbações. Durante todo o processo, o histórico da convergência é armazenado numa estrutura de dados. Assim, uma solução é considerada tabu, normalmente, quando já foi visitada. Esta estratégia de restrição tem por objetivo que o processo não entre em ciclos, visitando regiões já visitadas.

A **Figura 5** mostra o Algoritmo de Busca Tabu.

Passo 1: Gerar uma solução inicial
Passo 2: Repetir
Início
(a) Gerar a vizinhança
(b) Avaliar a vizinhança
(c) Determinar a melhor solução da vizinhança
(d) Atualizar a lista tabu
(e) Intensificar e Diversificar
Fim
(f) Até que o critério de parada seja atendido

Figura 5 – Algoritmo simplificado da BT implementada.

5.2.1 Aplicação da BT na Seleção de Comprimentos de Onda de Infravermelho

O funcionamento da BT, aplicada ao problema, obedece aos passos descritos na **Figura 5**:

Passo 1: Gerar uma solução inicial

Uma solução inicial é gerada aleatoriamente. A representação usada é uma cadeia (string) binária, onde 1 (um) representa um comprimento de onda selecionado e 0 (zero), caso contrário. O parâmetro ε determina o percentual de comprimentos de onda selecionados na cadeia (percentual de bits com valor 1 na cadeia). A posição do bit na cadeia corresponde ao comprimento de onda. Assim, o primeiro bit representa o primeiro comprimento de onda, o segundo bit da cadeia representa o segundo comprimento de onda e assim por diante, de forma análoga a um cromossomo, quando do uso do AG. O tamanho da cadeia (parâmetro μ) corresponde a todos os comprimentos de onda do espectro selecionado.

Passo 2a: Gerar a vizinhança

Este processo tem por objetivo encontrar γ soluções vizinhas a partir da solução corrente. Estas soluções são construídas fazendo-se movimentos, que são perturbações aleatórias de uma solução e que levam a outras soluções.

As perturbações aleatórias são controladas pelo parâmetro ϕ. Para ϕ pequenos, uma pequena quantidade de bits são alterados, conforme **Figura 6**. Neste caso, o algoritmo apresenta um processo de intensificação na busca. Para ϕ grandes, o número de bits alterados é grande e conseqüentemente o algoritmo apresenta um comportamento de diversificação da busca.

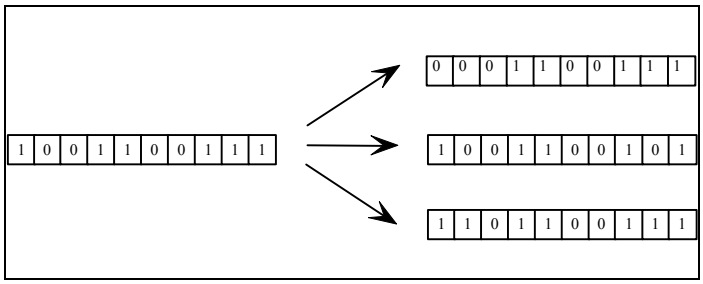

Figura 6 – Exemplo do processo de geração de vizinhança, a partir da perturbação da solução.

Passo 2b: Avaliar a vizinhança

Cada solução vizinha é avaliada através da técnica de regressão por mínimos quadrados parciais (PLS), conforme equação 10.

Passo 2c: Determinar a melhor solução da vizinhança

As soluções vizinhas são ordenadas de acordo com o valor da função objetivo obtido no passo 2b. A melhor solução, não classificada como tabu é selecionada. Caso a solução seja tabu, é verificado se a mesma satisfaz o critério de aspiração. Em caso afirmativo, a solução é selecionada e caso contrário, busca-se a próxima, melhor solução, do conjunto de soluções vizinhas.

Para que uma solução satisfaça o critério de aspiração, é necessário que o valor da sua função objetivo seja o menor obtido até aquela aquele momento.

Passo 2d: Atualizar a lista tabu

A solução selecionada no passo 2c, permanece η iterações proibida de ser selecionada, ou seja, permanece classificada como pertencente ao conjunto de soluções tabu.

