Universidade Federal de Pernambuco (UFPE)
Centro de Ciências da Saúde (CCS)
Departamento de Ciências Farmacêuticas (DCFAR)
Laboratório de Química Teórica Medicinal (LQTM)

Dissertação de Mestrado
no Programa de Pós-Graduação em Inovação Terapêutica
da Universidade Federal de Pernambuco
(PPGIT/UFPE)

Desenvolvimento de um software para planejamento *in silico* de fármacos: MultiMOL

Mestrando: Jorge Ferraz de Oliveira Filho
Orientador: Prof. Marcelo Zaldini Hernandes
Co-orientador: Prof. Wallace Duarte Fragoso

Recife, janeiro de 2011

UNIVERSIDADE FEDERAL DE PERNAMBUCO

Programa de Pós-Graduação em Inovação Terapêutica

JORGE FERRAZ DE OLIVEIRA FILHO

Desenvolvimento de um software para planejamento *in silico* de fármacos: MultiMOL

Recife

2011

JORGE FERRAZ DE OLIVEIRA FILHO

Desenvolvimento de um software para planejamento *in silico* de fármacos: MultiMOL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Inovação Terapêutica da Universidade Federal de Pernambuco, para a obtenção do Título de Mestre em Inovação Terapêutica

Orientador: Prof. Dr. Marcelo Zaldini Hernandes

Co-orientador: Prof. Dr. Wallace Duarte Fragoso

Recife

2011

Filho, Jorge	Desenvolvimento de um software para planejamento in silico de fármacos: MultiMOL	2,5 cm espaço reservado para etiqueta de localização	Mestrado PPGITUFPE 2011

Oliveira Filho, Jorge Ferraz
Desenvolvimento de um software para planejamento *in silico* de fármacos Multimol/ Jorge Ferraz Oliveira Filho. – Recife: O Autor, 2011.

73 folhas : il., fig., tab.
Orientador: Marcelo Zaldini Hernandes
Co-orientador: Wallace Duarte Fragoso
Dissertação (mestrado) – Universidade Federal de Pernambuco, Centro de Ciências Biológicas. InovaçãoTerapêutica, 2011.
Inclui bibliografia e anexos

1. Software 2. Fármacos 3. Modelos moleculares I. Título.

615.1 CDD (22.ed.) UFPE/CCB-2011-196

FOLHA DE APROVAÇÃO

Nome: FILHO, Jorge Ferraz de Oliveira

Título: Desenvolvimento de um software para planejamento *in silico* de fármacos: MultiMOL

Dissertação apresentada à Universidade Federal de Pernambuco para obtenção do título de Mestre em Inovação Terapêutica

Aprovada em: ___/___/____

Banca Examinadora

Prof. Dr.
Instituição:
Assinatura:_____

Prof. Dr.
Instituição:
Assinatura:_____

Prof. Dr.
Instituição:
Assinatura:_____

Prof. Dr.
Instituição:
Assinatura:_____

Ad Ecclesiam, sine Quae verae scientiae non daretur

AGRADECIMENTOS

A Deus, por tudo.

À Igreja, que sempre fez de mim um homem melhor.

Ao meu pai, à minha mãe, ao meu irmão e a toda a minha família, por terem sempre sido o alicerce sólido sobre o qual eu pude ter, com segurança, a ousadia de levantar o edifício da minha própria vida.

Ao Prof. Dr. Marcelo Zaldini, meu orientador, por todo o apoio concedido ao longo de todos os anos em que estivemos juntos. Por continuar acreditando em mim até mesmo para muito além do que era esperado - quando eu próprio já começava a duvidar.

Ao Prof. Dr. Wallace Fragoso, co-orientador deste trabalho, por toda a experiência repassada, por todas as dúvidas tiradas, por todas as energias empregadas neste projeto.

A Bruno Marques, colaborador deste projeto de Mestrado, pela enorme contribuição dada ao desenvolvimento da interface gráfica do MultiMOL.

Aos amigos que passaram pelo LQTM: Klaus, Lucas, Boaz, Sidney, Érica, Elisa, Marcelo. Pelo companheirismo manifestado a cada dia.

Ao Programa de Pós-Gradução em Inovação Terapêutica - PPGIT -, pela oportunidade de desenvolver este projeto. Pela amplitude e solidez dos conhecimentos que pude adquirir no programa. Pelos horizontes abertos. Pelos bens que não têm preço.

Vanitas

Cego, em febre a cabeça, a mão nervosa e fria,
Trabalha. A alma lhe sai da pena, alucinada,
E enche-lhe, a palpitar, a estrofe iluminada
De gritos de triunfo e gritos de agonia.

Prende a idéia fugaz; doma a rima bravia,
Trabalha... E a obra, por fim, resplandece acabada:
"Mundo, que as minhas mãos arrancaram do nada!
Filha do meu trabalho! ergue-te à luz do dia!

Cheia da minha febre e da minha alma cheia,
Arranquei-te da vida ao ádito profundo,
Arranquei-te do amor à mina ampla e secreta!

Posso agora morrer, porque vives!" E o Poeta
Pensa que vai cair, exausto, ao pé de um mundo,
E cai - vaidade humana! - ao pé de um grão de areia...

OLAVO BILAC

RESUMO

A área de Modelagem Molecular vem despertando interesse crescente desde que surgiu. O uso de métodos computacionais na previsão de estruturas e propriedades moleculares, particularmente aplicados na inovação terapêutica através do planejamento de fármacos, por exemplo, tem adquirido crescente espaço e confiabilidade na comunidade científica e na grande indústria farmacêutica mundial.

Dentre as diversas abordagens computacionais aplicáveis ao desenvolvimento de fármacos, destacam-se os estudos da relação quantitativa entre estrutura molecular e atividade biológica. Esta técnica permite construir modelos estatísticos de regressão capazes de oferecer, a partir das estruturas de moléculas, uma previsão confiável de uma propriedade de interesse, como por exemplo, a atividade biológica (QSAR), a toxicidade (QSTR) ou a solubilidade (QSPR).

O MultiMOL, desenvolvido no Laboratório de Química Teórica Medicinal (LQTM) da UFPE, é um software implementado em linguagem de programação C/C++, que oferece as técnicas de estatística multivariada mais comumente utilizadas nos problemas de QSAR. As suas funcionalidades incluem algoritmos de pré-processamento de dados (escalonamento, centrar na média, seleção de variáveis), diversos métodos de regressão (MLR, PCR, PLS e Q-PLS), a validação de modelos por validação cruzada (LOO-FCV) ou por utilização de série de testes e a exibição dos resultados dos modelos por meio de gráficos bidimensionais. Importa ressaltar que o método de PLS quadrático (Q-PLS) não é facilmente encontrado em outros softwares disponíveis, tornando-se assim um importante diferencial do MultiMOL. Todas estas funcionalidades encontram-se disponíveis para o usuário por meio de uma interface gráfica (GUI) de fácil utilização. O programa foi construído com ênfase em desempenho, robustez e precisão numérica, a fim de ser uma alternativa satisfatória como ferramenta de evidente interesse para a inovação terapêutica.

Os testes realizados, com três conjuntos de dados distintos (QSAR Tradicional, QSAR-3D e dados espectroscópicos), ofereceram indicativos satisfatórios da eficácia do software na construção de modelos de regressão. Dentre os modelos obtidos, podem ser destacados (i) PCR para QSAR Tradicional [$Q^2 = 0,70$]; (ii) PLS para QSAR-3D [$Q^2 = 0,75$]; e (iii) Q-PLS para os dados espectroscópicos [$Q^2 = 0,93$]. Estes resultados, aliados à precisão e ao bom desempenho do programa, demonstram que o MultiMOL é uma ferramenta adequada para tratar problemas típicos de estatística multivariada.

Versões futuras do software poderão incluir opções de processamento paralelo (Grid-Computing) para cálculos que exijam maiores demandas computacionais, bem como a implementação de algoritmos classificatórios (HCA, SIMCA, KNN), com o intuito de aumentar o leque de aplicabilidade do programa.

ABSTRACT

The molecular modeling area has an increasing interest since it appeared. The application of computational methods in order to predict molecular structures and properties, particularly applied in therapeutic innovation by drug design, for example, has acquired increasing space and reliability in the scientific community and big pharma industry.

Among the various computational approaches applied to drug development, it can be highlighted the study of quantitative structure-activity relationship (QSAR). This technique allows the building of statistical regression models which are capable to offer a reliable prediction of a property of interest, e.g., biological activity (QSAR), toxicity (QSTR) or solubility (QSPR).

The MultiMOL program, developed at the Laboratório de Química Teórica Medicinal (LQTM), UFPE, is implemented in the programming language C / C++, which offers the multivariate statistical techniques most commonly used in QSAR problems. The features available on it include algorithms for data preprocessing (scaling, mean-centering, variable selection), a variety of regression methods (MLR, PCR, PLS and QPLS), techniques to validate models by cross validation (LOO- FCV) or by use of series of tests, and features to display the results in a graphical way. It should be emphasized that the method of quadratic PLS (QPLS) is not easily found in other softwares, becoming so an important implementation for MultiMOL. All these features are available to the user through a graphical user interface (GUI) for easy use. The program was built with an emphasis on performance, robustness and numerical accuracy, in order to be a satisfactory alternative as a tool of obvious interest for medicinal chemistry and therapeutic innovation.

Tests performed with the program, using three different data sets (traditional QSAR, 3D-QSAR and spectroscopic data) have given satisfactory indications of its effectiveness in building regression models. Generated results were obtained with the performance and accuracy characteristics of the program. Among the models obtained, can be detached (i) PCR for traditional QSAR [$Q^2 = 0.70$], (ii) PLS for 3D-QSAR [$Q^2 = 0.75$] and (iii) Q-PLS for spectroscopic [$Q^2 = 0.93$]. These results demonstrate that MultiMOL is a suitable tool for solving typical problems of multivariate statistics, accomplishing therefore with the previously established objectives.

Future versions of software may include parallel processing options (Grid-Computing), for calculations that require greater computational demands, as well as the implementation of classification algorithms (HCA, SIMCA, KNN), in order to increase the range of applicability of the program.

Sumário:

Lista de ilustrações e tabelas

Lista de abreviaturas e siglas

APS: “Algoritmo de Projeções Sucessivas”; v. SPA.

CADD: “Computer-aided Drug Design”: Desenvolvimento de fármacos assistido por computador.

CoMFA: “Comparative Molecular Field Analysis”. Literalmente, “Análise comparativa de campos moleculares”. Nome próprio de um conhecido software de QSAR.

CP: “Componente Principal”.

GUI: “Graphical User Interface”. Interface Gráfica do Usuário.

MLR: “Multiple Linear Regression”: Regressão Linear Múltipla.

QSAR: “Quantitative Structure-Activity Relationship”: Relação Quantitativa [entre] Estrutura [molecular] e Atividade Biológica.

QSPR: “Quantitative Structure-Property Relationship”: Relação Quantitativa [entre] Estrutura [molecular] e Propriedade.

QSTR: “Quantitative Structure-Toxicity Relationship”: Relação Quantitativa [entre] Estrutura [molecular] e Toxicidade.

P&D: [Processo de] Planejamento e Desenvolvimento [de Fármacos].

PCA: “Principal Components Analysis”: Análise de Componentes Principais.

PCR: “Principal Componentes Regression”: Regressão em Componentes Principais.

PLS: “Partial Least Squares”: [Método dos] Mínimos Quadrados Parciais.

PRESS: “Predicted Residual Sum of Squares”. Soma dos Quadrados dos Resíduos de Predição.

Q-PLS: “Quadratic Partial Least Squares”: [Método dos] Mínimos Quadrados Parciais Quadráticos.

SEP: “Standard Error of Prediction”. Erro Padrão de Predição.

SPA: “Sucessive Projection Algorithm”: Algoritmo das Projeções Sucessivas.

VL: “Variável Latente”.

1. Introdução

A área de modelagem molecular vem despertando interesse crescente desde que surgiu. A sua evolução acompanha de perto a evolução das arquiteturas de computadores, cujo aumento da capacidade computacional permite o processamento, em tempo viável, de estruturas moleculares cada vez mais complexas. Com as plataformas de computadores mais modernas, já é possível calcular – com a qualidade desejada - os modelos computacionais de moléculas de interesse biológico.

O processo de desenvolvimento de fármacos [AMARAL *et. al.*, 2002] utiliza-se de métodos computacionais sofisticados para aumentar a rapidez e diminuir os custos inerentes ao planejamento de fármacos e medicamentos. Um conjunto de resultados obtidos por meio destas técnicas [ALLEN *et. al.*, 1987; BERNARD *et. al.*, 2001; HOPFINGER, 1985; FALCÃO, 2009] têm comprovado a sua viabilidade e confiabilidade, e contribuído para a sua popularização. Os chamados métodos de planejamento *in silico* de fármacos, assim, têm se mostrado uma alternativa adequada para a identificação de novos compostos protótipos, ou até mesmo para o melhoramento ou otimização de moléculas já conhecidas.

A utilização de computadores na área de desenvolvimento de fármacos tem, portanto, tomado proporções significativas [HOPFINGER, 1985; COHEN *et. al.*, 1990] , onde os formalismos teóricos como, por exemplo, química quântica e mecânica molecular, têm sido aplicados rotineiramente nas abordagens teóricas necessárias para o estudo e a previsão de estruturas e propriedades moleculares, principalmente. A incorporação de tecnologias de planejamento de fármacos assistido por computador (CADD – “Computer Aided Drug Design”) às abordagens de P&D pode levar a redução de até 50% dos custos de desenvolvimento de um fármaco [FDA, 2009; MCGEE, 2005]. Entre os motivos que justificam o crescente interesse da comunidade científica aos métodos *in silico* de desenvolvimento de fármacos pode-se destacar a facilidade crescente de acesso a arquiteturas computacionais poderosas o suficiente para abordar sistemas mais complexos, os bons resultados que têm sido obtidos na área, e a comodidade em se desenvolver novos compostos com uma fase prévia de modelagem molecular, precedente à obtenção experimental efetiva de novas moléculas, por exemplo, através de síntese orgânica.

O interesse associado à aplicação destes métodos tem, assim, estimulado fortemente o desenvolvimento de modelos e metodologias que possibilitem a descrição fiel de propriedades destes sistemas, além da previsão e eventual confirmação de novas informações sobre os compostos estudados. Dentre as diversas abordagens computacionais existentes que podem ser aplicadas ao problema de desenvolvimentos de fármacos, destaca-se o estudo da relação quantitativa entre estrutura e atividade [AMARAL *et. al.*, 2002; HANSCH *et. al.*, 1962; HANSCH *et. al.*, 1964; HANSCH, 1969], o qual pode ser “tradicional”, tipicamente denominado QSAR (“Quantitative Structure-

Activity Relationship") ou em três dimensões (QSAR-3D) [KUBINYI, 1997; WANG *et. al.*, 2003]. Estas abordagens têm sido utilizadas em larga escala nos mais diversos estudos de química medicinal encontrados na literatura. Com estas técnicas, é possível construir modelos de regressão que oferecem uma previsão confiável de uma certa propriedade de interesse (como por exemplo, a atividade biológica, a toxicidade, a lipossolubilidade, etc.) de uma molécula a partir somente de sua estrutura molecular. Deste modo, é possível selecionar os compostos com maior potencial de resposta biológica para serem sintetizados e testados (*in vivo* ou *in vitro)*, aumentando assim as chances de aumentar a rapidez e diminuir os custos do processo de desenvolvimento de novas moléculas bioativas. Tais técnicas vêm sendo cada vez mais amplamente utilizadas, justamente por causa da sua grande aplicabilidade e pela capacidade destes modelos em auxiliar nas decisões de modificação molecular necessárias para a potencialização dos efeitos farmacológicos das moléculas de interesse.

É importante salientar que, paralelamente aos modelos de *QSAR* ("Quantitative Structure-Activity Relationship"), também existem os modelos de *QSTR* ("Quantitative Structure-Toxicity Relationship") e *QSPR* ("Quantitative Structure-Property Relationship"). A diferença fundamental entre eles consiste apenas na função resposta utilizada nos modelos de regressão que, no caso de *QSAR*, é a atividade biológica ou farmacológica, em *QSTR*, é a toxicidade das moléculas e, em *QSPR*, pode ser qualquer propriedade físico-química, como, por exemplo, a solubilidade, ou uma propriedade farmacocinética qualquer. Desta forma, estes métodos são matemática ou estatisticamente equivalentes no que diz respeito à tentativa de correlação de descritores moleculares com Atividade, Toxicidade ou uma Propriedade. Assim sendo, ao longo deste texto, será utilizada muito freqüentemente a sigla *QSAR*, mas o programa MultiMOL possui generalidade suficiente para também ser utilizado para modelos *QSTR* e *QSPR*. Portanto, na grande maioria dos casos onde a sigla *QSAR* aparece, poderia ser naturalmente substituída por *QSTR* e *QSPR*, neste texto.

O maior desafio para a abordagem de QSAR Tradicional é a escolha das variáveis selecionadas para a construção do modelo de regressão, ou seja, os descritores moleculares. São comumente utilizadas propriedades físico-químicas, como coeficiente de partição, contagens de determinados tipos de átomos na estrutura, cargas sobre átomos específicos, eletronegatividade, e diversas outras propriedades topológicas, eletrônicas e empíricas. Preferencialmente, estes descritores selecionados guardam relação com a jornada do fármaco no corpo humano até o seu alvo biológico (farmacocinética). Verifica-se, no entanto, que é difícil estabelecer, para o caso geral, uma regra universalmente válida para a seleção das variáveis (descritores) relevantes, sendo geralmente a escolha feita em cada estudo específico.

Adicionalmente, muitas das abordagens atuais para o desenvolvimento ou a busca de moléculas bioativas utilizam modelos de *QSAR*-3D [GOLBRAIKH *et. al.*, 2000; HOU *et. al.*, 2001]. Estes métodos procuram estabelecer uma correlação estatística significativa entre dados de atividades biológicas experimentais e variáveis estruturais,

através da distribuição geométrica, no espaço tridimensional (3D), de propriedades associadas com eventos de reconhecimento molecular. A idéia fundamental por trás destas metodologias é que as diferenças nas propriedades moleculares dos compostos podem ser freqüentemente associadas às diferenças presentes em campos não-covalentes ("Campos Moleculares") ao redor das moléculas de interesse. Para a aplicação desta metodologia, as moléculas de interesse são colocadas dentro de uma "caixa" tridimensional discretizada. Em seguida, estas moléculas precisam ser alinhadas, a fim de garantir que o valor do campo molecular calculado em cada um dos pontos da caixa estará adequado ao campo correspondente em cada uma das moléculas da série. Por fim, por meio de uma sonda capaz de calcular o campo (p.ex., eletrostático ou estérico) criado por aquela determinada estrutura molecular, o seu valor em cada um dos pontos da "caixa" é registrado para todas as moléculas em estudo. Em outras palavras: a partir da estrutura tridimensional da série homóloga de moléculas a serem estudadas, são calculados os campos moleculares para cada um dos pontos espacialmente localizados na vizinhança da molécula, os quais são, então, utilizados como descritores para a aplicação de metodologias de identificação de relações entre a estrutura molecular e a atividade biológica (QSAR).

Existem alguns programas comerciais e acadêmicos que são utilizados para estes procedimentos, entre os quais merecem destaque o CoMFA ("Comparative Molecular Field Analysis") [CRAMER *et. al.*, 1988] e o GRID [GOODFORD, 1985].

Estes são baseados na utilização de sondas ("probes") para a geração de campos moleculares dos tipos eletrostático e estérico, calculados na intersecção da malha ou grade discretizada de pontos com a região tridimensional ao redor da molécula. Desta forma, cada descritor tridimensional é representado por valores de campo eletrostático ou estérico, em um conjunto de pontos da malha 3D e funcionam como variáveis independentes de uma análise QSAR usando-se técnicas quimiométricas [SHARAF *et. al.*, 1986], como por exemplo, regressão linear múltipla (*MLR* – "Multiple Linear Regression"), mínimos quadrados parciais (*PLS* – "Partial Least Squares"), regressão em componentes principais (*PCR* – "Principal Component Regression") ou análise de componentes principais (*PCA* – "Principal Components Analysis"). Para a sonda eletrostática, por exemplo, este procedimento funciona como um mapeamento gradual das mudanças nas propriedades de interação intermolecular, dada pelo cálculo da energia potencial em um malha de pontos regularmente espaçados que envolvem as moléculas de interesse. Isto funciona na prática como uma simulação do processo de interação molecular entre o fármaco e o seu alvo biológico, na qual o caminho percorrido pela sonda mede os campos moleculares da interação da molécula com o seu receptor. Naturalmente, o alinhamento das moléculas, baseado em critérios estruturais ou critérios de superposição de campos, é um requisito necessário e bastante importante para a obtenção do modelo QSAR-3D.

Note-se a importância, nestas abordagens, da utilização de uma série homóloga de moléculas (i.e., que se caracterizem como variações de substituintes sobre um

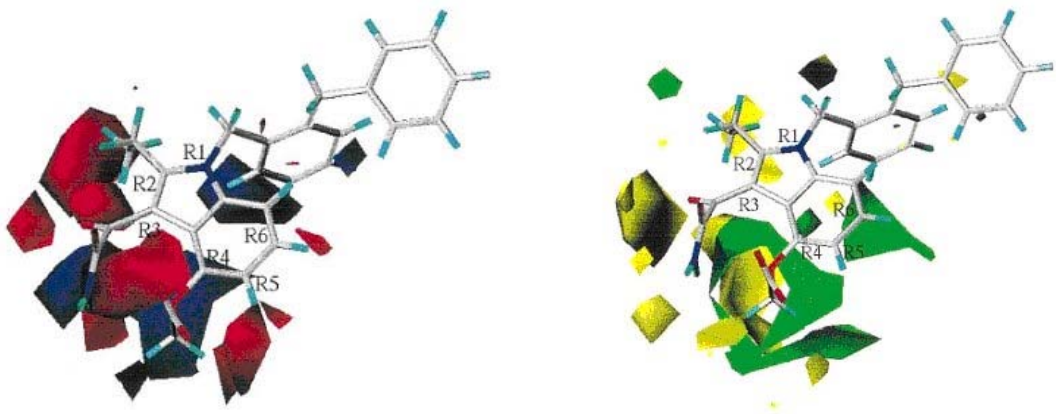
mesmo grupo farmacofórico, comum a todas), particularmente uma série que apresente o mesmo mecanismo de ação. Sendo homólogas, o alinhamento das moléculas também tende a ser mais simples e preciso, uma vez as diferenças entre as moléculas da série são periféricas, de substituintes, e não do grupo farmacofórico comum a todas.

Em outras palavras, em uma amostragem apropriada dos campos moleculares ao redor do conjunto de moléculas alinhadas, espera-se que esteja disponível a informação adequada para uma melhor compreensão da justificativa molecular da sua atividade biológica, principalmente nos casos onde a estrutura tridimensional do sítio ativo não é conhecida. Esta amostragem apropriada é conseguida pelos cálculos de energia de interação entre cada molécula e uma sonda apropriada que percorre a malha tridimensional regularmente espaçada ao redor das moléculas de interesse alinhadas. A energia resultante é calculada a partir de funções de potencial ("campos de força") simples e já bem estabelecidas na literatura, como, por exemplo, o potencial de Coulomb e de Lennard-Jones [ALLEN *et. al.*, 1987].