Passo 2e: Intensificar e Diversificar

Estes módulos têm o objetivo de controlar o escopo da busca. Podemos entender como diversificação o aumento da perturbação (ϕ grande) realizada para gerar a vizinhança e como intensificação a diminuição deste índice (ϕ pequeno). Quanto maior a perturbação, maior a aleatoriedade do processo, quanto menor a perturbação menor a aleatoriedade. Assim, a intensificação ajuda na busca em uma região local, enquanto a diversificação permite a saída da busca para outra região. Desta forma, tenta-se evitar mínimos locais e a aproximação do mínimo global.

Passo 2f: Até que o critério de parada seja atendido

O procedimento retorna ao passo 2a. Quando um número máximo, φ, de iterações é atingido o programa pára.

6. Modelagem Quimiométrica

Foram definidas através de um planejamento experimental 22 amostras contendo o princípio ativo hidroclorotiazida em amido, das quais 17 foram usadas para calibração (34 espectros) e 5 para validação (10 espectros). O principio ativo foi manipulado de forma a obter um conjunto de amostras com diferentes concentrações, através de pesagem e diluição em amido, correspondendo as amostras padrões (com valores de referência pré-determinados). Esta variação foi de 20% acima e abaixo das concentrações comumente comercializadas, sendo a variação de 0,5% para cada amostra. O conjunto de espectros dos padrões e das amostras de validação (**Figura 7**) foram obtidos utilizando-se um acessório de reflectância difusa acoplado a um espectrômetro Nicolet Magna 550 na região do infravermelho médio.

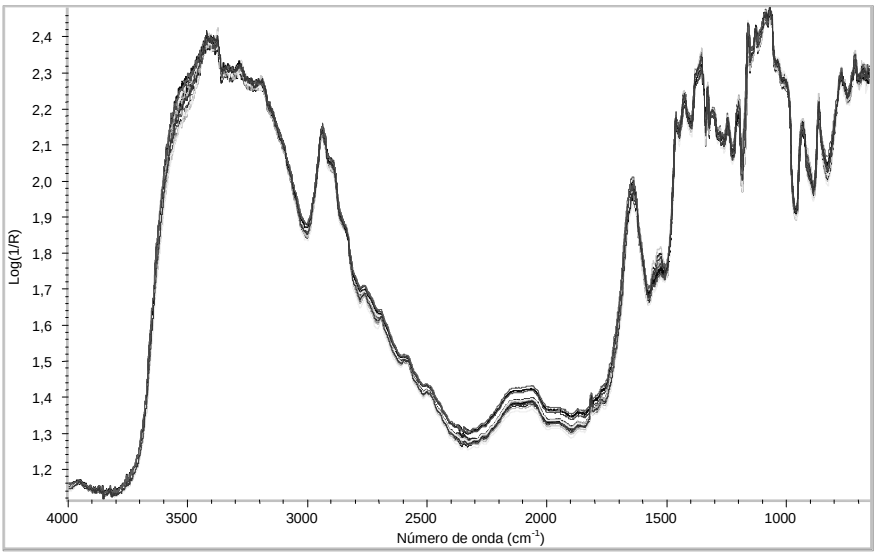

Figura 7 – Espectros das amostras de hidroclorotiazida em diferentes concentrações.

Os modelos de regressão e os algoritmos AG e BT foram implementados em ambiente MATLAB, empregando os dados brutos em forma de Log(1/R). Os programas foram executados em computador, com processador Pentium III, 800 Mhz, 256 RAM e sistema operacional Windows Me. A calibração foi realizada com o método de regressão por mínimos quadrados parciais (PLS) utilizando a correção do sinal multiplicativo (MSC) ao conjunto de espectros.

As faixas de comprimento de onda compreendendo 959-1799 e 2800-3849 cm^{-1} foram utilizadas no processo de otimização.

Na implementação do AG, foram usados os seguintes parâmetros:

$\alpha = 50$;
$\beta = 5\%$;
$\chi = 1889$;
$\lambda = 5\%$;
$\delta = 500$;

Nos testes realizados com a Busca Tabu os parâmetros a seguir foram usados:

$\varepsilon = 5\%$
$\mu = 1889$;
$\gamma = 50$;

O parâmetro ϕ inicia realizando perturbações em 50% dos bits da solução. A cada 100 iterações o valor de ϕ diminui em 50%.

$\eta = 20$;
$\varphi = 500$;

7. Resultados

Em ambos os métodos a função a ser otimizada utiliza dois parâmetros dependentes do modelo. O primeiro conhecido por SEC representa o erro de estimativa para as amostras empregadas na calibração; e o segundo, SEV, para as amostras de validação. Nas Tabelas 1 e 2 são apresentados os resultados referentes a cada geração, empregando-se um número de variáveis latentes variando de 1 a 15, SEC, SEV, R^2, o número de variáveis selecionadas (comprimentos de onda) e o tempo.

Na **Figura 8** podemos observar o comportamento de cada um destes parâmetros quando se faz variar o número de variáveis latentes (VL) empregadas na modelagem PLS. Independente do método de otimização utilizado, os valores de SEC e SEV diminuem com o aumento do número de variáveis latentes chegando a um valor mínimo para VL=10 a partir do qual o SEC e SEV oscilam entre valores compreendidos entre 0-0,5%. Por outro lado, os valores dos coeficientes de correlação convergem a 1 ($R^2 \cong 1$) quando são empregados valores crescente de VL para o PLS.

Coeficientes de correlação da ordem de 0,99 já são alcançados com VL=5, indicando que apesar de os valores de SEC e SEV ainda continuarem diminuindo com o emprego de mais variáveis latentes, para a calibração, não se verifica melhora.

Tabela 1– Resultados para os modelos PLS otimizados pela Busca Tabu.

VL	SEV	SEC	R^2	Número de variáveis selecionadas	Tempo (min)
1	3,0177	4,1387	0,5767	3	3
2	1,3835	1,3839	0,9211	22	3
3	1,1673	1,2023	0,9405	42	3
4	0,9079	0,9575	0,9622	38	2
5	0,2666	0,2677	0,9970	54	5
6	0,3283	0,3348	0,9954	80	6
7	0,1767	0,1872	0,9986	65	8
8	0,2438	0,2530	0,9974	73	6
9	0,2173	0,2319	0,9978	50	5
10	0,1143	0,1489	0,9991	56	6
11	0,1126	0,1197	0,9994	85	9
12	0,1385	0,1610	0,9989	71	7
13	0,2051	0,2374	0,9977	62	5
14	0,1196	0,1211	0,9994	67	6
15	0,0990	0,1008	0,9996	89	10

Tabela 2– Resultados para os modelos PLS otimizados pelo Algoritmo Genético.

VL	SEV	SEC	R^2	Número de variáveis selecionadas	Tempo (min)
1	2,972	4,0436	0,6242	6	7
2	1,3672	1,3669	0,9230	34	7
3	1,0399	1,0499	0,9546	44	8
4	0,6403	0,6773	0,9811	47	8
5	0,4644	0,4803	0,9905	71	9
6	0,4042	0,3946	0,9936	64	9
7	0,3128	0,3138	0,9959	56	9
8	0,2932	0,3080	0,9961	64	9
9	0,2518	0,2557	0,9973	56	9
10	0,1856	0,2190	0,9980	60	9
11	0,1674	0,2216	0,9980	64	10
12	0,2173	0,2602	0,9972	54	9
13	0,1861	0,1938	0,9985	61	10
14	0,1494	0,1521	0,9990	58	10
15	0,2418	0,2413	0,9976	68	10