O método QSAR-3D também pode ser utilizado para estudar diferenças fundamentais entre grupos específicos de enzimas, como foi feito por Kastenholz e colaboradores [KASTENHOLZ *et. al.*, 2000]. Neste trabalho, os autores realizaram uma Análise de Componentes Principais Consensual sobre os descritores tridimensionais obtidos através da varredura de determinadas sondas sobre serina proteases homólogas tipo trombina, tripsina e fator Xa. É importante salientar que as regiões identificadas através desta metodologia, que têm uma importante "capacidade de localização" (i.e., possibilidade de identificação dos pontos relevantes dentro da estrutura tridimensional da molécula), em muitos casos são importantes para o aumento da seletividade das moléculas que apresentam atividade contra tais enzimas. Isto acontece porque os ligantes investigados podem sofrer modificações estruturais que permitam aumentar a afinidade para uma classe de enzimas e ao mesmo tempo diminuir para outra, ou seja, direcionar a atividade de tal molécula. Este procedimento torna-se, portanto, uma ferramenta extremamente útil para combater problemas de efeitos colaterais, como toxicidade, onde, muitas vezes, o problema básico por trás do efeito tóxico é o da necessidade de aumento da seletividade de uma determinada droga para um tipo específico de enzima.

Pintore, Bernard e colaboradores [PINTORE *et. al.*, 2001] e [BERNARD *et. al.*, 2001] utilizaram o programa *CoMFA* para estudar uma série de derivados de indol com atividade inibidora reversível de fosfolipase pancreática humana, tentando estabelecer seu potencial como drogas para o tratamento de doenças inflamatórias. Os autores puderam estabelecer os critérios moleculares importantes correlacionados com a potencialidade das drogas, através do cálculo e da visualização dos campos estérico e eletrostático, que são representados na Figura 1, em conjunto com a molécula que apresentou a maior atividade neste estudo. De uma maneira geral, os autores chegaram a conclusão que o potencial inibitório das drogas desta natureza poderia ser

aumentado empregando-se grupos volumosos na posição “R5”, grupos carregados positivamente na posição “R6”, e grupos carregados negativamente na posição “R2”.



(a) **(b)**

Figura 1: (a) Representação do campo molecular eletrostático calculado com o programa *CoMFA*. A capacidade inibitória da droga é potencializada aumentando cargas negativas dentro da região representada em vermelho ou cargas positivas dentro da região em azul. (b) Representação do campo molecular estérico calculado com o programa *CoMFA*. A capacidade inibitória da droga é potencializada aumentando o impedimento estérico dentro da região representada em verde ou diminuindo-o dentro da região em amarelo.

Hou e colaboradores [HOU *et al.*, 2001] também usaram o programa *CoMFA* para analisar uma série de cinamamidas com atividade anticonvulsiva. Pode-se notar novamente que a ordem de atividade destas drogas é explicada a partir do preenchimento dos critérios moleculares descritos pelos diversos campos calculados com o programa *CoMFA*. Além disto, espera-se que este estudo forneça informação útil suficiente para o planejamento de drogas ainda mais potentes.

Isto posto, pode-se destacar, de uma maneira geral, a necessidade de desenvolvimento e implementação de metodologias teóricas próprias (*softwares*), implementadas com tecnologias nacionais, para o estabelecimento de modelos de *QSAR*, uma vez que estes modelos têm sido utilizados em larga escala nos mais diversos estudos de química medicinal encontrados na literatura, justamente por causa da sua grande aplicabilidade e pela capacidade destes modelos em auxiliar nas decisões de modificação molecular necessárias para a potencialização dos efeitos farmacológicos das moléculas de interesse.

O grupo de pesquisa em Modelagem para Inovação Molecular (MODiMOL) tem desenvolvido um pacote de programas denominado CAMOL, constituído por 4 (quatro) ferramentas fundamentais e voltado especificamente para a construção de modelos de relação estrutura-atividade utilizados em planejamento de fármacos. O programa MultiMOL, alvo deste projeto de mestrado, constitui uma das importantes ferramentas do pacote CAMOL. Apesar de ser parte integrante do CAMOL, o programa MultiMOL é

um software independente e com funcionalidades próprias, que podem, inclusive, serem utilizadas para outras finalidades correlacionadas, que particularmente envolvam análises estatísticas multivariadas.

Neste sentido, a proposta de desenvolvimento do programa MultiMOL é a de fornecer uma alternativa viável para a aplicação das técnicas de estatística multivariada mais utilizadas nos problemas de QSAR, oferecendo uma implementação precisa, robusta e com boa performance, de modo transparente e flexível para o usuário.

Desenvolvido no Laboratório de Química Teórica Medicinal (LQTM) do Departamento de Ciências Farmacêuticas (DCFar) da Universidade Federal de Pernambuco (UFPE), o MultiMOL apresenta-se assim como uma alternativa aos *softwares* tradicionais de estatística multivariada, situando-se em uma área de interesse crescente para a química medicinal e de grande importância para a Inovação Terapêutica.

2. Objetivos

2.1. Objetivo Geral

Este projeto tem o seu principal objetivo voltado para o desenvolvimento e a implementação do programa ("software") MultiMOL, que pode ser aplicado para a geração de modelos de regressão multivariada tipicamente presentes na área de inovação em planejamento de fármacos.

2.2. Objetivos Específicos

Os objetivos específicos deste projeto englobam o desenvolvimento e aprimoramento, através da implementação de novas características e funcionalidades, do *software* denominado MultiMOL, que é usado na geração de modelos de regressão tipicamente utilizados em química medicinal, para o planejamento de fármacos, contribuindo, desta forma, para a inovação terapêutica.

Os seguintes objetivos específicos fazem parte deste projeto de mestrado:

1. a implementação do algoritmo de regressão PLS, e sua comparação com os demais algoritmos do MultiMOL;
2. a implementação do algoritmo de regressão Q-PLS, utilizado para conjuntos de dados que possuam dependências não lineares, e a sua validação;

3. o desenvolvimento da Interface Gráfica do Usuário (GUI), integrada à parte numérica do MultiMOL, para que possa facilitar a utilização do programa;
4. teste das metodologias implementadas no MultiMOL com conjuntos de dados estatísticos multivariados, tanto na área de química medicinal, quanto na área de química analítica.

3. Metodologia

3.1. Metodologia geral

O presente projeto concentra esforços em novas e importantes implementações que foram realizadas no programa MultiMOL. Um dos pontos principais é o desenvolvimento de uma interface gráfica de usuário (*GUI* – "Graphical User Interface") para o software, de modo a aumentar a sua *usabilidade* (facilidade de interação entre o usuário e o software), tornando-o mais amigável e prático para o usuário final. Além disso, este projeto também pretende desenvolver e implementar novas funcionalidades úteis para a geração de modelos de regressão como, por exemplo, a disponibilização de gráficos bidimensionais informativos (usando a GUI) que possibilitem a interpretação visual e interativa dos resultados obtidos por meio dos cálculos com os modelos de regressão. Para desenvolver a *GUI* do MultiMOL, foi utilizada a biblioteca QT [TROLLTECH, 2008], cuja licença de uso é livre. Espera-se, desta forma, aumentar o potencial de aplicabilidade do programa nos diversos problemas de química medicinal e química analítica.

A versão do QT utilizada na implementação da interface gráfica do programa foi a 4.5. O QT é um *framework* (um conjunto de componentes de software – no caso, de componentes gráficos – que podem ser combinados para a construção de aplicações) desenvolvido em C/C++, o que facilita a sua integração com o restante do código do MultiMOL, uma vez que este vem sendo desenvolvido na mesma linguagem. Para a geração dos gráficos bi-dimensionais foi utilizada uma biblioteca chamada Qwt, versão 5.1.2 [QWT, 2009], completamente integrada ao QT e que oferece os componentes necessários à geração dos gráficos que o MultiMOL se propõe a disponibilizar em sua GUI. Com estas ferramentas, foi possível construir a interface gráfica do programa com as funcionalidades planejadas, utilizando-se uma solução *open source* (código aberto) de simples compatibilidade com a parte numérica do MultiMOL.

Por fim, é necessário buscar continuamente conjuntos de dados utilizados em pesquisas na área de química medicinal ou inovação terapêutica, para os quais o programa possa ser útil e com os quais ele possa ser testado. Neste sentido, serão

utilizados conjuntos de dados disponíveis em bases existentes na internet, como, por exemplo, a “cheminformatics” [CHEMINFORMATICS, 2008], que disponibiliza conjuntos de dados (*QSAR*, *QSTR* e *QSPR*) interessantes e validados para a obtenção de modelos de regressão, com enfoque especial em química medicinal.

3.2. Metodologia específica

Para o desenvolvimento do software MultiMOL, optou-se por usar a linguagem de programação C/C++. Isso foi feito porque esta linguagem possibilita o desenvolvimento de aplicativos robustos, de bom desempenho e com todas as facilidades que uma linguagem de alto nível oferece aos programadores. Com ela, é possível obter todas as vantagens inerentes ao paradigma de orientação a objetos, no qual a modelagem do problema é feita considerando-se como objetos os diversos componentes do sistema. Isto possibilita reuso de código, otimização de recursos, independência entre os módulos do sistema, entre outras características. O alto desempenho é inerente às linguagens de programação que utilizam código compilado, como C/C++ (i.e., transformado em operações computacionais que são específicas para a plataforma onde o código está sendo executado), ao contrário do código interpretado, como é o caso, por exemplo, do Matlab ou Scilab, onde existe um interpretador responsável por traduzir, em tempo de execução, o código escrito para as operações que são efetivamente realizadas pelo computador – o que, desta forma, provoca uma perda de desempenho. Foi escolhida a representação numérica computacional por meio de pontos flutuantes de precisão dupla ("double") ao invés de simples ("float"), com o objetivo de garantir a precisão numérica dos resultados gerados pelo software. Espera-se que o MultiMOL seja capaz de oferecer as ferramentas comumente usadas para regressão multivariada, em uma variedade que lhe confira versatilidade, e a um custo computacional acessível.

A etapa mais importante na construção de um modelo de *QSAR* (tradicional ou 3D) consiste, basicamente, na busca por uma correlação estatística (ver Figura 2) entre um determinado conjunto de características ou descritores moleculares (variáveis independentes) calculadas e uma função resposta (variável dependente) referente à atividade biológica (ou propriedade físico-química, ou toxicidade) observada experimentalmente para uma classe homóloga de moléculas. Do ponto de vista matemático, trata-se de um método de regressão multivariado onde uma seqüência de operações matriciais é realizada com o objetivo de se construir uma função que associe os descritores moleculares calculados à função resposta observada ou medida experimentalmente. A metodologia de *QSAR* aplicada no MultiMOL encontra-se fartamente descrita na literatura [BRERETON, 2000; FERREIRA *et. al.*, 1999; GAUDIO *et al*., 2001; ARAUJO *et. al.*, 2001] .

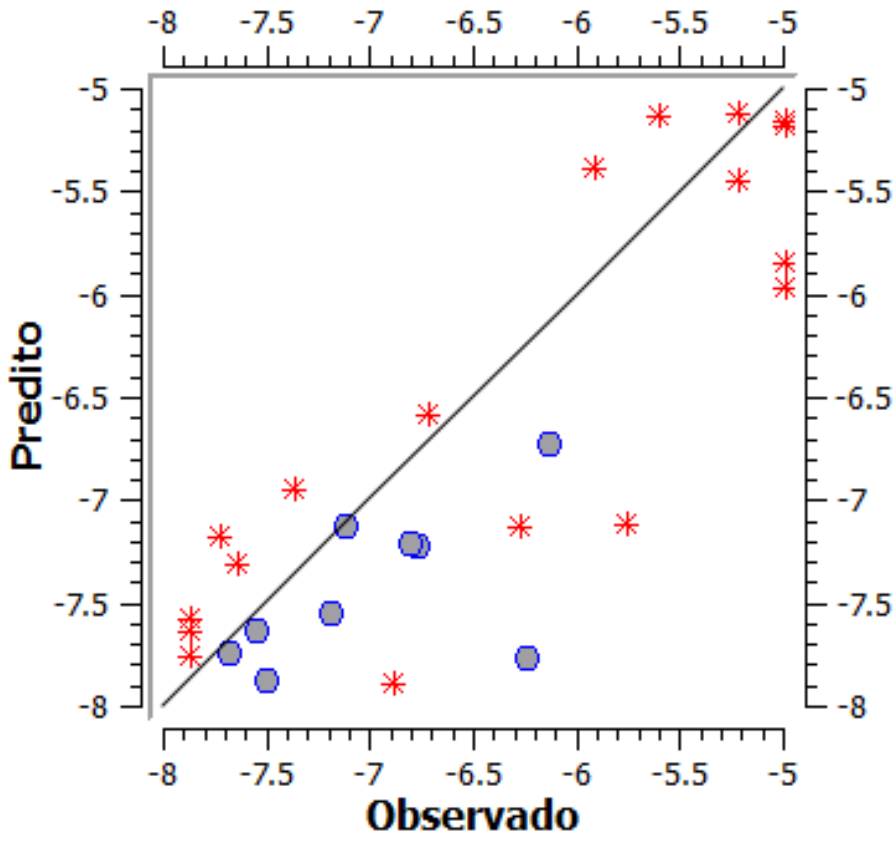


Figura 2: Exemplo de modelo *QSAR* obtido com o programa MultiMOL. Correlação entre a atividade (pk de 31 esteróides [WAGENER *et al.*, 1995] frente à globulina humana carregadora de corticosteróide) predita pelo modelo e a observada experimentalmente. Os pontos em vermelho indicam as moléculas que foram utilizadas como conjunto de calibração do modelo, enquanto os pontos em azul representam as amostras de validação ou teste. A linha representa a diagonal, apenas para facilitar a interpretação do gráfico.

Para o cálculo das operações matriciais elementares utilizadas nos algoritmos de regressão multivariada (tais como multiplicação e inversão de matrizes), foi feita a opção por uma *biblioteca* (conjunto de algoritmos de computador) pública, gratuita, com código aberto e com implementação já bem estável em C++, denominada NEWMAT [DAVIES, 1996; EDDELBUTTEL, 1996]. Essa decisão foi motivada pelo fato desta biblioteca já se encontrar otimizada e ter máximo desempenho já testado para cálculos matriciais. Ainda, o fato de ela ser escrita na mesma linguagem em que o MultiMOL foi desenvolvido (C++) possibilita uma integração simples entre os códigos sem maiores problemas de compatibilidade. Por fim, o fato de ser uma biblioteca de código aberto

(*open source*) permite que ela seja utilizada no MultiMOL sem nenhum problema de integração ou de direitos autorais.

3.2.1. Construção de Modelos de Regressão

A construção dos modelos de regressão compreende duas etapas distintas: a calibração (ou modelagem) e a validação. Conforme Ferreira e colaboradores [FERREIRA *et al.*, 1999], "[o] processo geral de calibração consiste de duas etapas: MODELAGEM, que estabelece uma relação matemática entre X e Y no conjunto de calibração e a VALIDAÇÃO, que otimiza a relação no sentido de uma melhor descrição do(s) analito(s) de interesse". Vale salientar que, para a química analítica, a otimização da relação está também compreendida na etapa de modelagem.

Na modelagem, são construídos diversos modelos de regressão para as amostras que fazem parte do conjunto de calibração. Para a construção destes modelos, é realizada uma busca sistemática no espaço das variáveis (p.ex., no caso do PLS, são construídos modelos com uma variável latente, duas variáveis latentes, e assim sucessivamente), após a qual os modelos construídos são comparados entre si, a fim de que seja identificado aquele que apresenta melhores resultados – ou seja, para a identificação do número ótimo de Variáveis Latentes. Para tanto, na modelagem, é necessário que seja feita a análise estatística dos modelos gerados, a fim de se verificar a qualidade deles e garantir a existência de correlação estatisticamente significativa, excluindo correlação casual. Com este intuito foram então implementados os cálculos de (a) *PRESS* ("Predicted Residual Sum of Squares") e (b) R^2 (coeficiente de determinação) conforme [BRERETON, 2000].

O primeiro é definido como a soma da diferença quadrática entre o valor da atividade biológica observada experimentalmente (e utilizado na construção do modelo) e o valor predito pelo modelo. Já o segundo é obtido quando se divide o valor de *PRESS* pelo quadrado dos desvios dos valores observados em torno a média, e este quociente é subtraído de 1. O R^2 consiste em um valor entre 0 e 1, sendo tanto mais significativo o modelo quanto mais próximo da unidade este valor esteja. As equações 1 e 2 mostram a forma como estes valores são calculados. Nestas equações, $y_i obs$ representa o valor experimental (observado) para a i-ésima amostra; $y_i cal$ é o valor calculado pelo modelo para a i-ésima amostra; e ymed é a média dos valores experimentais. A diferença entre o valor calculado e o predito é que o primeiro obtém-se quando se projeta no modelo uma amostra que foi utilizada para a sua construção e, o segundo, quando a amostra projetada não foi utilizada para a sua calibração: por exemplo, a amostra deixada de fora na LOO-FCV ("Leave-One-Out Full-Cross-Validation") ou as amostras de um conjunto de validação com série de teste.

$$PRESS = \sum (y_i obs - y_i cal)^2 \qquad \textbf{(equação 1)}$$

$$R^2 = 1 - [\sum (y_i obs - y_i cal)^2 / \sum (y_i obs - ymed)^2] \qquad \textbf{(equação 2)}$$

Foi igualmente implementado o cálculo do coeficiente de correlação da validação cruzada (Q^2). Este coeficiente tem significado idêntico ao R^2, mantendo a diferença de que os dados aqui obtidos são provenientes da predição das amostras que foram deixadas de fora durante o processo de calibração dos modelos, e não após a construção do modelo final. Na equação 3, pode ser observada a forma como o Q^2 é calculado; nesta equação, $y_i pred$ é o valor predito pelo modelo para a i-ésima amostra.

$$Q^2 = 1 - [\sum (y_i obs - y_i pred)^2 / \sum (y_i obs - ymed)^2] \qquad \textbf{(equação 3)}$$

A validação dos modelos de *QSAR,* seguindo a tradição encontrada na literatura, é feita de duas maneiras no MultiMOL: através da validação cruzada (conjunto interno de validação ou teste) ou da utilização de um conjunto externo de validação ou teste (série de teste). A opção por uma ou outra maneira vai depender, principalmente, da quantidade dos dados experimentais disponíveis para a construção dos modelos de regressão, ou seja, do número de amostras.

A prática de utilização dos modelos de regressão demonstra que sempre há o risco de que a correlação obtida entre os dados originais e sua função-resposta possa ser casual, i.e., não possuir significado físico verdadeiro. Como os formalismos matemáticos são aplicados sobre números, quaisquer que sejam eles, é forçoso que as metodologias de regressão forneçam sempre algum resultado, independente de haver ou não sentido físico aplicável ao problema concreto analisado. Verifica-se, assim, a necessidade de que, mesmo após obtido o melhor modelo de regressão para um determinado conjunto de dados, seja comprovado se a modelagem estatística está efetivamente condizente com a realidade experimental ou se, ao contrário, é meramente uma correlação numérica casual. Isto foi verificado, neste trabalho, através da utilização do teste F (F-Test) [GAUDIO *et al.*, 2001].

Este teste é utilizado na avaliação de modelos de regressão. Segundo Gaudio [GAUDIO *et al.*, 2001], "[o] teste F verifica o quanto da variabilidade de Y pode ser explicada pelas variáveis X_1, X_2, …, X_k, e o quanto pode ser atribuída ao efeito do erro aleatório". Este teste simples possibilita a obtenção de um valor que, quanto maior, indica que mais provavelmente a correlação encontrada não é fortuita. O teste F é definido como uma proporção entre a variância explicada pelo modelo de regressão e a variância inexplicada pelos erros aleatórios; o seu resultado é então comparado com uma tabela de distribuição estatística [THE F-DISTRIBUTION, 2009], para avaliar o grau de significância da correlação mensurada. A fim de fornecer este indicador quantitativo da qualidade do modelo gerado, o MultiMOL implementa o teste F tanto

para a calibração dos modelos quanto para a sua validação com séries de testes (quando for o caso).

3.2.2. Escolha dos conjuntos de testes

Para a validação dos algoritmos implementados no MultiMOL e a realização dos testes cujos resultados serão apresentados na seção seguinte, foram selecionados alguns conjuntos de dados da literatura que pudessem fornecer a base para um estudo comparativo adequado e confiável. Foram selecionados três conjuntos de dados, de tal maneira que cobrissem a maior parte das aplicações do programa, possibilitando testes representativos das principais funcionalidades do MultiMOL com a ênfase em robustez e desempenho que é proposta por este projeto. Estes conjuntos doravante serão denominados “Conjunto A”, “Conjunto B” e “Conjunto C”, conforme a Tabela 1.

Para os modelos de QSAR tradicional, foi utilizado um conjunto de inibidores de enzimas que possuem função conversora de Angiotensina (inibidores ACE, pIC50) estudados por Sutherland [SUTHERLAND *et al.*, 2004]. Este conjunto possui descritores que foram chamados de “2.5D”, i.e., descritores tradicionais de QSAR acrescidos de alguns descritores tridimensionais, como volume molecular, por exemplo. Estes descritores não devem ser confundidos com os descritores utilizados para problemas de QSAR-3D, conforme explicado abaixo.

Tabela 1: Conjuntos que foram utilizados nos testes do MultiMOL.

Conjunto	Breve Descrição	#Amostras	#Descritores	Aplicabilidade
“Conjunto A”	Série homóloga de inibidores de enzimas que têm função conversora de Angiotensina.	114	56	Usado para os testes de QSAR tradicional.
“Conjunto B”	Série homóloga de esteróides, com afinidade para globulina ligadora de corticosteróide.	31	1813	Usado para os testes de QSAR-3D.
“Conjunto C”	Conjunto de espectros simulado, gerados por meio da aplicação de funções numéricas quadráticas.	75	300	Usado para a validação do algoritmo Q-PLS.

Para os modelos de QSAR-3D, foi utilizado um conjunto bem conhecido de esteróides [WAGENER *et al.*, 1995], para gerar um modelo para a função-resposta "CBG affinity" (afinidade pela globulina ligadora de corticosteróide). Os descritores utilizados são os valores dos campos eletrostáticos calculados para cada um dos pontos da grade tridimensional em cujo interior as estruturas tridimensionais das moléculas foram colocadas. A figura 3 mostra esse conjunto de esteróides.

Para a validação do algoritmo Q-PLS, foi necessária a utilização de um terceiro conjunto de dados, de espectros simulados, onde as dependências não lineares da resposta (variável dependente) foram embutidas de maneira simulada. Os detalhes específicos deste conjunto de dados podem ser observados no Anexo I. A função-resposta foi construída por meio da aplicação de funções quadráticas nas variáveis independentes. O comportamento deste conjunto é bastante específico e conhecido *a priori*, o que permite uma avaliação robusta do modelo quadrático para ele construído, como será apresentado na seção seguinte.