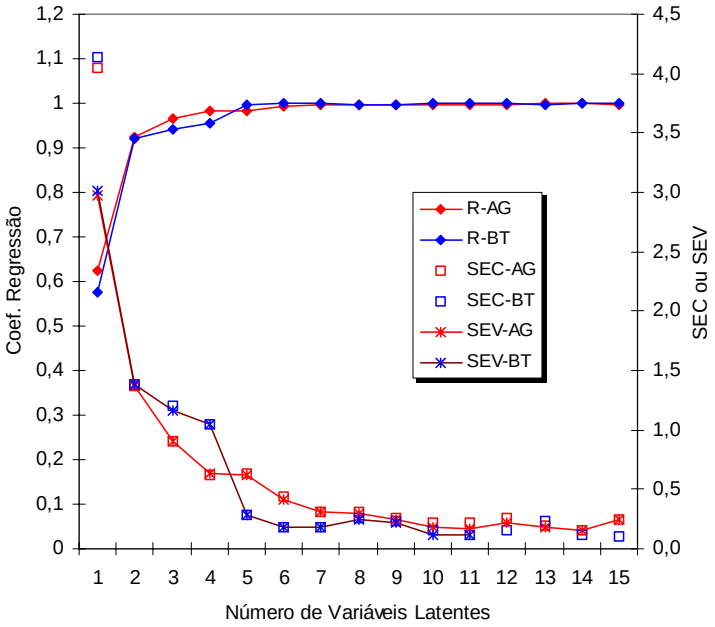

Figura 8 – Variação dos valores de SEC, SEV e R^2 em função do número de variáveis latentes empregados na modelagem PLS.

As curvas de calibração apresentadas nas **Figuras 9** e **10** ilustram bem o comportamento dos algoritmos de otimização empregados. Pode-se verificar um grande decréscimo dos valores de SEC e principalmente de SEV quando deixamos de utilizar todos os comprimentos de onda do espectro para a modelagem PLS (**Figura 9**), para serem utilizadas as combinações originadas pela aplicação dos algoritmos BT e AG (**Figura 10**). Pode-se observar que a tendência de modelar melhor o conjunto de amostras de calibração versus as de validação são minimizadas, principalmente para o algoritmo BT.

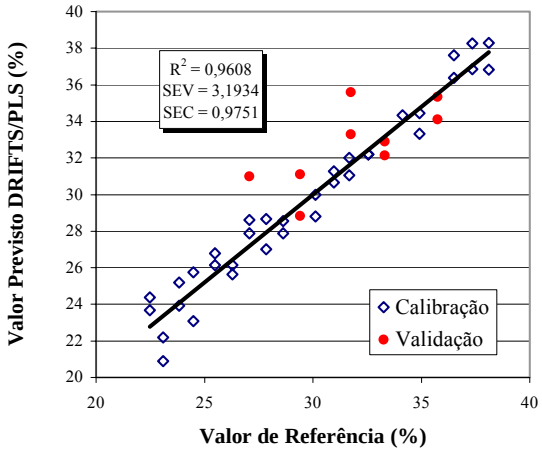

Figura 9 – Curva de calibração para o modelo que emprega todos os comprimentos de onda e 5 variáveis latentes.