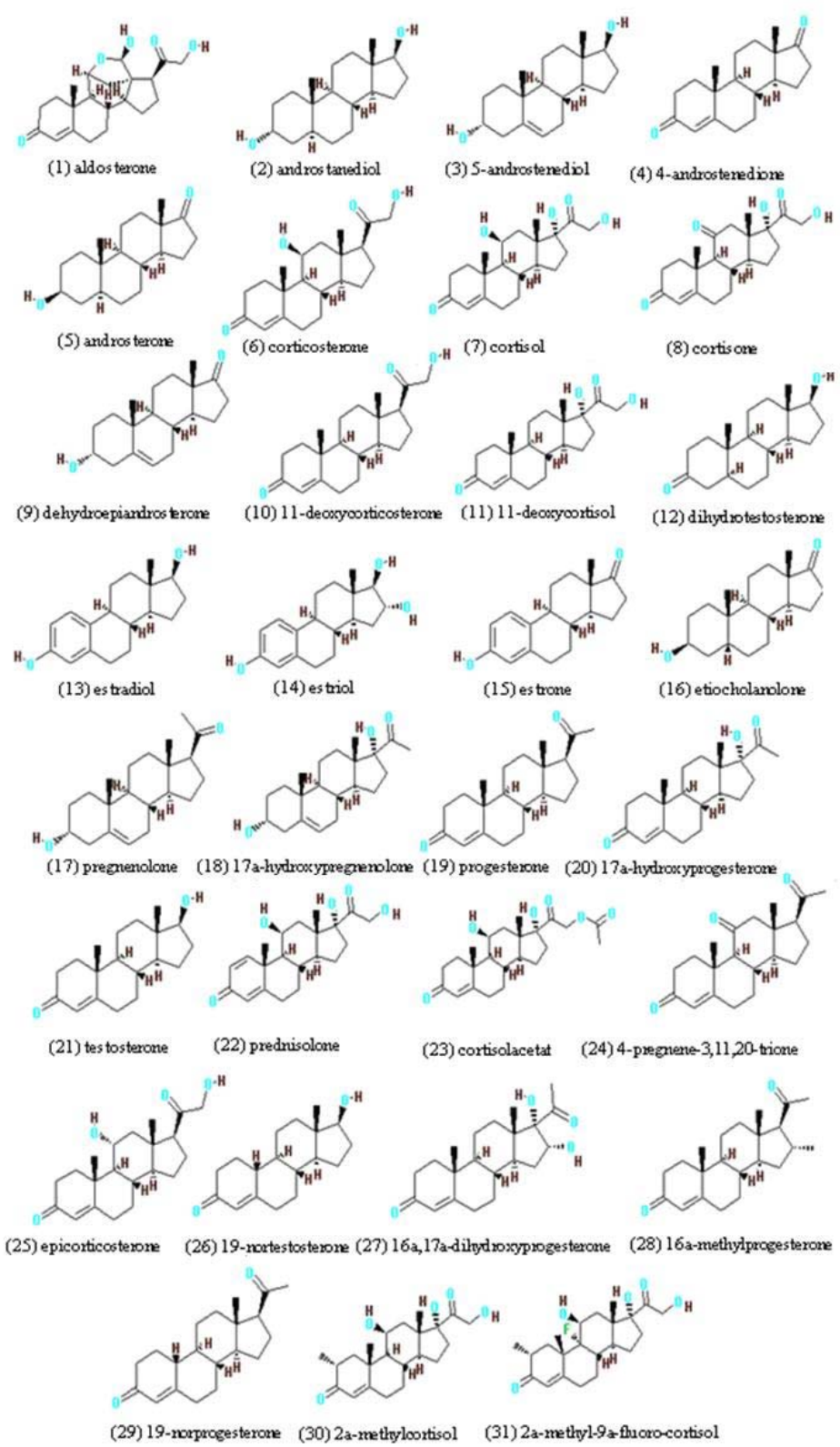


Figura 3: Esteróides utilizados nos testes. As amostras de 1 a 21 foram utilizadas como conjunto de calibração; da 22 até a 31, são as de validação.

3.2.3. Leitura de arquivos

Como o programa MultiMOL faz parte do pacote de programas CAMOL, ele precisa estar integrado com os demais programas, como por exemplo, nos formatos de arquivos de entrada e saída usados ou gerados nos cálculos realizados por ele. Para tanto, foi fundamental definir um formato padrão de leitura e de escrita de arquivos, que possa ser utilizado pelo MultiMOL e pelos demais programas do pacote.

Optou-se pela utilização de arquivos ASCII (caracteres de texto) no formato “csv” (“comma separated value”), o que permite clareza na interpretação dos dados e portabilidade do arquivo de entrada, uma vez que este pode ser processado por softwares populares de manipulação de planilha convencionais, como o Microsoft Excel e tantos outros. Isto facilita bastante a obtenção dos arquivos que serão utilizados pelo programa. Por uma questão de independência das opções de idioma selecionadas para cada computador no qual porventura o software venha a ser executado, convencionou-se que o caracter separador de valores é o ponto-e-vírgula (“;”), e não a vírgula. Isto porque, em português, a vírgula é convencionalmente utilizada como separador entre a parte inteira e valores decimais de um número real.

Os arquivos de texto que são reconhecidos pelo MultiMOL, assim, devem conter simultaneamente tanto os dados numéricos do problema quanto os nomes ("labels") das amostras (moléculas, em QSAR) e das variáveis (descritores, em QSAR) que compõem a matriz de dados original a ser utilizada pelo programa. A representação dos dados numéricos pode ser feita em notação normal ou científica, sendo ambas as formas reconhecidas pelo programa. A interface gráfica de usuário (“GUI”) permite a importação simples destes arquivos, os quais podem ser selecionados por meio da exploração dos diretórios do computador após selecionar-se a opção “importar”. Por fim, utilizando-se a representação dos dados no formato csv conforme acima explicado, o *layout* deste arquivo de entrada (“input”) é o seguinte (as cores indicam a legenda dos campos):

[dimensão da matriz; número de linhas e número de colunas] – 1ª linha
[nomes dos descritores] – 2ª linha
[nome da amostra] [valores numéricos] – linhas seguintes

```
Exemplo:
31;1813
1S000001;1S000002; (...) 1S001813;F_RESP
mag001_b;0.072471;-0.042452; (...) -0.041185;7.77
mag003_b;0.071409;-0.054886; (...) -0.05609;7.68
mag004_b;0.075778;-0.064991; (...) -0.056773;7.64
[...]
mag091_b;0.019453;-0.088005; (...) 0.011067;5.75
```

Note-se que o formato reconhecido pelo programa precisa seguir, rigorosamente, o formato acima especificado: na primeira linha, as dimensões da matriz; na segunda, os nomes dos descritores utilizados e da função-resposta observada, que deve ocupar a última posição da linha; e, da terceira linha em diante, por tantas linhas quantas forem as amostras do conjunto, o identificador de cada amostra, seguido dos seus descritores (na mesma ordem em que foram anteriormente nomeados) e do valor da função-resposta correspondente.

3.2.4. Pré-Processamento

Para a correta utilização dos dados em problemas de estatística multivariada, é muitas vezes importante a aplicação de algum tipo de processamento prévio nos dados que serão utilizados para a construção dos modelos de regressão [BRERETON, 2000; GAUDIO *et al.*, 2001]. Dois tratamentos matemáticos são bastante utilizados: **(a)** centrar na média, que consiste em subtrair de cada um dos elementos de cada coluna da matriz de dados o valor da média daquela coluna; e **(b)** escalonar, que consiste em dividir cada um dos elementos de cada coluna da matriz de dados pelo desvio padrão da respectiva coluna. Caso o escalonamento seja feito na matriz que já está centrada na média, tem-se então o que é chamado auto-escalonamento. Tais métodos têm a grande vantagem de não acrescentarem complexidade quase alguma ao código do programa, tendo, portanto, uma demanda computacional relativamente baixa, ao mesmo tempo em que possibilitam uma melhoria significativa na qualidade dos modelos obtidos. Além disso, certos tipos de problemas exigem algum tipo de processamento prévio como condição necessária para a correta construção dos modelos de regressão; por exemplo, em problemas de QSAR onde os descritores tenham significados físicos diferentes (que, inclusive, podem estar expressos em ordens de grandeza diferentes), o escalonamento da matriz de dados original é essencial para que não se verifiquem distorções nos resultados encontrados. Estes métodos de pré-processamento, por fim, possuem também a vantagem de serem procedimentos independentes do cálculo dos modelos de regressão, sendo possível aplicá-los apenas uma vez e, após isso, salvar a matriz já com os dados pré-processados, ganhando-se tempo no futuro. Por definição, o MultiMOL salva automaticamente uma cópia da matriz com os dados pré-processados antes de iniciar o cálculo da regressão.

O procedimento de centrar na média é recomendado para a maior parte dos modelos de regressão construídos com base no PCR. O escalonamento é necessário para quaisquer problemas – notoriamente os de QSAR Tradicional, como já indicado – onde as grandezas mensuradas em cada um dos descritores sejam de natureza distinta, com variâncias diferentes.

3.2.5. Seleção de Variáveis

Para minimizar a redundância da informação nas variáveis independentes e reduzir a demanda computacional exigida nestes problemas de natureza multivariada (por causa do elevado número de variáveis envolvidas no problema), foi implementado no MultiMOL um procedimento moderno e robusto (i.e. capaz de processar dados de grande volume – da ordem de milhares de variáveis) para uma seleção prévia das variáveis que serão utilizadas na construção dos modelos de regressão. Existe um algoritmo, utilizado normalmente em química analítica, que é denominado *SPA* ("Successive Projections Algorithm"; em português, APS, "Algoritmo de Projeções Sucessivas") [ARAÚJO *et al.*, 2001]. Ele funciona iterativamente, através da seleção, dentro do conjunto de vetores inicial, daqueles que possuem menos colinearidade entre si. Deste modo, é possível selecionar as 'x' primeiras variáveis menos correlacionadas, ou seja, aquelas que expliquem a maior variância possível dentre o conjunto original.

O algoritmo APS original compreende, além desta seleção das variáveis, a construção de diversos modelos de regressão com vistas a encontrar o número ótimo de variáveis a serem selecionadas. Especificamente, são selecionados 'n' conjuntos de variáveis por meio da aplicação das projeções sucessivas, cada um desses conjuntos com números distintos de variáveis selecionadas, contendo no máximo um número 'p' de variáveis previamente informado pelo usuário. Após isso, é aplicada uma regressão linear múltipla (MLR) para cada um destes conjuntos; os modelos gerados são, então, comparados entre si, e o que apresentar melhores resultados determinará o sub-conjunto das variáveis originais que será selecionado pelo APS como sendo o melhor para aquele conjunto e para os parâmetros informados.

No MultiMOL, entretanto, não foi implementado o algoritmo APS rigorosamente da forma como ele se encontra na literatura, e sim uma adaptação do mesmo, conforme está exposto a seguir. A utilização das projeções sucessivas para selecionar as variáveis que possuem menor colinearidade foi mantida na íntegra. No entanto, a parte da otimização do número de variáveis selecionadas não foi implementada. No MultiMOL, o usuário informa qual o número de variáveis que ele deseja selecionar; o programa utiliza, então, o mesmo critério de seleção que o APS emprega, para selecionar exatamente aquele número de variáveis que o usuário solicitou. Em outras palavras, enquanto que o APS descrito na literatura constrói diversos modelos de regressão para selecionar o melhor número de variáveis com no máximo o valor informado pelo usuário, o MultiMOL aplica somente o algoritmo de seleção do APS uma única vez, para selecionar exatamente o número de variáveis que o usuário informou.

Este algoritmo pode ser matematicamente expresso da seguinte maneira: seja X_{nxp} a matriz de dados original, onde 'n' é o número de objetos e, 'p', o de variáveis. Seja V={1,2,...,p} o conjunto das colunas de X que se referem às variáveis independentes. O que se almeja é selecionar q variáveis, $1 \leq q \leq p$, do conjunto V das

variáveis originais e, assim, construir uma matriz X'_{nxq}, onde q é um subconjunto de p com cardinalidade menor do que a de p.

Aqui, é importante salientar que há uma diferença essencial entre esta abordagem e a transformação do espaço original que se obtém por meio, por exemplo, da PCA: o resultado da Análise de Componentes Principais é um espaço transformado e, por isso, distinto do espaço que está representado nos dados originais e que possui um significado físico. Já a seleção de variáveis permite que se esteja trabalhando com o mesmo espaço dos dados originais, apenas considerando-se um número menor de variáveis (aquelas que, matematicamente, têm maior significado estatístico, por possuírem menor colinearidade). Isto significa que, embora haja alguma perda de informação, não há nenhuma transformação teórica que comprometa o significado empírico dos dados com os quais se está trabalhando.

É importante também ressaltar que correlação e colinearidade são conceitos próximos, mas não intercambiáveis; em um universo linear, as variáveis que possuem maior colinearidade serão também as mais correlacionadas entre si, mas isto deixa de ser válido quando se quebra a linearidade dos algoritmos matemáticos empregados na construção dos modelos de regressão. Então, por exemplo, na utilização do método quadrático Q-PLS (que será apresentando mais adiante neste texto), não necessariamente as variáveis menos colineares serão aquelas menos correlacionadas. É importante lembrar que o APS minimiza a colinearidade, e não a correlação; será válido dizer este último quando a relação entre as variáveis for de ordem linear, mas isto deixará de ser verdade quando, p.ex., a relação entre elas seja de ordem quadrática. Neste caso, pode acontecer que estas variáveis estejam correlacionadas sem, no entanto, serem colineares. O algoritmo de seleção de variáveis implementado no MultiMOL não trataeste problema; esta limitação precisa ser conhecida do usuário para que ele possa, assim, decidir pela oportunidade ou não de aplicar esta seleção de variáveis em cada caso concreto.

O algoritmo APS é um algoritmo iterativo que se baseia na descoberta, dentre as variáveis originais, daquelas que possuem menor colinearidade entre si. O procedimento pelo qual isso é feito é, sucintamente, descrito abaixo:

i) inicia-se com uma coluna da matriz de dados original (comumente, aquela que apresenta a maior norma);

ii) calcula-se, para cada uma das demais colunas da matriz, a sua projeção relativa àquela escolhida no passo anterior;

iii) verifica-se, dentre todas as projeções, qual aquela que apresenta maior norma;

iv) seleciona-se essa e repete-se o passo (ii) para as colunas restantes.

Ao final da execução deste procedimento, obtém-se, da matriz original, tantas variáveis quantas forem desejadas, dispostas em ordem crescente de colinearidade,

i.e., das menos colineares para as mais colineares. Em outras palavras, as primeiras variáveis obtidas por meio desse procedimento são – pelo menos no universo dos modelos de regressão linear, conforme já foi discutido – as mais independentes e menos correlacionadas. Essas variáveis são, portanto, as mais interessantes para serem selecionadas para a construção dos modelos de regressão.

3.2.6. Análise de Componentes Principais

O primeiro problema encontrado na construção de modelos de regressão, tipicamente para problemas de QSAR-3D, é a sua natureza multivariada. Esta categoria de problemas trabalha com um universo vasto de variáveis independentes (descritores), e deve-se buscar uma correlação matemática entre estes e a propriedade de interesse, ou seja, a variável dependente – atividade biológica, propriedade, toxicidade, dependendo do problema: se é atividade biológica, QSAR; se propriedade, QSPR; se toxicidade, QSTR, respectivamente.

O tamanho deste conjunto de dados original pode facilmente atingir a casa das dezenas (se o problema for de QSAR tradicional) ou até mesmo dos milhares (se o problema for de QSAR-3D). Caso a solução almejada consistisse simplesmente no estabelecimento de uma correlação entre uma única variável independente e a função-resposta de interesse, a estatística básica poderia oferecer as ferramentas suficientes para tal, por meio de uma regressão linear simples obtida através do método de mínimos quadrados, por exemplo. No entanto, quando a complexidade do problema é tal que se torna necessário olhar não apenas para um único descritor, mas para um conjunto deles, é necessário utilizar-se de estatística multivariada, o ramo da estatística específico para o tratamento desta categoria de problemas.

Torna-se assim necessário reduzir a dimensionalidade dos dados dos quais se dispõe originalmente. Um dos motivos pelos quais esta redução é necessária é a importância de reduzir a demanda computacional exigida para o cálculo dos modelos de regressão; isto pode ser conseguido por meio da aplicação de fórmulas matemáticas a um conjunto de dados transformado, menor do que o original, mas que pode ser neste diretamente mapeado. Outro motivo é para possibilitar a visualização gráfica dos dados com os quais se está trabalhando. Uma matriz de p colunas pode ser interpretada como um conjunto de pontos distribuídos em um espaço p-dimensional; para que seja possível obter uma representação gráfica desses pontos, é necessário que o número de colunas seja pequeno, especificamente, dois ou três, porque são estas as dimensões que podem ser visualizadas em gráficos tradicionais bi ou tridimensionais. Para resolver este problema é útil a aplicação de uma Análise de Componentes Principais (PCA – "Principal Components Analysis") aos dados originais. Há ainda, por fim, um terceiro motivo de ordem teórica para que seja necessário diminuir o espaço dos dados com os quais se deseja trabalhar: para a realização de

uma regressão linear múltipla, o número de variáveis selecionadas não pode ser maior do que o número de objetos disponíveis na construção do modelo de regressão. Esta limitação é inerente aos algoritmos matriciais utilizados para o cálculo dos modelos [FERREIRA *et al.*, 1999].

A análise de componentes principais consiste basicamente na decomposição da matriz original em duas outras matrizes, uma de pesos (*Loadings*) e outra de *Scores*, por meio da diagonalização da matriz de dados original, a qual pode ser feita de maneira direta (SVD – "Single Value Decomposition") ou iterativa (NIPALS – "Non-linear Iterative Partial Least Squares"). Cada componente principal é, assim, definida por uma combinação linear das variáveis originais.

Isto é feito por meio da definição de novos eixos orientados, ortogonais entre si, em relação aos quais são projetados os objetos originais. Isso é interessante porque, na prática, devido à similaridade que as moléculas utilizadas têm entre si e à grande colinearidade que as variáveis medidas apresentam, esses objetos tendem a se organizar, no espaço multidimensional original, em padrões e agrupamentos de dimensionalidade muito menor. Desse modo, é possível a representação da maior parte da informação contida nas variáveis medidas originais em um espaço de dimensão consideravelmente menor e, por conseguinte, muito mais fácil de ser tratado e interpretado. O algoritmo utilizado para a sua implementação foi o NIPALS [BRERETON, 2000], que é iterativo. Isto significa que ele é construído de tal maneira que obtém, a cada execução, a componente principal que explica a máxima variância restante do conjunto de dados original (por exemplo, a primeira componente principal é o vetor que tem a direção da maior variância dos dados analisados; a segunda componente principal é o vetor, ortogonal ao primeiro, que tem a direção da segunda maior variância destes dados; e assim sucessivamente). Outra característica sua é que a porcentagem da cobertura da variância é numericamente mensurável, de tal maneira que é possível obter tantas componentes principais quantas forem necessárias, limitadas, naturalmente, à quantidade da informação constante no conjunto de dados original que se deseje considerar.

Em resumo, a idéia por traz da utilização da Análise de Componentes Principais nos problemas de QSAR pode ser esquematizada da seguinte maneira: os objetos dispostos em um espaço h-dimensional (de um grande número de dimensões) tendem, por sua própria natureza, a organizarem-se em agrupamentos p-dimensionais (de um número muito menor de dimensões do que o espaço original); isto ocorre porque há muita variância entre os dados que é simplesmente devida ao ruído na mensuração experimental. Existem técnicas da Álgebra Linear que permitem encontrar um novo sistema de eixos orientados de tal maneira que a informação presente nos dados originais seja representada em um número menor de dimensões. Assim sendo, é interessante ter à disposição estas ferramentas matemáticas que possam ser utilizadas para a redução da dimensionalidade de problemas que são por sua natureza complexos, a fim de que eles possam ser mais eficazmente tratados. Na prática, o que

se tem é a necessidade de uma escolha ("*trade-off*"[1]), onde é necessário escolher o melhor ponto entre a perda de um pouco da informação contida na matriz de dados original e a troca por uma representação em dimensionalidade menor. Felizmente, é relativamente comum que seja encontrado, logo nas primeiras componentes principais, a maior parte da variância da matriz original já explicada pelo novo sistema de eixos ordenados e, dependendo da natureza do problema em análise, isso pode ser mais do que suficiente. O acréscimo de dimensões extras muitas vezes não é interessante, uma vez que eles acrescentam pouca informação à representação já obtida com um número menor de eixos, bem como podem acrescentar informação proveniente de ruído. Isso pode aumentar a complexidade, pode aumentar a demanda computacional dos cálculos, e pode fazer com que se perca a possibilidade de uma representação gráfica do problema.

Esta forte correlação entre os dados encontrada em problemas desta natureza, de modo especial na QSAR-3D, torna particularmente importante a implementação da metodologia de Análise de Componentes Principais. Diz-se que duas variáveis estão correlacionadas quando elas têm algum tipo de associação entre as suas variações entre elas; por exemplo, se, quando uma variável aumenta, uma segunda aumenta também, elas estão positivamente correlacionadas e, se quando uma variável aumenta, uma segunda diminui, elas estão negativamente correlacionadas. A correlação, assim, acaba por inserir redundância na informação com a qual se deseja trabalhar, uma vez que, se duas (ou mais) variáveis estão fortemente correlacionadas, a mesma (ou quase a mesma) informação que pode ser extraída delas poderia ser obtida de apenas uma delas – variáveis obtidas de maneira independente, mas que sejam fortemente correlacionadas (p.ex., massa e volume), não acrescentam ao modelo uma quantidade significativa de informação, podendo até mesmo gerar problemas advindos da introdução de ruído e de informações redundantes. A existência de muita correlação no conjunto de dados de trabalho implica também em uma maior demanda computacional exigida (devido ao número excessivo de variáveis com as quais se está, desnecessariamente, trabalhando),

Excessiva correlação entre os dados provoca também uma degradação dos modelos de regressão quando se usam as variáveis originais (devido ao fato de que eles estão sendo construídos levando em consideração informações que não são relevantes). A PCA consiste na representação dos dados originais em um espaço transformado, de tal maneira que este é construído como uma combinação linear das variáveis originais, buscando sempre, a cada componente principal, a máxima explicação da variância (variabilidade) restante; isto faz com que, como já foi dito, os dados transformados sejam significativamente mais descorrelacionados do que os dados originais.

---

[1] O termo, comum na ciência da computação, designa uma situação tal em que é necessário estabelecer um ponto de equilíbrio aceitável entre duas características que estão inversamente correlacionadas, i.e., uma será tanto melhor quanto pior for a outra, e isto de tal maneira que não é possível otimizar ambas.

Como já foi indicado anteriormente, a redução da dimensionalidade é também indispensável para a viabilização matemática da construção dos modelos de regressão, uma vez que a regressão linear múltipla [FERREIRA *et al.*, 1999] exige que o número de amostras presentes no conjunto original seja maior ou igual ao número de descritores analisados, situação que raras vezes se apresenta nos casos de QSAR tradicional e nunca é encontrada nos casos de QSAR-3D. Para se aplicar uma MLR em uma matriz de dados que possua n linhas e m colunas, é necessário que $m \geq n$ e, como a natureza física dos problemas analisados é de tal maneira que esta situação praticamente não se verifica, é quase sempre necessário decompor a matriz de dados original de tal maneira que seja matematicamente possível aplicar os algoritmos pertinentes. Uma das formas de se decompor a matriz original em uma matriz transformada de dimensionalidade menor é justamente por meio da Análise de Componentes Principais, e a aplicação de uma Regressão Linear Múltipla na matriz de Scores (novas coordenadas, nas novas variáveis, dos elementos originais) obtida a partir da aplicação da PCA é chamada de Regressão de Componentes Principais (PCR – "Principal Componentes Regression"), sendo esta a metodologia mais simples para a construção de modelos de regressão, como pode ser visto na seção seguinte. Existem outras metodologias que podem ser aplicadas na decomposição das matrizes de dados originais, como por exemplo o algoritmo PLS ("Partial Least Squares"), sobre as quais falaremos em detalhes mais adiante.

3.2.7. Regressão em Componentes Principais

Para o MultiMOL, optou-se pela implementação da metodologia PCR (que, como visto, consiste basicamente em uma Regressão Linear Múltipla precedida de uma Análise de Componentes Principais). Para a aplicação da PCA, optou-se pela implementação do algoritmo NIPALS, em detrimento de outros (como o SVD – "Singular Value Decomposition"-, por exemplo), porque o custo computacional do NIPALS ("Non-linear Iterative Partial Least Squares") é proporcional ao número de componentes principais desejado, uma vez que, com ele, apenas são calculadas as componentes principais que sejam necessárias para a construção do modelo. Com isto, não há desperdício de tempo computacional com o cálculo de componentes principais que não serão utilizadas na elaboração dos modelos de regressão. Por outro lado, o algoritmo SVD só pode ser aplicado na matriz inteira, para a sua decomposição em todas as componentes principais que ela possui, o que nem sempre será necessário para construir modelos QSAR.

O algoritmo NIPALS encontra-se amplamente documentado na literatura [BRERETON, 2000], o que facilita quer a sua implementação e a comparação do MultiMOL com outros softwares e pacotes que são implementados com NIPALS, como por exemplo o Matlab. Além disso, a PCR exige dois módulos distintos e bem

separados: (i) o módulo de decomposição da matriz original em uma matriz menor, que contenha as componentes principais desejadas; e (ii) o módulo de regressão linear múltipla. Ao escolher a implementação do PCR, portanto, existe a opção de utilizar o segundo módulo independentemente do primeiro, aplicado diretamente à matriz de dados original e, assim, adiciona-se a possibilidade de fazer, além da Regressão por Componentes Principais, uma Regressão Linear Múltipla, o que aumenta a abrangência e aplicabilidade do MultiMOL e enriquece o leque de possibilidades apresentado ao usuário do programa.

Uma das desvantagens da PCR em relação à MLR é que a primeira é aplicada em um espaço dimensional transformado; isto faz com que se perca a interpretação física direta dos resultados obtidos no modelo, uma vez que as variáveis utilizadas não são mais as variáveis originais do problema, e sim uma combinação das mesmas, projetadas em um espaço de dimensionalidade menor. Assim sendo, apesar do indiscutível ganho de desempenho advindo da aplicação de uma regressão em componentes principais, é necessário fazer uma minuciosa análise caso a caso, uma vez que, para o problema concreto que se está tratando talvez seja interessante manter o modelo construído tendo por base o espaço de variáveis original, mesmo às custas de uma maior demanda computacional.

3.2.8. PLS Tradicional

Uma alternativa à PCR é a utilização do algoritmo PLS (“Partial Least Squares”) [TOBIAS, 1999], que consiste também em uma decomposição da matriz de dados original em um conjunto menor que explique a maior variância possível. A principal diferença entre ambos os métodos é que, no PLS, é considerada também a função-resposta (variável dependente) na decomposição, enquanto que no PCR, é utilizada apenas a matriz de dados com as variáveis independentes. Embora os resultados obtidos por qualquer um dos métodos sejam normalmente equivalentes do ponto de vista estatístico (preditividade), existe um grande ganho de tempo de processamento (performance) associado à utilização do método PLS, de modo que foi julgada relevante a inclusão desta opção no MultiMOL. Às variáveis que definem o espaço transformado por meio da execução do PLS dá-se o nome de Variáveis Latentes, para distinguir daquelas que são obtidas por meio da aplicação da PCA, as quais são chamadas Componentes Principais. Componentes Principais são espécie do gênero Variáveis Latentes, de tal modo que toda componente principal é uma variável latente, mas a recíproca não é verdadeira.

Há dois algoritmos tradicionalmente utilizados para a decomposição da matriz original em uma matriz de Loadings e outra de Scores por meio do PLS: o PLS1 e o PLS2 [BRERETON, 2000]. O primeiro aplica uma seqüência única de transformações matriciais no conjunto de dados (variáveis e função-resposta), decompondo-o assim

nas matrizes desejadas. O segundo consiste em uma aplicação iterativa de operações matriciais, buscando a convergência dos valores obtidos, o que faz com que ele possua uma demanda computacional menor do que o PLS1. Historicamente, o PLS2 foi desenvolvido para que se pudesse fazer a decomposição simultânea de um conjunto de dados para os quais havia mais de uma função-resposta a ser considerada. Isto foi importante em um momento onde limitações computacionais tornavam excessivamente oneroso multiplicar as aplicações do PLS1 para a mesma matriz de dados, variando apenas a função-resposta. No entanto, com os avanços tecnológicos nos processadores e a facilitação do acesso a recursos computacionais mais potentes, hoje em dia já se torna viável aplicar diversas vezes o PLS1, para os casos em que isso se faça necessário, de modo que o PLS2 caiu em desuso.

A grande vantagem de desempenho obtida com a utilização do PLS é decorrente do fato de que o algoritmo escolhido para implementação – o PLS1 – é baseado em uma única execução por variável latente, ao contrário do NIPALS (utilizado na implementação do PCR), que pressupõe sucessivas iterações a cada componente principal buscando a convergência dos "Loadings" e "Scores".

Assim, o número de operações matemáticas executadas pelo algoritmo PLS é consideravelmente menor em comparação com o PCR, o que justifica a melhoria no desempenho.

3.2.9. PLS Quadrático

O PLS pode ser expandido para a obtenção de relações não-lineares entre as amostras e suas funções-resposta. É possível que encontremos, por exemplo, um determinado conjunto de descritores que, por sua própria natureza, esteja não linearmente, mas quadraticamente relacionado com a função-resposta das amostras que se está analisando. Para estes casos, é interessante que exista algum modelo que seja capaz de levar em consideração este elemento quadrático na elaboração das equações de regressão, a fim de apresentar uma melhor modelagem dos dados e possibilitar a construção de modelos que sejam mais adequados. Neste caso, é importante considerar a utilização do PLS Quadrático (Q-PLS) para obter modelos que sejam mais precisos, de preferência, com um menor número de variáveis latentes. É importante notar que o Q-PLS não está implementado em nenhum pacote de software comercial, de modo que o MultiMOL apresenta-se assim como uma ferramenta diferenciada e inovadora por disponibilizar esta funcionalidade, com benefícios para o usuário.

O Q-PLS [WOLD *et al.*, 1989], à semelhança do PLS tradicional, consiste também na decomposição do conjunto de dados original em duas matrizes de “Loadings” e “Scores”. Tais matrizes são aqui construídas de modo a considerarem a dependência quadrática entre as variáveis independentes e a dependente, de modo

que seja possível obter uma equação de regressão que inclua termos quadráticos, o que é capaz de fornecer uma melhora na modelagem para os problemas que têm intrinsecamente uma dependência desta natureza. Ao contrário do PLS tradicional (baseada em PLS1), no entanto, esta implementação (baseada em PLS2) exige a busca por convergência no cálculo de cada uma das variáveis latentes, o que torna a sua demanda computacional muito maior do que aquela apresentada pelo PLS.

A sugestão de implementação de um PLS quadrático feita por Wold possui as seguintes características. Um modelo PLS objetiva relacionar duas matrizes, X e Y. Aqui estamos nos referindo aos modelos preditivos em dois blocos, PLS2, uma implementação ligeiramente diferente do PLS que permite a calibração em blocos, ou seja, modela simultaneamente um conjunto de respostas yi, organizadas como colunas da matriz Y, diferente do PLS1, a implementação original onde apenas uma resposta y é modelada por vez. Assumindo a convenção de usar letras minúsculas para vetores colunas, e letras minúsculas seguidas de ’ para vetores linhas, onde ’ e o operador transposição de matriz, as equações do PLS, escritas de uma forma mais geral, expressam a decomposição das matrizes X e Y como apresentado a seguir:

$$\mathbf{X} = \mathbf{tp}' + \mathbf{E} \qquad \textbf{(equação 4)}$$

$$\mathbf{Y} = \mathbf{uq}' + \mathbf{F} \qquad \textbf{(equação 5)}$$

Tradicionalmente é assumido uma relação linear entre os escores t e u (h representa o resíduo):

$$\mathbf{u} = \mathbf{bt} + \mathbf{h} \qquad \textbf{(equação 6)}$$

Assim é possível modelar Y por t e q

$$\mathbf{Y} = \mathbf{tq}'\mathrm{b} + \mathbf{F} \qquad \textbf{(equação 7)}$$

uma vez que:

$$\mathbf{u} = f(\mathbf{t}) + \mathbf{h} \qquad \textbf{(equação 8)}$$

A ideia por tráz do QPLS é simplesmente reescrever a relação entre **u** e **t**. Assim, para uma relação polinomial quadrática:

$$\mathbf{u} = c_0 + c_1\mathbf{t} + c_2\mathbf{t}^2 \qquad \textbf{(equação 9)}$$

O desafio na implementação desta proposta é reescrever o algoritmo do PLS para construir modelos que levem em conta esta relação. No QPLS isto é feito modificando-se o algoritmo para o PLS2, fazendo a substituição de **t** por **Xw**, onde **w** é o vetor de pesos do PLS, e definindo **u** = *F*(**X**,**w**,**c**). A cada iteração, **t**, **q**, e **u** são calculados segundo o PLS ordinário, **c** é obtido por mínimos quadrados e **w** é incrementado de um valor d**w** calculado a partir da linearização de **u** = *F*(**X**,**w**,**c**).

Igualmente, a inclusão do termo quadrático só apresenta resultados satisfatórios quando o problema que se está tentando modelar é de natureza quadrática, por isto, torna-se importante possuir um bom conhecimento prévio dos dados com os quais se está trabalhando, a fim de que seja possível tomar uma decisão acertada sobre qual tipo de modelo será escolhido para a regressão.

Há várias situações em que as não linearidades podem estar presentes, como por exemplo: i) não homogeneidade na amostra; ii) não linearidades de detectores fotocondutivos; iii) não linearidades em medidas de transmitância/reflectância difusa em espectroscopia NIR; iv) não linearidades químicas devido a mudanças de interações moleculares em função da concentração, composição da matriz ou condições de medida; v) não linearidades na relação entre a propriedade a ser calibrada e a concentração, uma vez que a lei de Beer só garante a linearidade do espectro com a concentração, mas não do espectro com a propriedade dependente da concentração. Fora do contexto analítico, as relações estrutura-atividade e as superfícies de respostas usadas na otimização, que não são regidas pela lei de Beer, podem apresentar máximos ou mínimos no domínio investigado. A utilização de calibração multivariada não-linear continua sendo pouco aplicada, mesmo em situações onde seria vantajosa, como por exemplo, em modelos de previsão baseados em espectroscopia NIR ("Near-Infrared") e/ou QSAR. A ausência de ferramentas de fácil acesso para a realização dessas calibrações pode ser uma das responsáveis por esse panorama e a implementação atual presente no MultiMOL tenta fornecer uma alternativa viável.

Vale a pena distinguir o algoritmo Q-PLS, implementado no MultiMOL, do PLS linear com projeção quadrática de X. Este último, como explica Wold [WOLD *et. al.*, 1989], consiste na aplicação do PLS linear à matriz de dados original acrescida dos termos quadráticos (x_k^2) e dos termos cruzados ($x_j * x_k$). Foi demonstrado que a projeção desta matriz estendida em um plano corresponde à projeção da matriz original em uma superfície quadrática. Tratam-se de duas abordagens distintas. O Q-PLS aplica uma metodologia de regressão de natureza quadrática a um conjunto de dados real para obter uma equação de regressão quadrática. O PLS linear com projeção quadrática aplica uma metodologia de regressão linear a um conjunto de dados composto pelos dados reais acrescidos de termos quadráticos e cruzados, para obter no final uma equação de regressão linear. Assim, no Q-PLS, o efeito quadrático é considerado no algoritmo de regressão e, no PLS linear com projeção quadrática, o efeito quadrático é introduzido por meio de uma manipulação dos dados de entrada.

Entre as vantagens do Q-PLS em relação ao PLS linear com projeção quadrática, pode-se destacar o melhor desempenho do primeiro algoritmo, uma vez que o acréscimo dos termos quadráticos e cruzados das variáveis originais aos dados de entrada faz com que o programa precise trabalhar com uma matriz substancialmente maior. Nos problemas de QSAR, especialmente de QSAR-3D, o número de descritores (variáveis independentes) pode atingir facilmente a casa das centenas ou até mesmo dos milhares, fazendo com que a aplicação do PLS linear com projeção quadrática exija

recursos computacionais muito maiores, além de tornar os resultados obtidos mais difíceis de serem interpretados. Estes foram os motivos que guiaram a escolha do Q-PLS como algoritmo de regressão não-linear oferecido pelo MultiMOL.

É importante também salientar que a implementação do Q-PLS pode ser facilmente estendida para outros tipos de modelos não-lineares (exponenciais, logarítmicos, etc.), por meio de pequenas alterações no algoritmo utilizado, aproveitando-o em grande parte. Embora não haja previsão imediata para a implementação de outros modelos de regressão no MultiMOL, o programa foi concebido e implementado de tal maneira que pudesse ser estendido facilmente quando se identificar a necessidade ou a oportunidade da disponibilizar outros modelos de regressão.

3.2.10. Validação cruzada e Validação com série de teste

Para a obtenção dos modelos de regressão, é freqüentemente útil dividir o conjunto de dados original em dois, um para a calibração do modelo e outro para aferir a qualidade do modelo construído (validação). O MultiMOL fornece duas opções para dividir o conjunto de entrada. Na primeira delas, o usuário informa expressamente quais as amostras (ou o intervalo de amostras) que ele deseja reservar para o conjunto de validação. Na segunda opção, o programa separa automaticamente uma a cada x amostras, onde x é um número inteiro, informado pelo usuário, e que vai de 1 até o número máximo de amostras (n) do conjunto original. Como exemplo, se o usuário escolhe x=4, então, 1 (uma) a cada 4 (quatro) amostras vai ser escolhida para validação, o que significa que 25% do conjunto original de amostras vai ser destinado para este fim.

Para fazer a calibração do modelo, é possível utilizar o método tradicional de validação cruzada “leave-one-out-full-cross-validation” (LOO-FCV), ou utilizar a série de teste selecionada nos procedimentos descritos logo acima. No LOO-FCV [BRERETON, 2000], que se trata de um método comumente utilizado pela comunidade científica de química medicinal para validação de modelos de QSAR (principalmente quando se tem poucas amostras (moléculas) disponíveis), o número de componentes principais (variáveis latentes) é otimizado construindo-se ‘n’ modelos, cada um deles com ‘n-1’ amostras (onde ‘n’ é o número de amostras), e calculando-se o coeficiente de correlação da validação cruzada (Q^2) para o conjunto dos modelos. Esta opção é geralmente utilizada quando o número de amostras é pequeno o bastante para impedir a separação dos dados em um conjunto de calibração e outro de validação. São bem conhecidas as dificuldades para obtenção de um conjunto de dados de porte tal que permita a sua divisão em dois conjuntos independentes e, por isso, a validação cruzada provavelmente será a alternativa utilizada em grande parte dos casos.

As funções estatísticas que podem ser usadas para mensurar a qualidade de um modelo estatístico construído com validação cruzada são as seguintes: (a) *SEP* ("Standard Error of Prediction") e (b) Q^2 (coeficiente de correlação da validação cruzada), conforme [GAUDIO *et. al.*, 2001]. O SEP é calculado conforme a equação 10, onde n é o número de amostras do conjunto de dados original. A partir de tais quantidades estatísticas, pode-se determinar quais são os modelos mais promissores dentre os que foram gerados, por exemplo, os modelos construídos com uma CP ou VL têm o seu valor de SEP e Q^2; os modelos construídos com duas vão ter outros valores, e assim por diante. O MultiMOL opta, como critério de identificação de melhor número de Variáveis Latentes, por aquele que apresenta o menor valor de SEP (que corresponde ao maior valor de Q^2) durante o processo de calibração.

$$SEP = \sqrt{[\sum (y_i obs - y_i cal)^2 / n]} \qquad \textbf{(equação 10)}$$

O processo iterativo de validação dos modelos por meio de LOO-FCV pode ser resumido da seguinte forma:

i. Calcula-se a primeira CP ou VL;
ii. Constroem-se tantos modelos quanto forem as amostras, deixando cada uma delas de fora (LOO-FCV);
iii. Calculam-se os valores estatísticos para estes modelos;
iv. Calcula-se a CP ou VL seguinte;
v. Caso não haja mais CPs ou VLs a serem calculadas, o algoritmo termina aqui; em outro caso, volta-se ao passo "ii".

Obtendo-se os valores de SEP e Q^2 para cada um destes conjuntos de modelos, é possível identificar qual é o número ótimo de CPs ou VLs para um determinado problema.

Um dos maiores problemas identificados na utilização da LOO-FCV é a alta demanda computacional exigida por estes procedimentos: é necessária a construção de tantos modelos quantos forem o número de amostras do conjunto de calibração, e isto se aplica a tantas CPs ou VLs, quantas se deseje obter. Num processo exploratório tradicional, onde são construídos todos os modelos com todas as CPs ou VLs para que seja identificado o número ótimo delas, o número de regressões realizadas é geralmente da ordem de N^2, onde N é o número total de amostras contidas no conjunto de calibração.

Por outro lado, se o usuário escolhe a opção de validação com série de teste, um único modelo é construído para cada número de VLs com o conjunto de calibração, e projeta-se neste modelo o conjunto de validação (série de teste). Neste caso, o número de VLs é escolhido como melhor modelo por meio da otimização do coeficiente de correlação da predição (R^2pred) sobre a série de teste. Esta opção, naturalmente, só

se torna disponível se o usuário escolhe uma das duas formas de seleção de amostras que o programa oferece para a escolha da série de teste. Esta alternativa é geralmente mais vantajosa em relação à validação interna por causa da maior independência entre os conjuntos de calibração e validação.

Ela consiste na construção de um único modelo, para cada número de variáveis latentes, com as amostras que fazem parte do conjunto de calibração; neste modelo, serão projetadas as amostras da série de teste (o conjunto de validação) e, por meio da comparação entre esses vários modelos, será identificado o número ótimo de variáveis latentes para o conjunto informado. Neste caso, o algoritmo de calibração do modelo será como segue:

i. Calcula-se a primeira CP ou VL;
ii. Constrói-se um único modelo, com as amostras do conjunto de calibração, no qual serão projetadas as amostras do conjunto de validação;
iii. Calculam-se os valores estatísticos para este modelo;
iv. Calcula-se a CP ou VL seguinte;
v. Caso não haja mais CPs ou VLs a serem calculadas, o algoritmo termina aqui; em outro caso, volta-se ao passo “ii”.

O critério utilizado para a determinação do número ótimo de variáveis latentes é o mesmo da validação cruzada: aquele modelo que apresentar o menor valor de SEP será o escolhido. Há uma diferença de terminologia entre os dois algoritmos que é digna de nota: enquanto que, na validação cruzada, calcula-se o coeficiente de correlação da validação cruzada (Q^2), nesta validação com série de testes utiliza-se o coeficiente de determinação para as amostras preditas (R^2_{pred}), cuja fórmula é exatamente a mesma apresentada na Equação 3. A diferença é meramente conceitual: enquanto que, para o cálculo do Q^2, o “valor predito” de cada amostra foi obtido de um modelo diferente (i.e., do modelo construído sem ela), para o cálculo do R^2_{pred} todos os valores preditos são obtidos de um mesmo modelo (o modelo construído com o conjunto de calibração inteiro). Matematicamente, as fórmulas são idênticas e o significado de ambos os valores é também o mesmo: trata-se de um coeficiente que ilustra quanto da variância do modelo é explicada deterministicamente.

É, por fim, importante notar que o MultiMOL permite também que seja feita a divisão do conjunto original de amostras e, mesmo assim, seja aplicada a validação LOO-FCV no conjunto de calibração para otimização do número de variáveis latentes (VLs), com posterior projeção nele das amostras da série de teste. Neste caso, a série de testes é utilizada para aferir a qualidade de um modelo que foi obtido por meio de validação cruzada. Esta metodologia consiste também na separação do conjunto de dados em dois subconjuntos: um deles será utilizado para a construção do modelo de regressão (conjunto de calibração), por meio da validação cruzada LOO-FCV, e o outro

será usado exclusivamente para a validação ou teste da capacidade preditiva do modelo de regressão.

Aqui é importante notar que a avaliação crítica do modelo fica ao encargo do responsável pela pesquisa: o melhor modelo obtido através da validação cruzada pode apresentar uma capacidade preditiva aquém da que seria desejada e, portanto, não ser um modelo adequado, a despeito dos bons resultados porventura obtidos durante na calibração e auto-predição. Note-se que se distinguem aqui três coisas: a calibração, tradicionalmente feita através da validação cruzada; a auto-predição (projeção das amostras que foram utilizadas na construção do modelo), que indica a capacidade explicativa do modelo; e, por fim, a predição com a série de teste (projeção de amostras que não foram utilizadas na construção do modelo), onde está indicada a capacidade preditiva do modelo.

No caso particular do QSAR, um problema que surge desta abordagem é o seguinte: como garantir que as amostras deixadas de fora estão compreendidas, interpolativamente, entre aquelas que foram utilizadas para a construção do modelo? Não se tem conhecimento, até onde foi possível averiguar, de uma resposta definitiva para este problema. Por isto, torna-se importante enfatizar, mais uma vez, a necessidade da escolha de um conjunto de moléculas que formem uma série homóloga, pois a presença de um grupo farmacofórico em todas elas, ao qual são meramente adicionados (ou subtraídos) alguns substituintes, tem garantido resultados experimentais que corroboram a tese de que é possível trabalhar desta maneira.

4. Resultados e Discussões

O registro dos resultados obtidos neste projeto foi feito com o intuito de comprovar a realização dos objetivos propostos originalmente. Para os resultados, foram utilizados os conjuntos de testes apresentados na seção 3.2.2. Por motivos de espaço, não serão apresentados na presente seção os arquivos de saída, gerados pelo programa, correspondentes a cada um dos modelos apresentados na seção 4.3. Todas as figuras apresentadas nesta seção foram obtidas do próprio MultiMOL.

4.1. Características do programa MultiMOL

O MultiMOL possui mais de 8.000 linhas de código, distribuídas entre 29 arquivos C++, sem levar em consideração aqueles que foram obtidos de bibliotecas públicas e incorporados ao software (NEWMAT). Foram escritos procedimentos para diversas funcionalidades: não somente na codificação dos algoritmos de regressão, mas também leitura e escrita de arquivos, interação com o usuário, definição e controle de componentes gráficos, estruturas de controle, processamento interno de dados,

contabilização de tempo de processamento, visualização de resultados, entre outras funcionalidades necessárias ao funcionamento do programa.