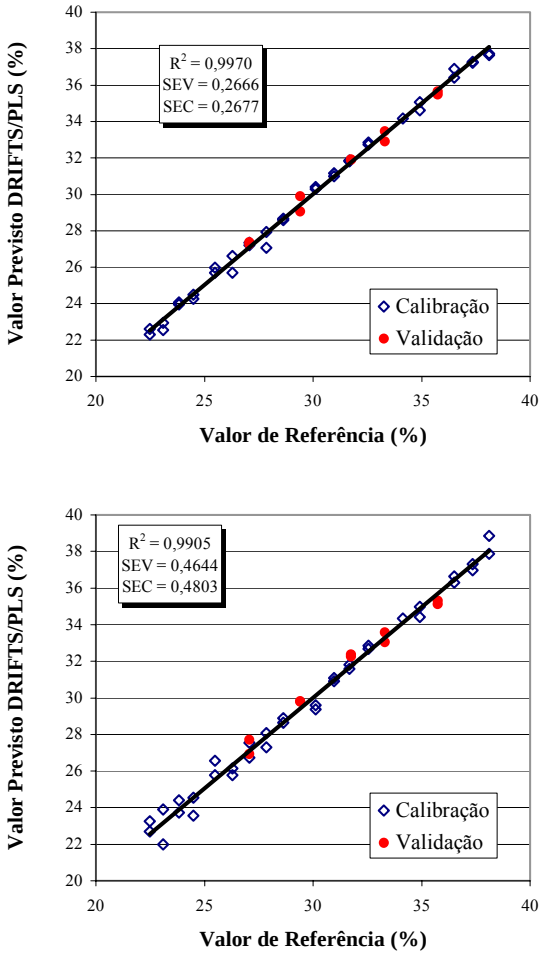

Figura 10 – Curva de calibração para as técnicas de otimização BT (primeira curva) e AG (segunda curva) empregando somente 54 e 71 comprimentos de onda, respectivamente, e 5 variáveis latentes.

8. Conclusões

Para todos os modelos otimizados empregando 4, 5 ou 6 variáveis latentes foi observada uma significativa melhora na habilidade de previsão destes, quando são selecionadas os comprimentos de onda dos espectros através dos métodos BT e AG. Este comportamento pode ser explicado, uma vez que esta redução de variáveis (seleção) deve estar sendo acompanhada pela redução do antagonismo entre as diferentes freqüências dos espectros modelados, principalmente porque são eliminadas as freqüências relacionadas com os ruídos experimentais inerentes a técnica espectroscópica empregada.

Nesta abordagem, observa-se que quando são utilizadas muitas variáveis latentes os modelos de regressão tendem a apresentar valores de SEC muito menores do que dos valores de SEV, configurando um problema conhecido por *overfitting*.

Também é observada que a melhoria da capacidade de previsão dos modelos é acompanhada pelo acréscimo no coeficiente de correlação das amostras de calibração (R^2), o que indica maior robustez nos modelos onde os métodos de otimização são empregados.

Um atrativo ao emprego da otimização utilizando técnicas meta-heurísticas reside no fato destas serem facilmente implementadas e apresentarem baixo tempo computacional, quando usadas em problemas cuja melhor solução global é difícil de ser obtida em função do elevado número de combinações possíveis.

De uma forma geral, a meta-heurística busca tabu se mostrou mais adequada na otimização do conjunto de dados em questão, seja pelos menores erros observados (SEC e SEV), ou pelo menor tempo computacional necessário para garantir soluções adequadas.

Estes resultados demonstram as potencialidades dos métodos DRIFTS/PLS/TS e DRIFTS/PLS/AG aqui propostos, como excelentes alternativas para o monitoramento do processo produtivo de empresas e farmácias que produzam ou manipulem estes princípios ativos em larga escala, apresentando baixo tempo de análise, não destruição da amostra e a não geração de resíduos.

Agradecimentos

Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul – FAPERGS.

Secretaria de Ciência e Tecnologia do Estado do Rio Grande do Sul – SCT/RS.

Referências Bibliográficas

(1) Abdullah, A.H. & Sherman, W.F. (1997). Kramers-Kronig Type Analysis of Short Spectral Range Reflection Spectra. *Vibrational Spectroscopy*, **13**, 133-142.

(2) Adams, M.J. (1995). *Chemometrics in Analytical Spectroscopy.* University of Wolverhampton, Wolverhampton.

(3) Birth, G.S. & Hecht, H.G. (1987). The Physics of Near-infrared Reflectance. **In:** *Near-infrared Technology in the Agricultural and Food Industries* [edited by Phil Wiliams & Karl Norris], American Association of Cereal Chemists, St. Paul.

(4) Coates, J. (1998). Vibrational Spectroscopy: Instrumentations for Infrared and Raman Spectroscopy. *Applied Spectroscopy Reviews*, **33**, 267-425.

(5) Costa Filho, P.A. & Poppi, R.J. (1999). Algoritmo Genético em Química. *Química Nova*, **22**, 405-411.

(6) Culler, S.R. (1993). Diffuse Reflectance Infrared Spectroscopy: Sampling Techniques for Qualitative/Quantitative Analysis of Solids. **In:** *Practical Sampling Techniques for Infrared Analysis* [edited by Patricia B. Coleman], CRC Press, Boca Raton.