A opção pela representação dos valores numéricos com uma precisão dupla ("double") foi feita por meio da utilização do tipo de dado específico que a própria linguagem de programação (C++) disponibiliza; como já foi dito, esta escolha foi motivada pela necessidade de se obter resultados numéricos mais precisos. As matrizes foram representadas com as estruturas disponíveis na biblioteca NEWMAT, nas quais já estão incluídas as operações matriciais mais comumente utilizadas, como transposição, inversão e multiplicação de matrizes. Estas operações possuem já precisão e desempenho otimizados, o que se confirmou nos testes que foram executados; portanto, não foi necessário fazer nenhuma adaptação no código da biblioteca para atender às necessidades do MultiMOL. Como foi explicado na seção 3.2., a biblioteca NEWMAT é *open source*, o que significa que o seu código-fonte é aberto, é disponibilizado junto com a biblioteca e pode ser modificado para atender às necessidades específicas de cada aplicação que a utilize. No caso do MultiMOL não foi necessário fazer nenhuma alteração no código. No entanto, para que esta decisão fosse tomada, foi realizado um trabalho detalhado de análise do código-fonte desta biblioteca matemática para, enfim, chegar à conclusão de que ele, na forma como estava, atendia bem aos propósitos do software. Cabe lembrar que “open source” significa que não há custo monetário associado à licença de software, e que o código-fonte do software é disponibilizado na íntegra, podendo ser analisado e, se necessário, modificado. A biblioteca NEWMAT é gratuita e “open source”. Todos os demais algoritmos utilizados pelo programa MultiMOL para a obtenção dos modelos de regressão (MLR, NIPALS, PLS, Q-PLS) foram implementados por nós.

Para a leitura dos arquivos de entrada, optou-se por uma representação em ASCII (caracteres de texto). Isto permite uma manipulação dos dados mais simples pelo usuário, que pode ser feita através de programas de edição de texto comuns, como foi explicado na seção anterior (ver seção 3.2.3).

Todas as opções referentes à construção do modelo estão acessíveis ao usuário por meio da Interface Gráfica (GUI). Após selecioná-las, basta clicar no botão de executar para que o programa execute o cálculo dos modelos com base nos parâmetros informados pelo usuário (ver seção 4.2. para mais detalhes).

O procedimento de validação cruzada tradicionalmente utilizado para a identificação automática do número ótimo de variáveis latentes exige que sejam construídos diversos modelos de regressão. Para conjuntos de dados que possuam um grande número de amostras, a quantidade de memória exigida para os cálculos destes modelos pode facilmente ultrapassar a memória física disponível no sistema. Por conta disso, o MultiMOL permite que os cálculos da validação cruzada sejam realizados tanto em memória (opção padrão) quanto através do armazenamento dos modelos intermediários em arquivos temporários no disco rígido. Esta última opção permite que sejam processados conjuntos de dados que, de outra maneira, exigiriam recursos de

hardware que talvez não estivessem disponíveis. O processamento realizado em disco faz com que um único modelo esteja carregado na memória de cada vez e, ao final dos cálculos, ele é armazenado no disco em um arquivo temporário, que será carregado em memória quando for necessário recuperar e atualizar este modelo. Devido ao grande número de operações de leitura e escrita de disco ("I/O") envolvidas neste procedimento, o desempenho do programa diminui em relação à opção padrão de cálculos realizados totalmente na memória. No entanto, o armazenamento temporário em disco permite que sejam processados conjuntos de dados muito maiores, o que incrementa grandemente a robustez do programa e aumenta o seu leque de aplicações possíveis.

Os arquivos de saída ("output") são também escritos em ASCII, como os de entrada, de modo a facilitar a sua interpretação pelo usuário. Têm o formato "exit_AAAAMMDD_HHMMSS.txt", onde "AAAA" são os quatro dígitos do ano, "MM" são os dois dígitos do mês (de 01 a 12), "DD" são os dois dígitos do dia dentro do mês, "HH" são os dois dígitos da hora (em formato 24 horas), "MM" são os dois dígitos dos minutos e, "SS", os dois dígitos dos segundos. Estes dados são obtidos da data e hora atuais do sistema.

Para os arquivos de saída, foram levadas em consideração as informações sobre a construção do modelo mais relevantes, como as informações escolhidas para o modelo (pré-processamento, tipo de regressão, utilização de validação com série de teste, etc.), os resultados estatísticos obtidos (R^2, Q^2, F-Test, etc.) e os valores de predição obtidos pelo modelo. No caso em que seja reservado um subconjunto dos dados para a validação com série de teste, esta informação é também exibida no "output" do programa, em local específico e identificado como tal. Um exemplo[2] destes arquivos de saída pode ser visto a seguir:

[*informações sobre o modelo; contém as opções que foram selecionadas pelo usuário para a regressão*]

```
########################
Informações do Modelo:
Removeu as colunas com variância igual a zero: NÃO.
Centrou os dados na Média: NÃO.
Escalonou os dados: SIM.
Aplicou o APS: NÃO.
```

[2] Dados deste modelo: "conjunto A" (QSAR "2.5D"), PCR limitado a dez componentes principais, dados escalonados mas não centrados na média, LOO-FCV para calibração e divisão sistemática do conjunto original, reservando-se uma amostra a cada 4 (quatro) para fazer a validação com conjunto de teste.

```
Usou intercept: NÃO.
Deixou amostras para validação com série de teste: SIM.
Algoritmo de Regressão: PCR.
########################
```

[*tempo total de processamento, seguido dos resultados da validação-cruzada; indica os valores de SEP e de Q² para cada número de variáveis latentes que foi utilizado na obtenção do melhor modelo*]

```
Processamento concluído!
Tempo total de processamento: 16.406 segundos.

Resultados Gerais:
```

Número PCs	SEP	Q^2
1	2.15604886	0.07832135463
2	1.529422274	0.5362145174
3	1.520649568	0.5415197692
4	1.534585876	0.5330775865
5	1.56107862	0.5168167394
6	1.559711718	0.5176625331
7	1.551674575	0.5226206637
8	1.304893407	0.6623922545
9	1.302979109	0.663382079
10	1.311074151	0.6591864617

[*resultados do melhor modelo obtido pela validação cruzada*]

```
Melhor modelo construido:
#PCs: 9
#Amostras: 86
PRESS: 119.4265143
SEP: 1.302979109
Q^2: 0.663382079
R^2: 0.7246629633
F: 2.097500539
s: 1.253555545
sd: 2.2589603
```

[*tabela que mostra o resultado da validação interna, LOO-FCV*]

Amostra	V. Obs.	V. Pred.	Res. Pred.	V. Cal.	Res. Cal.
MOL_07	6.1100001	6.7633101	-0.6533099	6.7355530	-0.6255529
MOL_08	9.0000000	6.6622025	2.3377975	7.0814386	1.9185614
MOL_09	7.6399999	5.3034915	2.3365084	5.4133345	2.2266653
MOL_14	7.3099999	6.3525666	0.9574334	6.3981046	0.9118953
[...]					
SQ29852_2H	7.1900001	9.5244980	-2.3344979	9.6322859	-2.4422858
THIOL_22	8.7700005	5.8540369	2.9159636	5.6877070	3.0822934
THIOL_28	9.6400003	6.7031399	2.9368605	6.4624844	3.1775159
sd:	2.2835044	1.8650802	2.7871960	1.9018282	2.8006006

[*dados da validação com série de teste]*

```
Melhor modelo construido - Dados validacao com serie de teste:
#Amostras: 28
PRESS: 73.19038
SEP: 1.616769
R^2: 0.5118976
F: 2.097501
```

[*tabela que mostra o resultado da validação com série de teste*]

Amostra	V. Obs.	V. Pred.	Diferenca.
MOL_12	7.3099999	6.2783344	1.0316655
MOL_18	8.9200001	6.8706991	2.0493010
MOL_22	9.2200003	7.9456324	1.2743678
MOL_29	8.1499996	7.0729399	1.0770598
[...]			
MOL_67	5.0799999	6.7033762	-1.6233763
SQ29852_2P	6.4699998	9.2128233	-2.7428235
THIOL_12	3.5899999	5.4845326	-1.8945327
sd:	2.331705	1.724782	1.599726

Como pode ser visto, no cabeçalho do arquivo de saída encontram-se as informações gerais sobre todas as opções que foram utilizadas na construção do modelo. Os principais parâmetros que foram selecionadas pelo usuário – tais como pré-processamento dos dados, remoção de colunas com variância igual a zero, divisão do conjunto de entrada, etc. – podem ser encontrados nesta área do arquivo de saída gerado pelo programa. Em seguida, são apresentados os resultados do processo de calibração do modelo, em uma tabela na qual são exibidas, em número crescente, as variáveis latentes, com seus respectivos valores de SEP e de Q^2. Note-se que, conforme dito anteriormente, o número de variáveis latentes que tenha menor valor de SEP é aquele escolhido pelo programa como sendo o melhor modelo encontrado. Em seguida, após o programa calcular e prever os dados no modelo que tenha este número

ótimo de variáveis latentes, são apresentados os dados deste modelo (auto-predição): o valor do coeficiente de determinação, de PRESS, do Teste F, entre outros. É também apresentada uma tabela que resume a etapa de construção do modelo, contendo o valor observado, predito (na calibração – no caso, LOO-FCV) e calculado (após a identificação do melhor modelo: auto-predição) para cada uma das amostras que fazem parte do conjunto original. Por fim, caso exista um conjunto de validação como série de teste, é apresentado um pequeno cabeçalho com informações estatísticas sobre este conjunto (contendo o seu número de amostras, o valor do PRESS, do SEP, do R^2 e do Teste F), seguido de uma tabela. Esta é análoga àquela exibida para a validação interna, contendo os valores observados e preditos para estas amostras que ficaram de fora do cálculo do modelo.

O paradigma de linguagem de programação escolhido, orientado a objetos, facilita a modelagem do problema. No MultiMOL existem cerca de 15 (quinze) classes. Em linguagens de programação orientadas a objeto, “classes” designam um conjunto de linhas de código que modelam um determinado comportamento desejado pelo desenvolvedor do software. No caso específico do MultiMOL, por exemplo, o comportamento “executar regressão linear simples” pode ser efetivamente modelado como sendo uma classe que executa esta operação, a qual pode ser utilizada em diversos pontos do programa, conforme seja necessário. No contexto do presente trabalho, é este o sentido no qual o termo é aplicado e deve ser entendido, não sendo aplicável o conceito clássico da estatística e/ou da quimiometria, no qual “classes” definem agrupamentos de objetos que possuam algumas propriedades de interesse em comum. As classes do MultiMOL foram modeladas para a execução dos algoritmos utilizados pelo programa (MLR, PCR, NIPALS, PLS, etc), para a geração de arquivos de saída, para o cálculo das variáveis estatísticas, para a representação dos componentes gráficos, etc. O conceito de herança foi utilizado para a otimização do código-fonte, permitindo a reutilização de procedimentos e evitando a duplicidade de linhas de código. Desta forma, por exemplo, a regressão linear múltipla (MLR) pode ser usada como ferramenta de regressão aplicada diretamente nos dados de entrada, ou pode ser utilizada nas componentes principais obtidas destes dados de entrada, para fazer uma regressão em componentes principais (PCR). A mesma implementação do algoritmo é utilizada em ambas, eliminando a redundância. Durante o trabalho de desenvolvimento do software foi também levada em consideração a modularização do código, pois ela facilita o trabalho de manutenção futura e possibilita o reuso dos algoritmos, como foi explicado acima. O conceito, bastante utilizado em linguagens de programação, tem um significado bastante intuitivo: um programa está estruturado em “módulos” quando ele é composto por diversas partes, relativamente autônomas, que se comunicam para a execução de procedimentos mais complexos. Assim, por exemplo, o MultiMOL possui um módulo de escrita de arquivos de saída (“output”), que em si é independente dos algoritmos de cálculo do modelo, encarregando-se somente de exibir os dados de uma maneira tal, em uma ordem tal, etc. – como foi mostrado no exemplo de output

anteriormente mostrado. Isso possibilita que o formato do arquivo de saída seja sempre o mesmo, qualquer que seja o algoritmo de regressão utilizado.

O resultado de tudo isso é, enfim, um programa baseado em um código-fonte de considerável complexidade, robusto e bem estruturado, planejado e implementado de maneira modularizada e otimizada, que é capaz de produzir os resultados que serão apresentados nesta seção.

Segue um exemplo prático de como as características do código do MultiMOL, acima apresentadas, comportam-se em uma utilização típica do programa. Suponha-se um usuário que tenha em mãos, em uma planilha do Excel, uma série de moléculas com conhecida atividade inibidora de enzimas ACE, tendo calculado uma série de descritores para cada molécula da série e possuindo também, ao final, a atividade inibitória expressa em termos de IC_{50}[3], e deseje construir um modelo QSAR para tentar explicar a atividade biológica destas moléculas em termos dos descritores que ele selecionou. Estes dados estão praticamente prontos para serem utilizados pelo MultiMOL, bastando para isso exportá-los no formato .csv e acrescentar, na primeira linha do arquivo, as dimensões da matriz (número de amostras e número de descritores).

Por meio da interface do programa, o usuário seleciona, do seu computador, o arquivo .csv que contém os dados com os quais quer trabalhar, que serão carregados e exibidos em uma tabela do programa. Internamente, esta tabela é representada como uma matriz de ‘m’ linhas por ‘n’ colunas. Sabendo que os seus descritores representam variáveis de diferentes grandezas e unidades de medida, e querendo evitar que um determinado descritor tenha um peso maior na construção do modelo de regressão unicamente por estar expresso em uma variável numericamente maior, o usuário escolhe a opção de escalonar os dados. Sabendo que estas amostras estão homogeneamente dispostas ao longo do seu conjunto e querendo reservar uma parcela dele para fazer a validação do modelo com uma série de teste, o usuário decide fazer uma separação sistemática do conjunto original, reservando seqüencialmente uma amostra a cada quatro para compor o conjunto de validação. Seleciona, assim, esta opção na interface gráfica.

Por se tratar de um conjunto relativamente pequeno de amostras, o usuário acredita que o seu computador possui memória suficiente para realizar as operações necessárias e, portanto, marca a opção de “Usar a Memória” para os cálculos. O algoritmo de regressão apresentado por “default” (opção padrão) é o PCR e, neste primeiro teste, o usuário não o modifica. Aperta o botão para executar o cálculo do modelo.

---

[3] Trata-se precisamente do “Conjunto A” exposto na seção 3.2.2., o qual foi também usado para a obtenção do “output” exibido na seção 4.1.

O programa, então, lê todos os dados informados na interface gráfica e os repassa para as classes[4] nas quais estão implementados os algoritmos de regressão, que realizam a calibração do modelo[5]. Após os procedimentos de Validação Cruzada "Leave-One-Out", é identificado o número ótimo de componentes principais, com o qual é construído o modelo final. Neste, são projetadas as amostras que ficaram de fora da calibração do modelo, e os seus valores estatísticos são calculados. Estes cálculos são todos realizados com precisão e bom desempenho computacional. Os resultados obtidos são armazenados em um arquivo de saída ("output"), em cujo nome está o dia e a hora no qual foi executado o MultiMOL.

O usuário acompanha, no console da interface gráfica, o procedimento de construção do modelo. Ao final, são exibidos na tela dois gráficos, entre os quais o usuário pode alternar clicando nas abas que o programa oferece: um que mostra os valores "observados e preditos" tanto para o conjunto de calibração quanto para o de validação; e, outro, que mostra o valor do Q^2 obtido para cada número de componentes principais durante o processo de calibração. Após analisar estes resultados, o usuário abre o arquivo de saída para ver mais detalhes do modelo que obteve, em um editor de texto padrão.

Caso deseje, por exemplo, comparar este modelo PCR com outro, PLS, o usuário pode simplesmente alterar o tipo de algoritmo de regressão na interface gráfica, mantendo inalteradas todas as outras opções. Após todos os cálculos, serão exibidos na tela os novos gráficos referentes ao modelo construído com o PLS. O novo arquivo "output" terá um nome diferente por causa do horário diferente em que o programa foi rodado, desta forma, não sobrescrevendo o resultado anterior e possibilitando, assim, que estes resultados sejam recuperados facilmente no futuro. Os dois modelos poderão também ser facilmente comparados com base nos arquivos de saída, uma vez que o formato de ambos será exatamente o mesmo, variando apenas os valores numéricos gravados em cada um.

4.2. A Interface Gráfica (GUI)

A parte gráfica do MultiMOL foi desenvolvida com o duplo objetivo de facilitar a utilização do software pelo usuário e, ao mesmo tempo, oferecer novas funcionalidades capazes de aumentar a aplicabilidade do programa. É por meio dela que se realiza toda a interação entre o usuário e o programa, desde a importação/edição dos dados até a visualização dos resultados gerados pelo programa.

---

[4] "Classes", vale lembrar, no contexto de linguagens de programação.

[5] Este é outro exemplo da aplicabilidade da modularização do programa: os parâmetros do modelo são aqui obtidos por meio da interface gráfica, mas para as classes que executam os procedimentos numéricos é indiferente de onde os parâmetros venham: da execução do programa em "linha de comando", com os parâmetros digitados manualmente pelo usuário, ou de maneira automática, através da interface gráfica (GUI).

As figuras 4-a, 4-b e 4-c mostram as telas principais do MultiMOL. São as três “abas” do programa, onde o usuário pode visualizar e/ou editar os dados de entrada do programa, bem como obter os resultados dos modelos de regressão. A primeira aba consiste em uma tabela ou planilha, aos moldes dos softwares mais populares de estatística multivariada, na qual as linhas representam as amostras (moléculas, em QSAR) e, as colunas, representam as variáveis (descritores, em QSAR). Ao lado esquerdo de cada linha, encontra-se o nome ("label") da amostra correspondente; acima de cada coluna, encontra-se o "label" da variável correspondente. A última coluna armazena os valores da função-resposta ou variável dependente (atividade biológica, em QSAR); as colunas subseqüentes são as variáveis independentes que serão utilizadas para a construção do modelo de regressão.

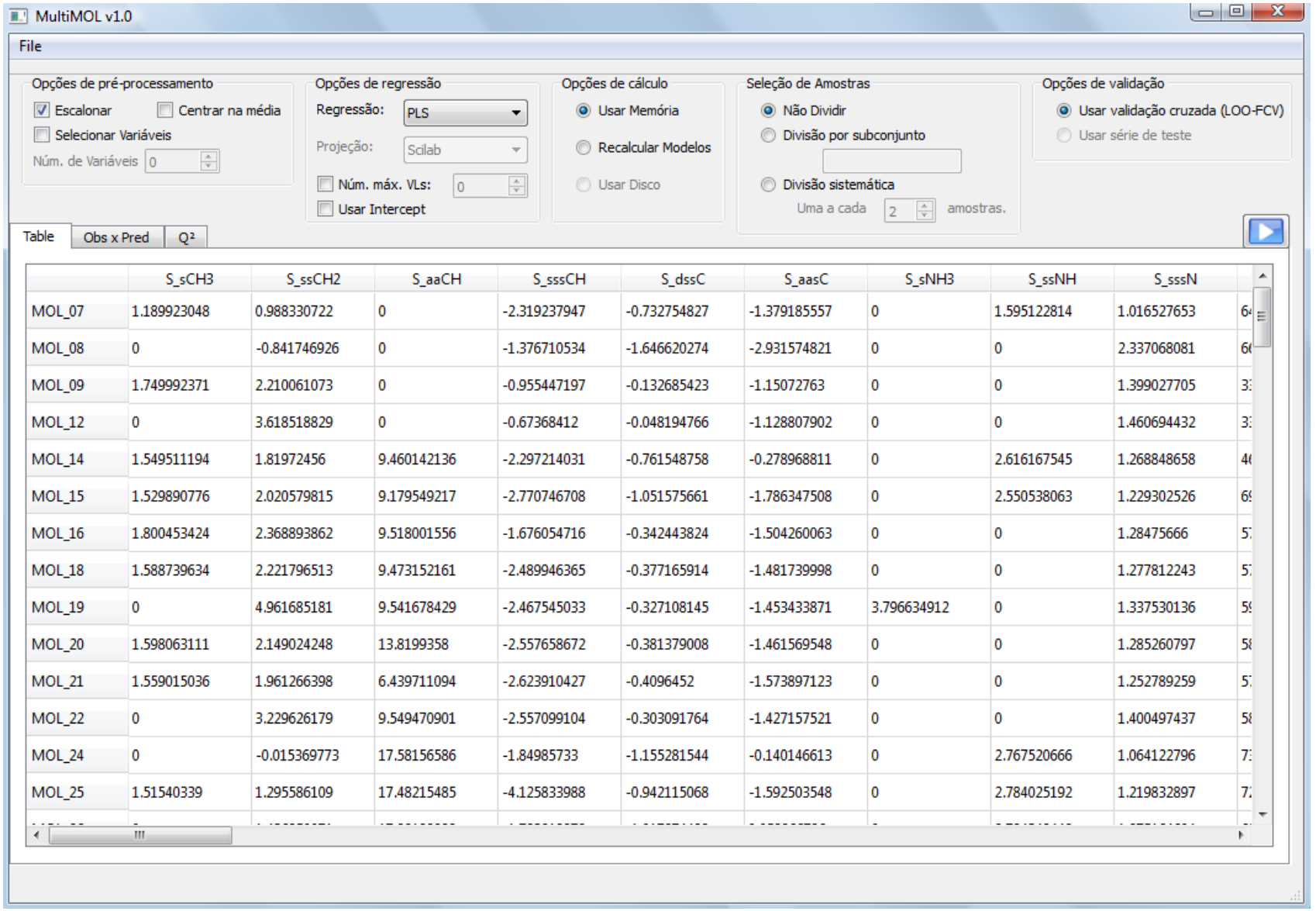
Na segunda aba, podem ser visualizados os dados mais importantes do modelo que foi construído. No lado direito, existe um console através do qual o usuário pode acompanhar o andamento do cálculo, durante o procedimento de calibração do modelo. As principais informações sobre o que o programa está fazendo no momento são exibidas neste console. Do lado esquerdo, há um gráfico onde, após o término do cálculo do modelo de regressão, são apresentados os valores observados e os preditos para todas as amostras que constam na tabela da aba anterior. Os dados que foram utilizados para calibrar o modelo (o conjunto de calibração) são apresentados com a cor vermelha; os dados externos (o conjunto de validação) são exibidos com a cor azul. O gráfico possui opções de escala para ambos os eixos, bem como opções de customização (tamanho, formato) dos pontos que são nele exibidos. A linha diagonal que aparece na Figura 4-b, ligando o canto inferior esquerdo do gráfico ao superior direito, é opcional e pode ser exibida caso a opção correspondente seja marcada pelo usuário. Existe ainda a opção de exportar a imagem no formato “.png”[6], de modo a salvá-la no computador e, posteriormente, recuperá-la com qualquer programa de edição de imagens.

Por fim, na terceira aba (Figura 4-c), são apresentados os resultados da calibração do modelo. O gráfico representa o valor do Q^2 para cada um dos números de componentes principais que foi utilizado na validação cruzada ou validação com série de teste. As mesmas opções de customização que existem para o gráfico de valores observados por valores preditos (Figura 4-b) estão também disponíveis aqui: assim, é possível alterar a escala dos dois eixos, a figura que representa cada um dos pontos do gráfico e também salvar a imagem no computador. Naturalmente, neste gráfico não existe a opção de exibir a diagonal, uma vez que ela não faz sentido para o tipo de informação nele representada.

Após o cálculo da regressão, o MultiMOL exibe na tela da GUI os resultados mais importantes do modelo. São apresentados os valores do coeficiente de correlação

[6] “Portable Network Graphics”, um formato de dados para imagens amplamente utilizado na internet. Como o .jpg, é perfeitamente reconhecível pela quase totalidade dos softwares visualizadores / editores de imagem.

da autopredição (R^2) e do coeficiente de correlação da validação (Q^2 ou R^2pred, dependendo da opção escolhida), bem como o tempo total de processamento. Há dois gráficos que são imediatamente carregados pelo programa após o término do processamento: i) “Observado x Predito”, que mostra, para cada uma das amostras, a diferença entre a função-resposta (informada no conjunto de dados de entrada) e o valor previsto pelo modelo de regressão; ii) “Q^2”, que mostra a evolução dos valores do coeficiente de correlação da validação em função do número de variáveis latentes utilizadas. Igualmente, é gerado um arquivo de saída (“output”) no formato TXT (ASCII), com informações mais detalhadas, ou seja, as opções utilizadas para a regressão, os valores observados, preditos e calculados para cada uma das amostras, o tempo de processamento, além de outras informações estatísticas importantes do modelo (“Standard Error of Prediction” – SEP, “Prediction Error Sum of Squares” – PRESS, “Standard Deviation” – SD, etc.).



	S_sCH3	S_ssCH2	S_aaCH	S_sssCH	S_dssC	S_aasC	S_sNH3	S_ssNH	S_sssN
MOL_07	1.189923048	0.988330722	0	-2.319237947	-0.732754827	-1.379185557	0	1.595122814	1.016527653
MOL_08	0	-0.841746926	0	-1.376710534	-1.646620274	-2.931574821	0	0	2.337068081
MOL_09	1.749992371	2.210061073	0	-0.955447197	-0.132685423	-1.15072763	0	0	1.399027705
MOL_12	0	3.618518829	0	-0.67368412	-0.048194766	-1.128807902	0	0	1.460694432
MOL_14	1.549511194	1.81972456	9.460142136	-2.297214031	-0.761548758	-0.278968811	0	2.616167545	1.268848658
MOL_15	1.529890776	2.020579815	9.179549217	-2.770746708	-1.051575661	-1.786347508	0	2.550538063	1.229302526
MOL_16	1.800453424	2.368893862	9.518001556	-1.676054716	-0.342443824	-1.504260063	0	0	1.28475666
MOL_18	1.588739634	2.221796513	9.473152161	-2.489946365	-0.377165914	-1.481739998	0	0	1.277812243
MOL_19	0	4.961685181	9.541678429	-2.467545033	-0.327108145	-1.453433871	3.796634912	0	1.337530136
MOL_20	1.598063111	2.149024248	13.8199358	-2.557658672	-0.381379008	-1.461569548	0	0	1.285260797
MOL_21	1.559015036	1.961266398	6.439711094	-2.623910427	-0.4096452	-1.573897123	0	0	1.252789259
MOL_22	0	3.229626179	9.549470901	-2.557099104	-0.303091764	-1.427157521	0	0	1.400497437
MOL_24	0	-0.015369773	17.58156586	-1.84985733	-1.155281544	-0.140146613	0	2.767520666	1.064122796
MOL_25	1.51540339	1.295586109	17.48215485	-4.125833988	-0.942115068	-1.592503548	0	2.784025192	1.219832897

Figura 4-a: Tela principal do MultiMOL.

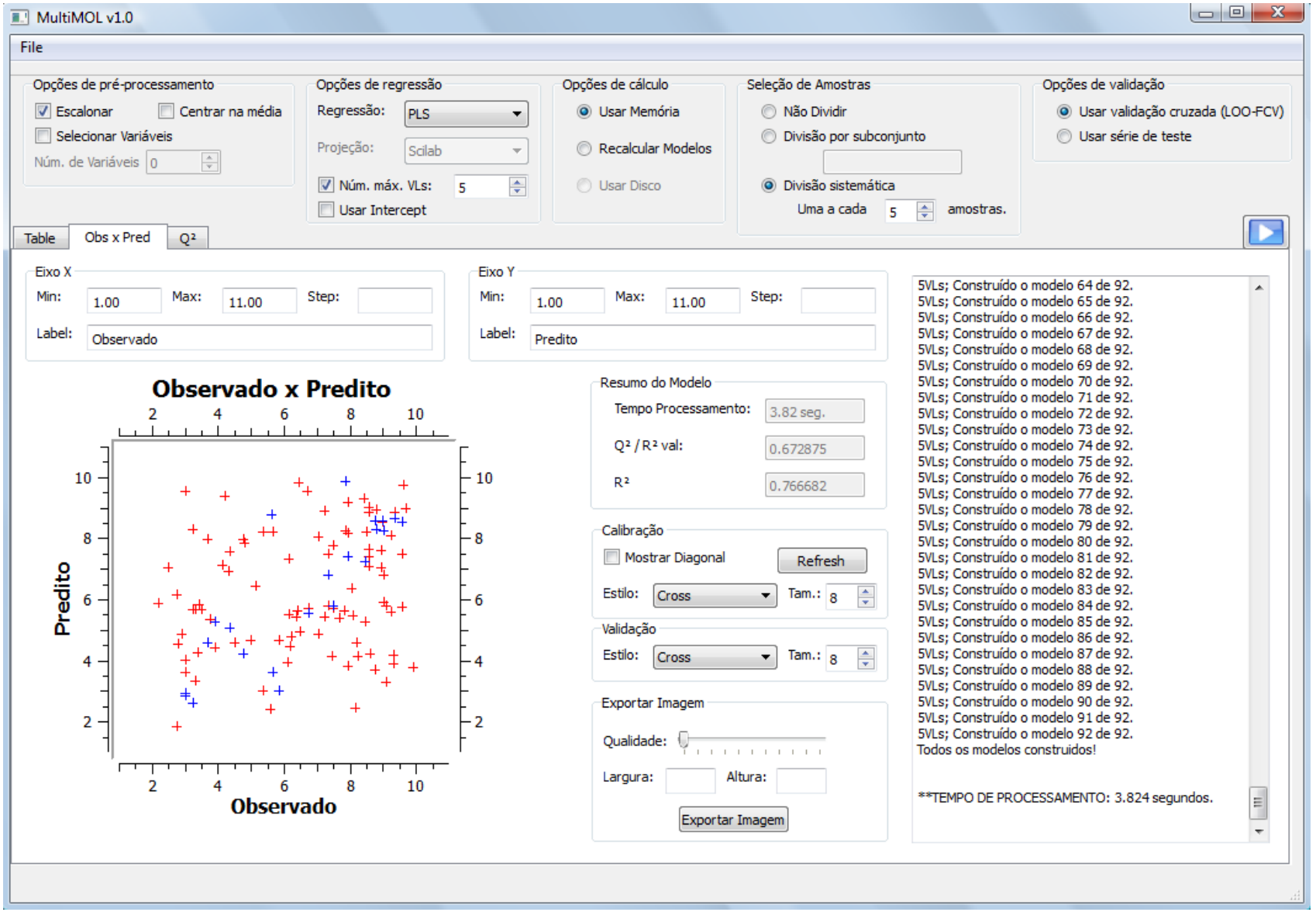


Figura 4-b: Tela principal do MultiMOL.

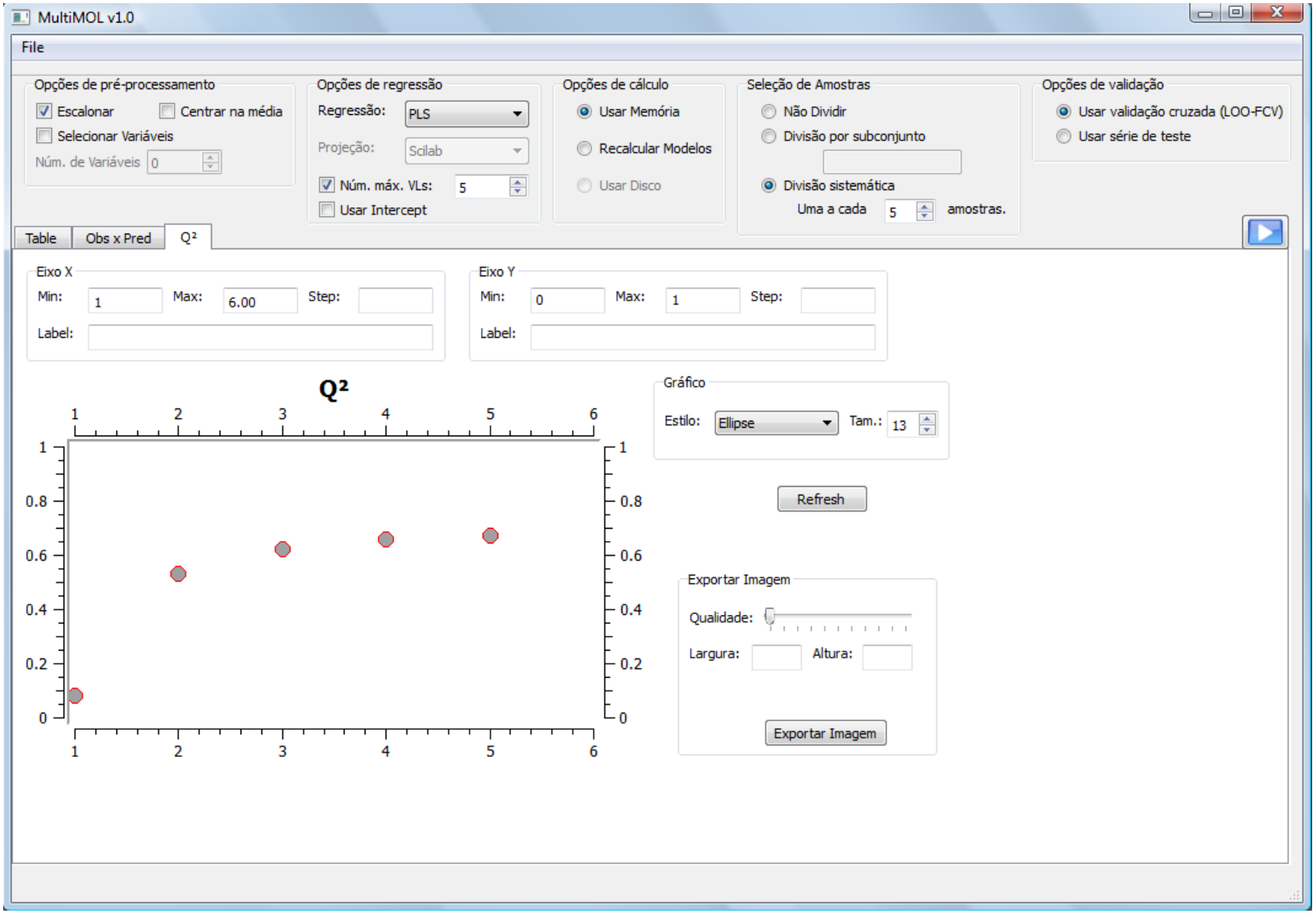


Figura 4-c: Tela principal do MultiMOL.

Os dados podem ser importados de um arquivo no formato csv (“comma separated value”), que possua o layout que já foi descrito anteriormente (v. seção 3.2.3). Uma vez importado o arquivo, os dados podem ser editados pelo usuário: os valores numéricos podem ser alterados, podem ser incluídas novas linhas ou novas colunas, e é possível remover linhas ou colunas existentes. Após a realização de todas as alterações necessárias, é possível salvar os dados editados, a fim de reutilizá-los no futuro, utilizando-se o menu “File”.

Por meio da parte superior da tela principal do programa (que pode ser vista na Figura -4-a, na Figura 4-b e na Figura 4-c, pois é a mesma para todas as abas do software), o usuário tem acesso a todas as opções de regressão que o MultiMOL oferece. Do lado superior esquerdo estão as opções de pré-processamento: o usuário pode selecionar se deseja escalonar os dados, centrá-los na média ou aplicar o algoritmo de seleção de variáveis. Em seguida (da esquerda para a direita), estão as opções de algoritmos de regressão: o usuário escolhe se deseja utilizar MLR, PCR, PLS ou Q-PLS. Ele também pode selecionar se deseja limitar o número máximo de variáveis latentes (componentes principais) a algum valor pré-estabelecido; caso ele o faça, o processo de calibração só irá construir modelos com, no máximo, o número informado. Caso esta opção não seja marcada, o programa vai construir todos os modelos de calibração, até que a variância da matriz original esteja esgotada - número que, conforme já exposto, é limitado pelo menor número entre o de objetos e o de descritores [FERREIRA et al., 1999].

Para MLR e PCR, é também possível marcar a opção de usar o “intercept”, que consiste em incluir uma constante na equação de regressão, forçando a reta de regressão a passar pela origem dos eixos. Isto foi feito acrescentando-se uma coluna (variável independente) com valores unitários na matriz de dados original.

Em seguida, é possível escolher a forma que o programa utilizará para a calibração dos modelos. Por limitações de tempo, não foi implementada a opção de salvar, em arquivos temporários, os modelos intermediários gerados pelo programa: estão somente disponíveis as opções de armazená-los na memória ou refazer, para cada número de componentes principais, os cálculos de todos os modelos precedentes.

É também possível escolher as opções de divisão de conjunto de dados. O usuário, então, escolhe se não deseja dividir os arquivos ou, caso o deseje, se quer fazê-lo informando expressamente quais são as amostras que deseja deixar de fora do processo de calibração ou se quer deixar o programa reservar, automaticamente, uma amostra a cada ‘x’ (valor que ela informa) para compor o conjunto externo. Por fim, o usuário escolhe também a forma que será utilizada pelo programa para fazer a calibração do modelo: pode optar por fazer uma validação cruzada (LOO-FCV) ou por utilizar o conjunto externo para otimizar o número de variáveis latentes (validação com série de teste – v. seção 3.2.10).

Após selecionar todas as opções que deseja, o usuário simplesmente clica no botão “run” (canto superior direito) e, então, o MultiMOL vai construir o modelo com os parâmetros selecionados. Durante os cálculos, é apresentada uma barra de progresso indicando que o programa está em execução, impedindo que o usuário submeta novamente um cálculo que já se encontra em processamento. Também durante a execução do programa, como já foi dito, o console da segunda aba da tela principal do MultiMOL informa o andamento do cálculo. Por fim, após o seu término, a interface gráfica carrega automaticamente os resultados obtidos na segunda e na terceira aba, os quais podem ser facilmente visualizados pelo usuário. A partir daqui, a exibição dos gráficos pode ser customizada e os mesmos podem ser salvos no computador, como já foi explicado.

Com estas facilidades, espera-se oferecer ao usuário as opções mais comuns de visualização e manipulação de dados numéricos, aumentando a usabilidade do programa e tornando-o mais amigável (“user-friendly”[7]).

4.3. Resultados dos testes

Para testar a qualidade e a robustez do código do MultiMOL, foram realizados vários testes, cujos resultados serão apresentados a seguir.

4.3.1. Conjunto ‘A’: QSAR Tradicional

O primeiro conjunto de dados testado consiste em uma série de 114 amostras para as quais foram utilizados 56 descritores de QSAR Tradicional. Há claramente a necessidade de que os dados sejam pré-processados, escalonando-os, uma vez que os descritores representam variáveis de grandezas distintas (unidades de medida diferentes) e, portanto, estão também mensurados em escalas distintas. Para os testes realizados neste conjunto, utilizou-se o auto-escalonamento da matriz de dados original.

O conjunto de dados apresenta um número de amostras (114) maior do que o número de escritores (56) e, por isso, é possível obter um modelo por meio de uma regressão MLR simples. Os resultados “observados x preditos” obtidos por este modelo estão expostos na figura 5.

[7] A expressão é comum na área de ciência da computação e designa um software cuja interface gráfica é construída de tal maneira que facilita a interação do usuário com o programa. Há toda uma área, na informática, associada ao estudo de interfaces e de usabilidade de software.

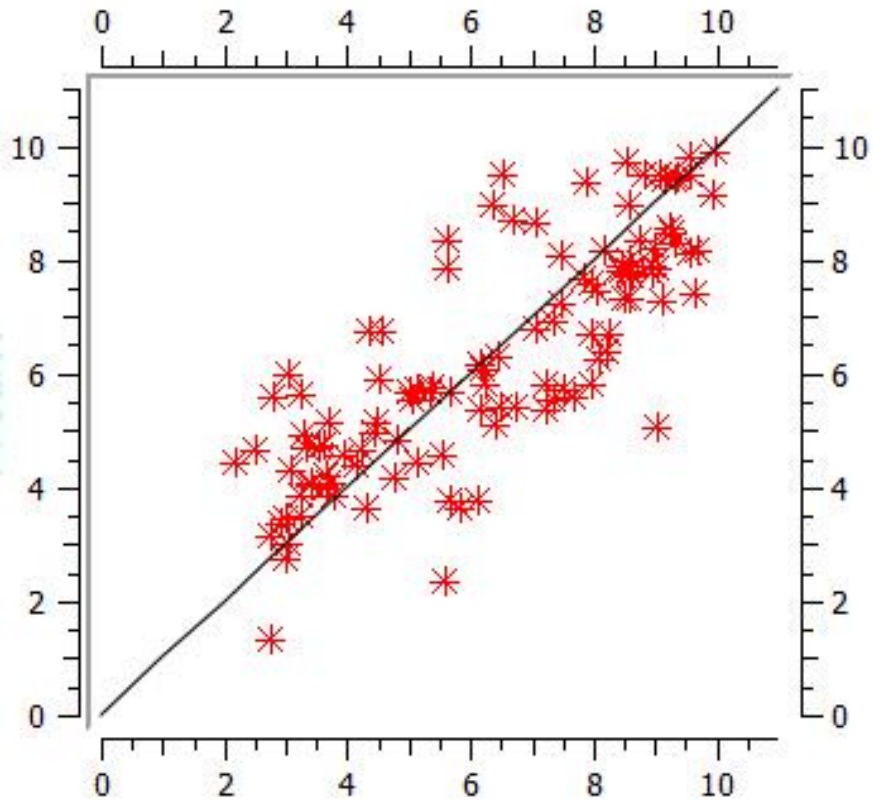


Figura 5: MLR aplicado sobre matriz auto-escalonada, conjunto A

Este modelo possui as seguintes características:

Número de variáveis selecionadas: 18
Número de Amostras: 114
PRESS: 456.1206029
SEP: 1.350359817
Q^2: 0.6459049577
R^2: 0.223045346

Percebe-se que o modelo obtido apresenta uma boa explicabilidade dos dados analisados; com as primeiras dezoito variáveis, obtém-se um Q^2 (obtido com Validação-Cruzada “LOO-FCV”) de 0,65. A figura 6 mostra a evolução do valor de Q^2 para as primeiras quinze variáveis.

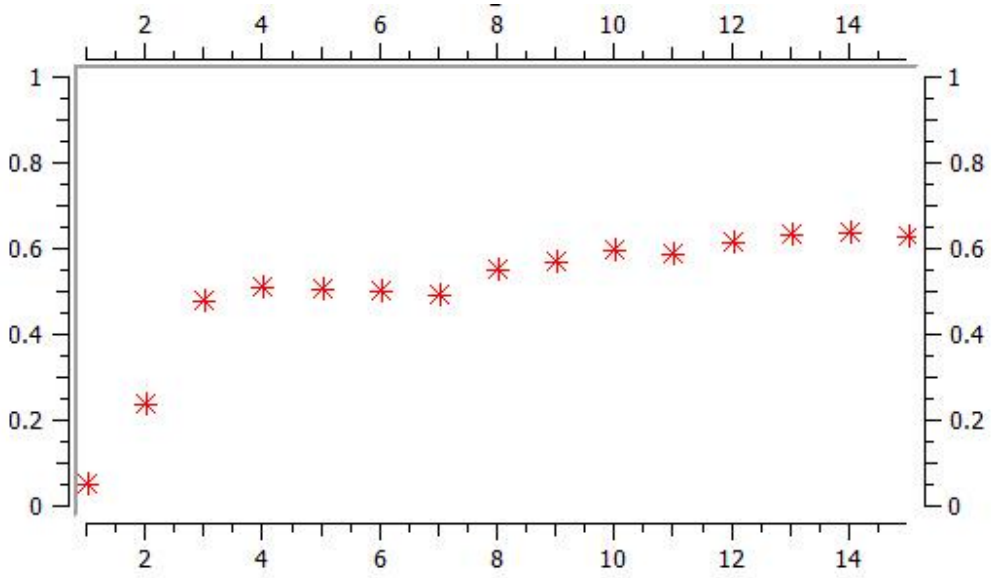


Figura 6: Valores de Q² para as primeiras quinze variáveis utilizadas do Conjunto A.

Embora as ferramentas da estatística utilizadas sejam comum a ambos, é importante diferenciar um problema da Química Medicinal (em particular, QSAR) de outro da Química Analítica. Este exemplo é de QSAR; nesta metodologia, está-se geralmente interessado em resultados de certo modo qualitativos. Mais importante do que saber exatamente o quanto cada um dos descritores contribui para a propriedade de interesse analisada, estamos interessados em identificar quais são os descritores (as propriedades, no caso do QSAR tradicional, ou os pontos da molécula no caso de QSAR-3D) que estão relacionados com a função-resposta. Isto acontece porque é já bastante relevante uma conclusão que identifique (p.ex.) quais as áreas da molécula estudada que estão mais relacionadas com a atividade biológica do fármaco, mesmo que não se saiba o valor quantitativo preciso desta contribuição.

Ao contrário, na Química Analítica (um exemplo da qual pode ser visto na seção 4.3.3., mais adiante), os compostos que fazem parte de uma mistura já são geralmente conhecidos *a priori*. O que se busca com a construção de modelos de regressão, neste caso, é identificar a concentração exata de cada um deles e, portanto, os critérios quantitativos são aqui muito mais rigorosos. É por isso que um valor de Q² de 0,65 como o apresentado é considerado bom para um problema de QSAR, mas no entanto seria inaceitável em um problema da Química Analítica. Como será visto mais adiante, os números obtidos em problemas que envolvam espectroscopia são diferentes dos obtidos na construção de modelos de QSAR.

Um dos problemas com a regressão linear múltipla, como já foi dito, é que a qualidade dos modelos de regressão está diretamente relacionada com a relevância dos descritores que foram selecionados para o conjunto de calibração. O MultiMOL não oferece nenhum algoritmo para otimizar a escolha dessas variáveis, de modo que a

otimização do número de variáveis é feita de maneira sistemática. O modelo que tem uma variável é aquele que utiliza a primeira variável; o modelo com duas variáveis, é aquele que usa as duas primeiras variáveis. Se houvesse, digamos, uma combinação de variáveis particularmente explicativa das propriedades de interesse da série de moléculas analisadas (digamos, o primeiro descritor junto com o quarto), o programa não seria capaz de identificá-la: um modelo que usasse estes dois descritores iria usar também, o segundo e o terceiro.

Isto posto, é relevante enfatizar a importância de que, para os problemas de QSAR Tradicional nos quais serão aplicados modelos de regressão construídos a partir de MLR, os descritores estejam dispostos em ordem de prioridade. Isto acontece porque o programa busca o melhor modelo utilizando-se os 'p' primeiros descritores da matriz de dados original. Ou seja, o modelo com um descritor vai utilizar o primeiro descritor; o modelo com dois descritores vai utilizar o primeiro e o segundo; e assim sucessivamente.

Neste caso, com os descritores originais colocados em ordem de importância (primeiro o mais significativo, em seguida o segundo mais significativo, etc.) garante-se que a busca sistemática irá acrescentar, a cada iteração, o próximo descritor mais explicativo do fenômeno que está sendo estudado. Se os descritores estiverem dispostos aleatoriamente, a busca sistemática implementada no MultiMOL não garantirá a convergência dos modelos para aquele que é mais explicativo. Ao contrário do que acontece, por exemplo, nos modelos construídos com PCR ou PLS (onde a primeira variável latente é a mais importante na explicação do modelo), não existe nenhuma garantia de que o primeiro descritor utilizado para o MLR seja “o melhor descritor”. Assim, pode ser útil também a construção de diversos modelos, nos quais a ordem dos descritores esteja alterada.