(7) Durig, J.R. & Sullivan, J.F. (1990). Vibrational Spectroscopy, Fourier Transforms and Analytical Chemistry. *Trends in Analitycal Chemistry*, **9**, 104-106.

(8) Eickhoff, T.; Grosse, P. & Theiss, W. (1990). Diffuse reflectance spectroscopy of powders. *Vibrational Spectroscopy*, Elsevier Science Publishers, Amsterdam, **1**, 229-233.

(9) Eikrem, L.O. (1990). Process Fourier Transform Infrared Spectroscopy. *Trends in Analitycal Chemistry*, **9**, 107-109.

(10) Farmacopéia Brasileira. (1988). *Atheneu.* São Paulo, SP, 33-47.

(11) Ferrão, M.F. (2001). Técnicas de Reflexão no Infravermelho Aplicadas na Análise de Alimentos. *Tecno-lógica*, **5,** 65-85.

(12) Geladi, P. (2002). Some recent trends in the calibration literature. *Chemometrics and Intelligent Laboratory Systems*, **60**, 211-224.

(13) Glover, F. (1989a). Tabu Search – Part I. *ORSA Journal on Computing*, **1**, 190-206.

(14) Glover, F. (1989b). Tabu Search – Part II. *ORSA Journal on Computing*, **2**, 4-32.

(15) Hart, J.R.; Norris, K.H. & Golumbic, C. (1962). Determination of the Moisture Content of Seeds by Near-Infrared Spectrophotometry of Their Methanol Extracts. *Cereal Chemistry*, **39**, 94-99.

(16) Harvey, R.A. & Champe, P.C. (1998). *Farmacologia Ilustrada.* Artmed, Porto Alegre, RS.

(17) Hibbert, D.B. (1993). Genetic Algorithms in Chemistry. *Chemometrics and Intelligent Laboratory Systems*, **19**, 277-293.

(18) Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems.* MIT Press, Ann Arbor, Michigan.

(19) Kalasinsky, K.S. (1990). Industrial Applications of Vibrational Spectroscopy. *Trends in Analitycal Chemistry*, **9**, 83-89.

(20) Leardi, R. (2001). Genetic Algorithms in Chemometrics and Chemistry: a Review. *Journal of Chemometrics*, **15**, 559-569.

(21) Lorena, L.A.N. & Furtado, J.C. (2001). Constructive Genetic Algorithm for Clustering Problems. *Evolutionary Computation*, **9**, 309-327.

(22) Lucasius, C.B. & Kateman, G. (1994). Understanding and Using Genetic Algorithms Part 2. Representation, Configuration and Hybridization. *Chemometrics and Intelligent Laboratory Systems*, **25**, 99-145.

(23) Osborne, B.G.; Fearn, T. & Hindle, P.H. (1993). *Practical NIR Spectroscopy With Applications in Food and Beverage Analysis.* Longman Scientific & Technical, Singapure.

(24) Pimentel, M.F. & Neto, B.B. (1996). Calibração: Uma Revisão para Químicos Analíticos. *Química Nova*, **19**, 268-277.

(25) Shaffer, R.E.; Smal, G.W. & Arnold, M.A. (1996). Genetic Algorithm-Based Protocol for Coupling Digital Filtering and Partial Least-Squares Regression: Application to the Near-Infrared Analysis of Glucose in Biological Matrices. *Analytical Chemistry*, **68**, 2663-2675.

(26) Sjöström, M.; Wold, S.; Lindberg, W.; Persson, J.A. & Martens, H. (1983). A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables. *Analytical Chimica Acta*, **150**, 61-70.

(27) Van De Voort, F.R. (1992). Fourier Transform Infrared Spectroscopy Applied to Food Analysis. *Food Research International*, **25**, 397-403.

(28) Wetzel, D.L. (1983). Near-Infrared Reflectance Analysis Sleeper Among Spectroscopic Techniques. *Analytical Chemistry*, **55**, 1165A-1176A.

(29) Wilson, R.H. (1990). Fourier Transform Mid-infrared Spectroscopy for Food Analysis. *Trends in Analitycal Chemistry*, **9**, 127-131.

(30) Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, **20**, 397-405.