Para este conjunto, executou-se também uma Análise de Componentes Principais antes da aplicação da MLR, ou seja, um procedimento PCR, a fim de compará-la com a regressão linear múltipla. Os mesmos parâmetros de pré-processamento e de calibração do modelo foram adotados: a matriz de dados foi auto-escalonada e, a validação, foi feita por meio de LOO-FCV. Nas Figuras 7 e 8 é possível ver, respectivamente, os gráficos de valores observados por preditos e o da evolução do Q^2 ao longo das primeiras quinze componentes principais.

A diferença na explicabilidade do modelo faz-se sentir: com a PCR, é possível obter um Q^2 de 0,70. Entretanto, não existe ganho real no número ótimo de variáveis latentes utilizadas na construção do modelo, pois este Q^2 é obtido com a utilização de dezoito componentes principais, o que é um número alto.

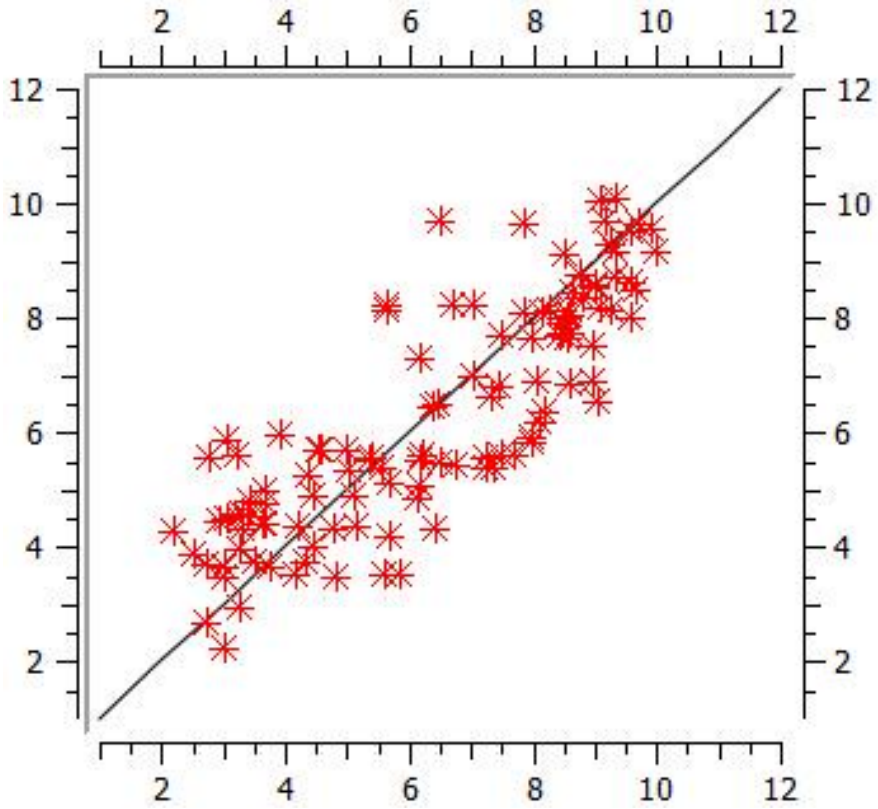


Figura 7: Valores “observados x preditos” para o conjunto A; PCR aplicada sobre conjunto de dados auto-escalonado.

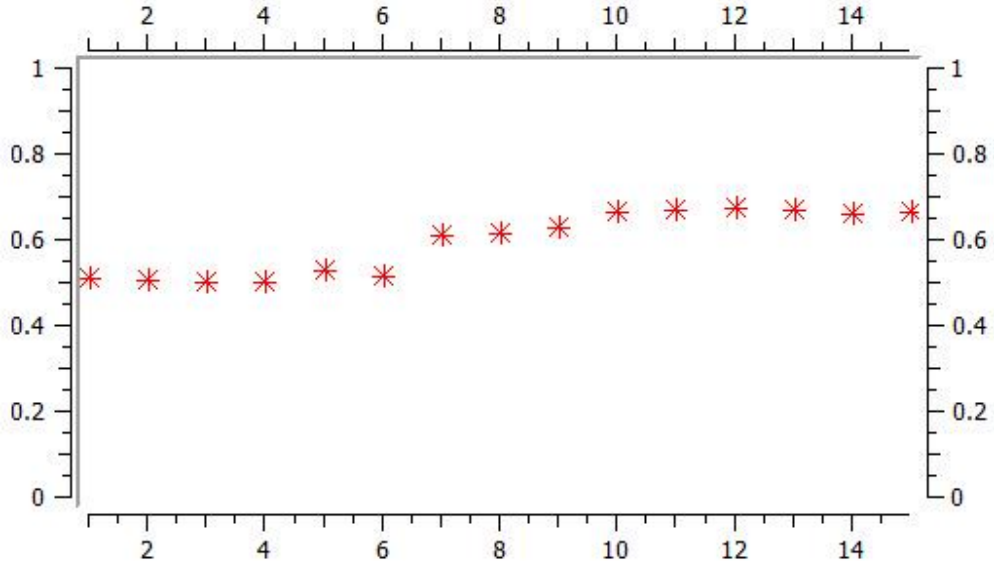


Figura 8: Evolução dos valores de Q² para as primeiras quinze componentes principais.

As características deste modelo são as seguintes:

Número de PCs: 18
Número de Amostras: 114
PRESS: 134.6991481
SEP: 1.249123409
Q^2: 0.6970077441
R^2: 0.7705538199
F: 17.72446953

Em especial, a Figura 8 mostra a vantagem de se construir modelos com variáveis que sejam combinações dos descritores originais. Enquanto que, na Figura 6, o valor do coeficiente de correlação aumenta aos saltos e oscila demasiadamente ao longo do número de variáveis utilizados, na Figura 8 nós vemos uma evolução paulatina e constante. Isso se explica porque o ganho de informação relevante quando se acrescenta a próxima componente principal é verdadeiro. Por outro lado, o simples acréscimo de um descritor aleatório pode acrescentar, ao modelo, mais ruído do que informação verdadeiramente relevante.

4.3.2. Conjunto 'B': QSAR-3D

O conjunto de QSAR-3D é composto por 31 esteróides, para cada um dos quais foi calculado o valor do campo eletrostático em 1813 pontos localizados espacialmente no interior de uma caixa discretizada. Os detalhes deste procedimento de obtenção dos descritores de QSAR-3D serão mostrados em um manuscrito que encontra-se atualmente em preparação, envolvendo o trabalho de outros integrantes do nosso grupo de pesquisa. Devido às dimensões do conjunto de entrada, não é possível realizar uma Regressão Linear Múltipla. Como é comum nos casos de QSAR-3D, é necessária a decomposição da matriz de dados original por meio de uma PCR ou do PLS.

Todos os descritores são da mesma natureza e, portanto, representam valores de mesma grandeza, de modo que não é necessário realizar o escalonamento dos dados. Para que fosse verificada a eficácia da seleção de variáveis do programa, uma Regressão em Componentes Principais foi aplicada sobre o conjunto de dados original e, depois, sobre as primeiras vinte e seis (26) variáveis selecionadas pelo software através do algoritmo APS.

A fim de conseguir melhores modelos, de ambos os modelos foram retirados do conjunto de calibração os esteróides número "1" e número "31" da série apresentada na Figura 3, pois se tratam de amostras que normalmente se comportam como "outliers" (amostras com comportamento fora do padrão) na maioria dos modelos de regressão publicados na literatura [COATS, 1998] e, portanto, que prejudicariam a calibração do modelo. Diz-se que uma determinada amostra de um conjunto é um "outlier" quando ela

tem um comportamento que se distingue do das demais amostras da série. Amostras “outliers” prejudicam a qualidade dos modelos de regressão construídos, o que torna importante a sua identificação. Os resultados estão dispostos na Figura 9.

Ao lado de resultados numericamente equivalentes, a grande vantagem de aplicação da seleção de variáveis é a economia de demanda computacional. De fato, a utilização do algoritmo para reduzir a dimensionalidade do conjunto de dados original reduz drasticamente o tempo computacional gasto pelo programa para o cálculo dos modelos. Isto mostra que (i) os descritores utilizados nos problemas de QSAR-3D são extremamente correlacionados; (ii) o algoritmo de seleção de variáveis implementado no MultiMOL é eficaz na escolha das variáveis que são as menos correlacionadas entre si e, portanto, cumpre aquilo a que se propõe; e (iii) portanto, esta redução da dimensionalidade é uma alternativa válida para a execução de cálculos onerosos sem comprometer a qualidade dos modelos de regressão obtidos.

Desta forma, o procedimento de seleção prévia de variáveis mostrou-se uma excelente maneira de reduzir a demanda computacional exigida pelos problemas de QSAR, viabilizando a execução de cálculos complexos em um tempo viável.

Os resultados obtidos por estes dois modelos pode ser visualizado na Tabela 2. Para cada um dos dois conjuntos foram aplicados os mesmos procedimentos: otimização do número de componentes principais por meio de validação cruzada, e regressão feita com PCR.

Note-se que, para ambos os modelos, o valor de R^2 é o mesmo. Além disso, foram colocadas na legenda da figura também as variáveis estatísticas de avaliação de modelos de regressão – R^2, F-test, número de componentes principais e tempo de processamento – para os dois modelos.

Pode-se notar também que o modelo construído com as 26 variáveis selecionadas pelo APS é ainda ligeiramente melhor do que o modelo construído com as variáveis originais, em todos os aspectos nos quais ambos podem ser comparados: é melhor em performance, pois o seu tempo de execução é menor, como foi dito anteriormente; é melhor em robustez, porque utiliza-se de um terço a menos de componentes principais do que o modelo que foi construído a partir dos dados originais; apresenta um melhor coeficiente de correlação de validação cruzada (0.81, contra 0,70 do modelo com todas as 1813 variáveis) e, por conseguinte, um menor erro padrão de predição (SEP), enquanto que a sua explicabilidade permanece rigorosamente à mesma, o que pode ser visto tanto pela comparação dos dois gráficos da Figura 8 quanto analisando-se os valores de R^2 dos dois modelos.

Tabela 2 – Variáveis estatísticas dos modelos QSAR construídos sem seleção de variáveis e com seleção de variáveis.

	1813 Variáveis	26 Variáveis
Tempo de Processamento	3 minutos e 9 segundos	7.5 segundos
Q^2	0,70	0,81
SEP	0,60	0,47
R^2	0,90	0,90
Número de PCs	12	8
F-Test	11,82	23,47

Estes resultados são bastante interessantes, uma vez que, por meio deles, é possível concluir que a utilização de uma seleção prévia de variáveis pode reduzir a demanda computacional exigida para o cálculo do modelo, mantendo, no entanto, a sua explicabilidade. Portanto, é possível verificar que esta abordagem de seleção de variáveis aplicada aos problemas de QSAR-3D é, ao mesmo tempo, coerente, robusta e satisfatória para a sua resolução.

Para este conjunto dos esteróides, foi também construído um modelo de regressão por meio do algoritmo PLS. Os resultados obtidos estão expostos nas Figuras 9 e 10.

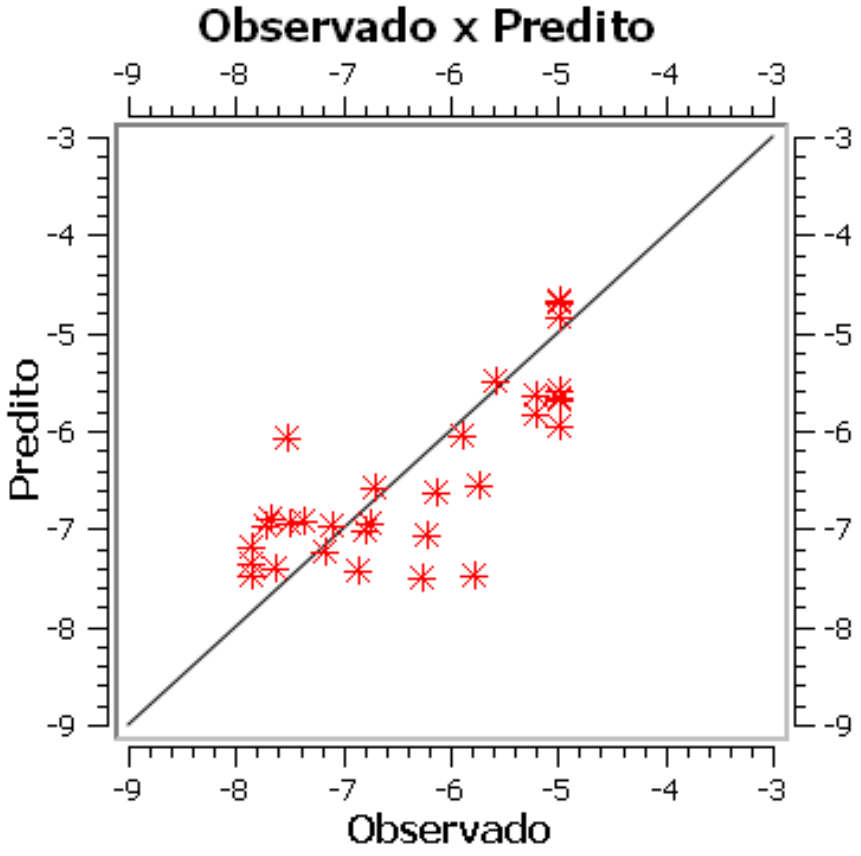


Figura 9: Gráfico “Observados x Preditos”, Conjunto ‘B’, PLS

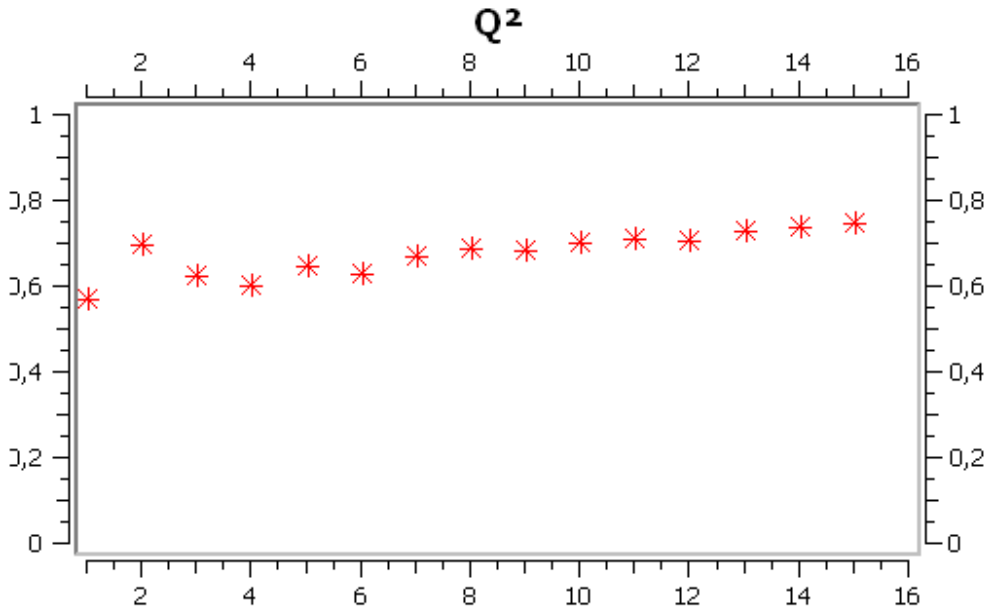


Figura 10: Gráfico de Q² para as primeiras 15 Variáveis Latentes

Nota-se que, conforme esperado, os resultados obtidos pelo algoritmo PLS são próximos daqueles obtidos por meio da Regressão em Componentes Principais (PCR). No entanto, o PLS é significativamente mais rápido do que o PCR, sendo executado, em média, em 25% do tempo. Vale também salientar que, quanto maior o número de componentes principais (variáveis latentes), maior o tempo gasto pelos algoritmos, mas este aumento de demanda computacional não é linear, uma vez que cada variável latente adicional calculada pelo PLS é obtida por meio de um algoritmo mais simples (o PLS1) do que o utilizado para adicionar uma componente principal no PCR (o NIPALS). O algoritmo PLS1 executa uma única vez para cada variável latente, conforme pode ser visto na seção 3.2.8. Já o NIPALS executa uma série de vezes para cada componente principal, buscando convergência de valores. Isto é o que justifica o melhor desempenho do primeiro.

4.3.3. Conjunto 'C': Espectro simulado

A validação do Q-PLS utilizou o conjunto de dados de espectro simulado, que foi apresentado na seção anterior. Isto foi feito para que se pudesse demonstrar a aplicabilidade do PLS Quadrático na descrição do comportamento de um conjunto de dados onde a relação existente entre as variáveis independentes (X) e a variável dependente (Y) é de natureza sabidamente não-linear. Foram realizados testes comparativos entre os métodos PLS e Q-PLS, para o conjunto de dados de espectros simulados – denominado **conjunto (C)**. O número ótimo de variáveis latentes do modelo, para ambos os casos, foi identificado automaticamente por meio de validação cruzada LOO-FCV. Podem-se observar nos gráficos da Figuras 11: os valores de Q^2 para as primeiras vinte variáveis latentes dos modelos construídos com PLS tradicional e quadrático (Q-PLS).

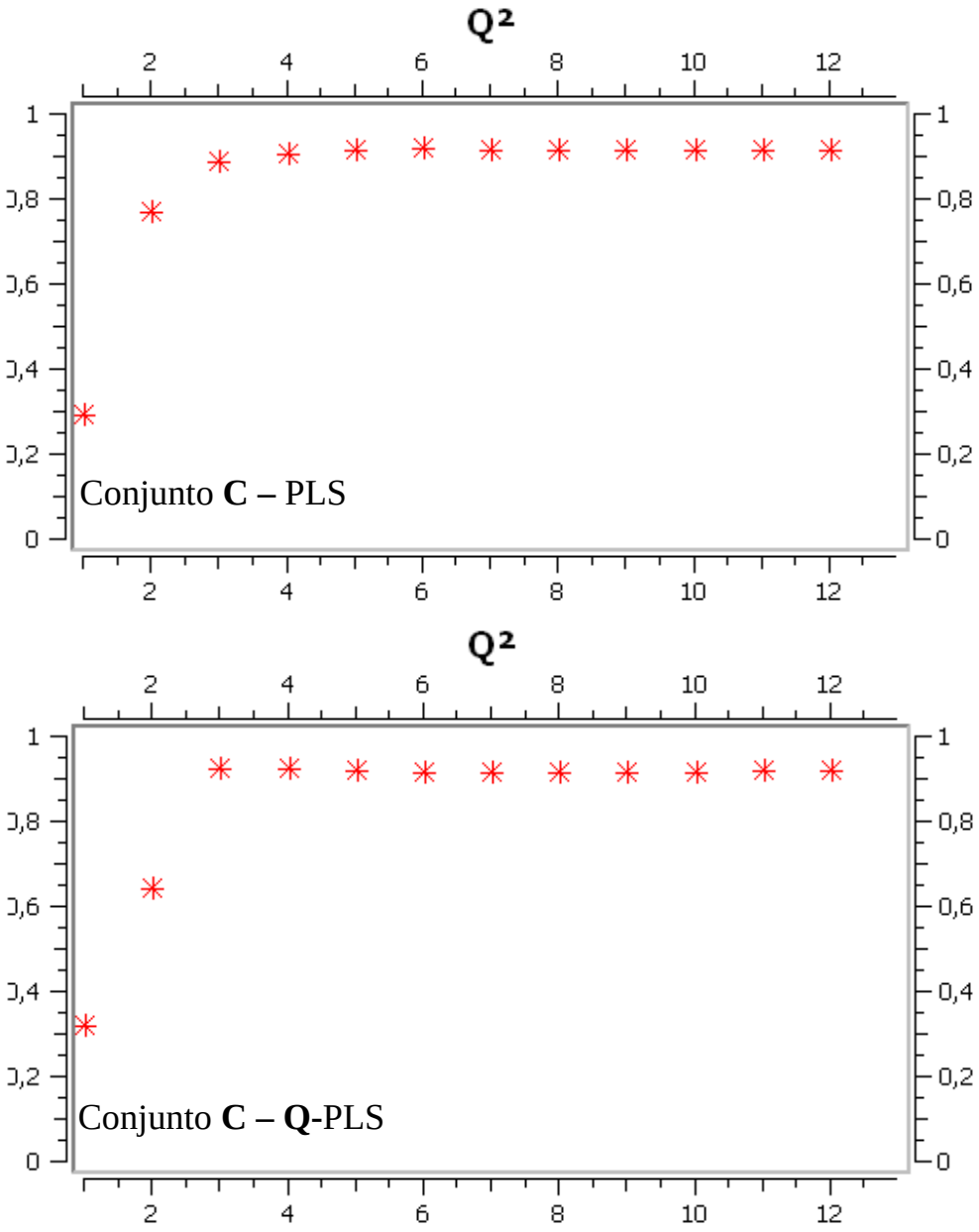


Figura 11: Comparação entre valores de Q² ao longo do número de componentes principais para o "Conjunto C".

Nos gráficos da Figura 11 podem-se observar a evolução da qualidade dos modelos de regressão obtidos por cada um dos algoritmos, conforme se aumenta o número de variáveis latentes utilizadas. Dada a natureza quadrática do conjunto de dados com o qual se está trabalhando, torna-se perceptível que a aplicação do Q-PLS é mais adequada do que a do PLS tradicional: o primeiro algoritmo consegue, para cada número de variáveis latentes, valores maiores de Q² do que a sua versão tradicional.

Isso corrobora a implementação do algoritmo feita pelo MultiMOL, uma vez que o comportamento previsto foi efetivamente observado nos testes realizados.

A capacidade preditiva dos modelos PLS e Q-PLS este conjunto pode ser observada nos gráficos da Figura 12, onde estão apresentadas as comparações entre os valores preditos e observados para as variáveis dependentes de cada conjunto. Trata-se novamente de um gráfico de valores preditos *versus* valores observados, onde a maior capacidade preditiva do modelo é indicada pela maior proximidade dos pontos em relação à reta diagonal. Desta vez, foi utilizado um conjunto de validação com série de teste para aferir a qualidade do modelo de regressão.

A utilização da validação com série de teste é importante porque, como já foi dito, geralmente a capacidade do modelo de prever amostras que não foram utilizadas na calibração de um modelo de regressão é uma característica mais importante do que a sua capacidade de calcular as amostras usadas para construí-lo. Também é desejável que um modelo possa ser validado através de amostras que não foram usadas para a sua construção, pois isto oferece um indicativo mais seguro de que os resultados encontrados não são fortuitos.

Os dados destes modelos são os apresentados na Tabela 3.

Tabela 3 – Dados dos modelos PLS e Q-PLS contruídos para o Conjunto C.

PLS	Q-PLS
Conjunto de calibração: Número de VLs: 6 Número de Amostras: 25 PRESS: 3.775160383 SEP: 8.195463265 Q^2: 0.9204952437 R^2: 0.9998212519 F: 35.28282514 **Conjunto de validação**: Número de Amostras: 50 PRESS: 4368.357 SEP: 9.347039 R^2: 0.8311719 F: 35.28283	**Conjunto de calibração**: Número de VLs: 4 Número de Amostras: 25 PRESS: 473.1156878 SEP: 7.759438809 Q^2: 0.9287300064 R^2: 0.977598689 F: 288.4557707 **Conjunto de validação**: Número de Amostras: 50 PRESS: 971.2496 SEP: 4.407379 R^2: 0.9624632 F: 288.4558

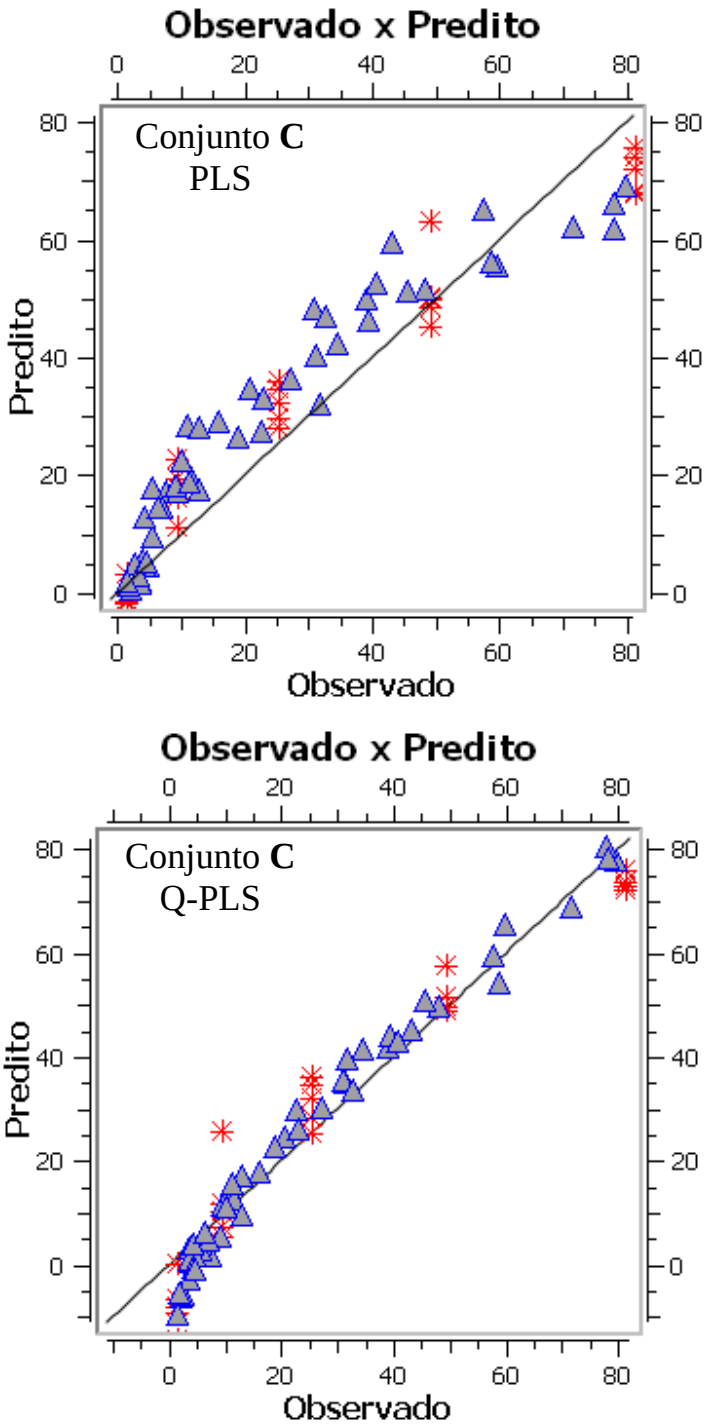


Figura 12: Gráfico de predição para o conjunto de espectro simulado (“Conjunto C”). Em azul, amostras de validação e em vermelho, de calibração.

Os modelos construídos com os dois algoritmos são relativamente próximos, mas é clara a vantagem que leva o PLS Quadrático na previsibilidade dos dados. Como os dados utilizados para o teste são de natureza quadrática, fica clara a insuficiência do

modelo linear (PLS) para a explicação deste fenômeno e a pertinência da aplicação de um modelo não-linear, como o Q-PLS.

É oportuno lembrar que este algoritmo funciona por meio da introdução, na equação de regressão, de um termo quadrático capaz de modelar a natureza não-linear do conjunto de dados, nos casos em que o processo de calibração não identifica dependência quadrática entre os descritores informados e a função resposta, o termo quadrático tende a ser anulado. Assim, o modelo Q-PLS tende a se aproximar do modelo PLS para conjuntos que não possuem dependência quadrática, não havendo prejuízo na explicabilidade dos dados. A única perda significativa, neste caso, é de desempenho: como o Q-PLS exige uma demanda computacional muito maior do que o PLS, os resultados obtidos por aquele levam um tempo consideravelmente maior para estarem disponíveis ao usuário.

Verifica-se, assim, a importância do conhecimento prévio a respeito do problema que está sendo tratado. A opção pelo uso de um algoritmo não-linear em um problema simples, passível de ser modelado adequadamente por meio dos métodos de regressão lineares, pode acrescentar uma complexidade desnecessária ao estudo do problema e alongar o tempo exigido para os testes dos modelos de regressão. Os benefícios advindos do emprego de uma metodologia de regressão quadrática, nestes casos, não seriam proporcionais ao ônus inerente a ela.

Na figura 12, a aplicação do algoritmo de Q-PLS para o modelo de regressão mostrou-se satisfatória e obteve resultados melhores do que o modelo linear (PLS). Isto é refletido também nos números obtidos para os coeficientes de correlação, melhores em 0,1 para o Q-PLS, em comparação com o PLS.

Estes testes mostraram que o Q-PLS é adequado para a modelagem de problemas de natureza quadrática, os quais não são tratados de maneira adequada pelos modelos lineares. O comportamento do algoritmo em relação ao conjunto de dados de espectro simulado foi o esperado: como tal conjunto possui uma relação quadrática previamente conhecida entre as variáveis independentes e a variável dependente, o modelo quadrático apresenta melhores resultados do que o linear, justamente por levar em consideração um componente quadrático na construção da equação de regressão.

O comportamento do algoritmo Q-PLS nos conjuntos de dados que não possuíam dependência quadrática foi também o esperado: a ausência de correlação quadrática entre os descritores e a função-resposta modelada pelo software faz com que o termo quadrático tenha, na equação de regressão, um valor bem próximo de zero. Assim, o modelo Q-PLS reduz-se, na prática, a uma regressão de PLS linear, com resultados que se aproximam daqueles obtidas pela aplicação direta do algoritmo PLS. Para estes conjuntos que não possuem dependência não-linear, os modelos lineares apresentaram melhores resultados, uma vez que levam em consideração somente aspectos que são relevantes, por encontrarem-se justificados na natureza física do

problema com o qual se está trabalhando. Por conta disso, o desempenho deles é substancialmente melhor.

Tais resultados mostraram, assim, a importância de um conhecimento aprofundado sobre a natureza do problema em estudo: os modelos que apresentam melhores resultados de predição são aqueles matematicamente adaptados aos dados que se deseja modelar.

Estes testes foram suficientes para demonstrar que o MultiMOL é aplicável, com performance e robustez, tanto a problemas de QSAR Tradicional quanto de QSAR-3D. Partindo-se do mesmo princípio, i.e., de que a natureza matemática dos algoritmos implementados pelo programa independem da aplicação concreta que se lhes será dada, pode-se expandir as conclusões obtidas dos resultados recém-expostos e afirmar que o MultiMOL é também aplicável a outros problemas de natureza distinta daqueles envolvidos na relação entre estrutura molecular e atividade biológica (QSAR), como por exemplo os problemas de regressão multivariada em química analítica, utilizando dados espectroscópicos, por exemplo. Os modelos de regressão obtidos pelo MultiMOL podem ser, assim, utilizados em quaisquer áreas de conhecimento às quais a metodologia estatística multivariada seja aplicável.

5. Conclusões

A área de desenvolvimento de fármacos assistido por computador tem apresentado um interesse crescente para a comunidade científica, devido à segurança das metodologias utilizadas e aos benefícios associados à incorporação destas técnicas ao processo de desenvolvimento de fármacos. O programa MultiMOL, desenvolvido no departamento de Farmácia da UFPE, pelo grupo de pesquisa em Modelagem para Inovação Molecular (MODiMOL), encaixa-se nesta área do conhecimento, propondo-se a fornecer as ferramentas computacionais exigidos para a realização de estudos de modelagem molecular, notadamente o estabelecimento dos modelos de QSAR, de forma precisa e robusta, com bom desempenho e fácil utilização. Conforme foi mostrado, o software cumpre com os objetivos que se propõe a atingir, apresentando-se como uma interessante alternativa para regressão multivariada na quimiometria.

O presente projeto de mestrado consistiu na otimização e no desenvolvimento de novas funcionalidades para o programa MultiMOL, especificamente no que se refere à incorporação de novas metodologias ao programa e ao desenvolvimento de uma Interface Gráfica do Usuário, de modo a ampliar o leque de aplicações do programa e tornar a sua utilização mais amigável para o usuário. Ao final do projeto, podemos dizer que o software apresenta as características mais importantes de uma ferramenta para aplicações que envolvam problemas de estatística multivariada, tanta na química

medicinal como em outras áreas da ciência onde estas ferramentas se façam necessárias. Os testes que foram realizados no programa garantem a sua ampla aplicabilidade, com o desempenho, a robustez e a precisão originalmente propostas.

Merece destaque o excelente desempenho obtido pelo PLS, por meio do qual é possível obter modelos tão precisos quanto aqueles obtidos com PCR utilizando-se, no entanto, de um processamento computacional muito menor. Estes bons resultados são devidos à escolha do algoritmo PLS1, que foi implementado no MultiMOL.

O Q-PLS foi implementado de maneira inovadora, pois não é comum encontrá-lo nem mesmo nos pacotes de software comerciais. A validação do algoritmo foi feita por meio de testes feitos com um conjunto de dados específico, de natureza quadrática previamente conhecida, por meio do qual foi possível atestar a aplicabilidade do Q-PLS.

Foram também realizadas discussões comparativas entre os dois métodos, PLS e Q-PLS, mostrando os pontos positivos e negativos de cada um deles e estabelecendo, assim, em suas linhas mais gerais, os critérios que devem ser utilizados na escolha do algoritmo de regressão que deve ser aplicado em cada problema concreto. Em particular, foi mostrado que o custo computacional do algoritmo Q-PLS é bem maior do que aquele exigido pelo PLS Tradicional; no entanto, o PLS Quadrático é capaz de modelar melhor do que o PLS os problemas onde a relação entre as variáveis e a função resposta não é de natureza linear.

Todos os testes apresentados nesta dissertação de mestrado foram realizados em três conjuntos de dados, escolhidos de modo a cobrir as principais áreas de aplicação que o MultiMOL se propõe a atender. Foram testados conjuntos de QSAR Tradicional, de QSAR-3D e da Química Analítica (espectros simulados), de diversos tamanhos e distintas naturezas, a fim de abranger um grande leque de aplicações para as quais o MultiMOL pode ser útil. As funcionalidades do programa foram aqui apresentadas através das exigências concretas de cada um dos problemas que foi tratado, em uma harmônica cooperação entre precisão e custo computacional.

Foi também realizada a completa integração entre a Interface Gráfica do programa (GUI) e o seu módulo numérico, aumentando assim a usabilidade do software. Foram apresentadas as principais telas do programa, com uma explicação detalhada de cada uma das opções que elas oferecem ao usuário do MultiMOL. Foram apresentados diversos gráficos, obtidos dos testes realizados e, todos eles, gerados pelo próprio programa.

O esforço associado à pesquisa na literatura sobre os algoritmos, à busca contínua sobre a sua otimização, à sua implementação em C++ para utilização no MultiMOL e à sua integração ao conjunto do programa perfizeram uma parte considerável deste projeto de mestrado, correspondendo a uma parcela significativa do trabalho por ele demandado. Ao final do projeto, estando concluídas as tarefas de codificação e testes dos novos algoritmos de construção de modelos de regressão, podemos concluir que hoje o programa MultiMOL representa uma opção viável para grupos de pesquisa que precisem trabalhar com ferramentas de estatística multivariada.

6. Perspectivas Futuras

Além das novas funcionalidades, é igualmente necessária a elaboração de um manual do usuário, que contenha uma descrição detalhada de todas as funcionalidades do MultiMOL, de modo a facilitar a sua utilização. Este manual precisará conter as opções existentes na versão atual do programa, e o modo como estas opções estão implementadas. A confecção de tutoriais práticos que ensinem os procedimentos de uso do software também é desejada. Toda esta informação poderá ser disponibilizada através de acesso on-line ao website que será desenvolvido e dedicado a este projeto.

Conforme a proposta de projeto de Mestrado submetida ao Programa de Pós-Graduação em Inovação Terapêutica (PPGIT) e por este aprovada, os resultados apresentados nesta dissertação correspondem àqueles que foram originalmente propostos e, portanto, pode-se considerar o presente projeto concluído com sucesso. Não obstante, no decorrer do projeto, foram identificadas diversas funcionalidades e melhorias que poderiam ser agregadas ao MultiMOL, aumentando assim o seu leque de aplicabilidade e contribuindo para consolidá-lo como uma ferramenta quimiométrica de porte e de ampla usabilidade.

A primeira perspectiva futura que o MultiMOL apresenta é o módulo de Grid-Computing. Foi identificado no MultiMOL a oportunidade de uma aplicação de computação distribuída no procedimento de validação interna dos modelos de regressão, notadamente na validação cruzada (LOO-FCV). A computação distribuída (”GRID-Computing”) consiste em uma arquitetura computacional onde duas ou mais máquinas estão interligadas e dividem a execução de tarefas independentes cujos resultados, depois, serão utilizados para a composição do resultado total da computação. Distingue-se o GRID-Computing da arquitetura de “clusters” porque esta última exige compartilhamento de recursos (exige, por exemplo, que a memória seja compartilhada por todas as máquinas que fazem parte do cluster), enquanto que no GRID-Computing essa exigência não existe: é suficiente que as máquinas estejam interligadas numa rede de computadores para que possam ser utilizadas. Além disso, verificou-se que, no GRID-Computing, é possível obter, em alguns casos, um escalonamento das tarefas praticamente linear em proporção ao número de máquinas utilizadas: duas máquinas realizam a atividade na metade do tempo em que a realizaria uma máquina única, quatro máquinas, em 25% do tempo, etc. Identificou-se que grande parte da demanda computacional utilizada nos cálculos computacionais realizados para a construção dos modelos de regressão é proveniente do procedimento de identificação do número ótimo de variáveis latentes, por meio da validação cruzada. Foi identificado também que os diversos modelos de regressão construídos nesta etapa da execução do programa são temporal e logicamente independentes entre si, constituindo um problema conhecido como B-O-T (“Bag-Of-Tasks”), o que os torna passíveis de serem executados em um ambiente de computação distribuída, especialmente através da

utilização de um “GRID Computing”. A possível utilização de “GRID Computing” ocasionaria um ganho enorme de desempenho, diminuindo o tempo de processamento precisamente onde a demanda computacional é mais elevada, ou seja, no gargalo computacional da geração dos modelos de regressão. Foram pesquisadas as tecnologias que podem dar suporte a esta aplicação distribuída, e foi feita a revisão conceitual da arquitetura do MultiMOL de modo a adequá-la a uma aplicação de “GRID Computing”. Decidiu-se pela utilização da tecnologia OurGrid [OURGRID COMMUNITY, 2009], um ambiente para computação distribuída que já foi testado no nosso grupo de pesquisa e já apresentou resultados bastante satisfatórios, possibilitando um escalamento quase linear na redução do tempo de processamento em relação ao número de máquinas utilizadas na distribuição do cálculo. Além disso, este ambiente é de fácil utilização e grande portabilidade, permitindo a construção de sistemas distribuídos sobre arquiteturas computacionais diversas com bastante simplicidade e praticidade. No presente momento, o projeto encontra-se em um estágio adequado para iniciar o desenvolvimento desta nova funcionalidade.

Dentre as outras futuras atividades que podem ser realizadas para dar continuidade ao desenvolvimento do software e que poderão ser, em conjunto ou individualmente, adotadas como objetivos de futuros projetos, julgamos importante destacar as seguintes:

- Implementar o módulo de escrita em disco dos modelos temporários utilizados (principalmente) durante o processo de validação cruzada, a fim de viabilizar o cálculo de conjuntos de dados de porte bem acentuado, em um tempo computacional aceitável.

- Aprimoramentos na Interface Gráfica do MultiMOL, acrescentando opções de visualização de outros tipos de gráficos ou, até mesmo, dando ao usuário a possibilidade de escolher, dentre os diversos resultados gerados pelo modelo, quais exatamente ele deseja graficar.

- Desenvolver um sistema de registro de atividades realizadas pelo programa (arquivo “.log”), por meio do qual pudessem ser registradas as regressões que o usuário realizou e, através da qual, resultados de cálculos feitos anteriormente pudessem ser reproduzidos sem a necessidade de se refazer novamente toda a configuração das opções desejadas para o cálculo do modelo de regressão na interface gráfica.

- Desenvolver um sistema de armazenamento e recuperação dos modelos produzidos pelo MultiMOL, a fim de que um modelo, uma vez calculado, possa ser salvo no computador e, posteriormente, recuperado sem que haja necessidade de se o recalcular; e possibilitando, ainda, que novas amostras possam ser projetadas em modelos anteriormente calibrados.

- Disponibilizar uma versão em língua inglesa do programa, a fim de atingir potenciais usuários que se encontrem fora do Brasil.

- Implementar o princípio da parcimônia (Navalha de Ockham), para otimizar a convergência do número de variáveis latentes utilizadas nos modelos de regressão.

- Implementar o algoritmo APS conforme se encontra na literatura, que não somente faça a seleção de um número ‘x’ de variáveis mas que, por meio da realização de sucessivas regressões, seja capaz de encontrar o número ótimo de variáveis a serem selecionadas no conjunto de dados original.

- Implementar algoritmos classificatórios e exaploratórios (PCA, HCA, KNN, SIMCA, dentre outros), que possam ser utilizados em um módulo do programa independente da construção de modelos de regressão.

- Expandir o algoritmo PLS Quadrático para que ele seja capaz de modelar também outras funções não-lineares de natureza não-quadrática, e verificar a qualidade dos modelos gerados com esta nova implementação.

7. Referências Bibliográficas

ALLEN, M. P.; TILDESLEY, D. J. “Computer Simulation of Liquids”; Oxford University Press: Oxford, **1987.**

AMARAL, A.T.; MONTANARI, C. A.; *Química Nova*, **2002**, *25*, 39-44.

ARAÚJO, M. C. U.; SALDANHA, T. C. B.; GALVÃO, R. K. H.; YONEYAMA, T. Y.; CHAME, H. C.; VISANI, V.; *Chemometrics and Intelligent Laboratory Systems*, **2001**, *57*, 65–73

BERNARD, P.; PINTORE, M.; BERTHON, J.-Y.;CHRÉTIEN, J. R.; Eur. *J. Med. Chem.,* **2001**, *36*, 1

BRERETON, R.G.; *Analyst*, **2000**, *125*, 2125.

CHEMINFORMATICS; http://cheminformatics.org/datasets/index.shtml; Último acesso: 14/08/2008

COATS, E. A.; *Perspectives in Drug Discovery and Design*, **1998**, *12/13/14*, 199.

COHEN, N.C.; BLANEY, J.M.; HUMBLET, C.; GUND, P.; BARRY, D.C.; *J. Med. Chem.* **1990**, *33*, 883.

CRAMER, R. D.; PATTERSON, D. E.; BUNCE, J. D.; *J. Am. Chem. Soc.* **1988**, *110*, 5959.

DAVIES, R.B.; *The second annual object-oriented numerics conference*, **1996**, 213.

EDDELBUTTEL, D.; *Journal of Applied Econometrics*, **1996**, *11*, 199.

FDA, *Challenge and opportunities on the critical path to new medical products*, (http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html), **2009**.

FALCÃO; DE MELO; DA SILVA; NETO; RAMOS; JBCS, 2009.

FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L. O.; *Química Nova*, **1999**, *22*, 724.

GAUDIO, A. C; ZANDONADE, E.; *Química Nova*, **2001**, *24*, 658-671.

GAUDIO, A. C; OLIVEIRA, D. B.; *Quant. Struct.-Act. Relat.*, **2000**, *19*, 599-601

GOLBRAIKH, A.; BERNARD, P.; CHRÉTIEN, J.R.; *Eur. J. Med. Chem.* **2000**, *35*, 123.

GOODFORD, P. J.; *J. Med. Chem.* **1985**, *28*, 849.

HANSCH, C. ; *Acc. Chem. Res.* **1969**, *2*, 232.

HANSCH, C.; FUJITA, T.; *J. Am. Chem. Soc.* **1964**, *86*, 1616.

HANSCH, C.; MALONEY, P.P.; FUJITA, T.; MUIR, R.M.; *Nature*, **1962**, *194*, 178.

HOPFINGER, A.J.; *J. Med. Chem.* **1985**, *28*, 1133.

HOU, T.; XU, X.; *Chemom. and Intellig. Lab. Syst.* **2001**, *56*, 123.

ZHENG, A.; AUSTIN, W. C. P.; *Proceedings of the National Academy of Sciences of the United States of America*, **2006**, *103 (31)*, 11473.

KASTENHOLZ, M.A.; PASTOR, M.; CRUCIANI,G.; HAAKSMA, E.E.J.; FOX,T.; *J. Med. Chem.,* **2000**, 43, 3033.

KUBINYI, H.; *Drug Discovery Today* **1997**, *2*, 457.

KUBINYI, H.; *Drug Discovery Today* **1997**, *2*, 538.

MCGEE, P.; *Drug Disc. Today*, **2005**, *8*, 23.

OURGRID COMMUNITY; http://www.ourgrid.org/; Último acesso: 20/10/2009

PINTORE, M.;BERNARD, P.;BERTHON, J.-Y.;CHRÉTIEN, J. R.; Eur. *J. Med. Chem.,* **2001**, *36*, 21.

QT; http://qt.nokia.com/products; Último acesso: 15/11/2009

QTW; http://qwt.sourceforge.net/; Último acesso: 15/11/2009

SHARAF, M. A.; ILLMAN, D. L.; KOWALSKI, B. R.; “Chemometrics”; Wiley: New York, **1986**

SUTHERLAND, J. J.; O'BRIEN, L. A.; WEAVER, D. F.; *J. Med. Chem*.; **2004**, *47*, 5541-5554

THE F-DISTRIBUTION; http://uregina.ca/~gingrich/f.pdf; Último acesso: 15/11/2009

THE UNSCRAMBLER; http://www.camo.com/; Último acesso: 20/11/2009

TOBIAS, R.D.; *Technometrics*, **1999**, *41*, 375.

TROLLTECH®; http://www.trolltech.com/; Último acesso: 04/02/2011

WAGENER, M.; SADOWSKI, J.; GASTEIGER, J.; *J. Am. Chem. Soc*.,**1995**, *117*, 7769.

WANG, X.; NIU, Y.; CAO, X.; ZHANG, L.; ZHANG, L.-H.; YE, X.-S.; *Bioorg. & Med. Chem.* **2003**, *11*, 4217.

WOLD, S.; KETTANEH-WOLD, N.; SKAGERBERG, B.; *Chem. & Intelligent Lab. Sys.*; **1989**, *7*, 53-65

ANEXO I – Detalhes da geração do conjunto de dados de espectro simulado, utilizados para a validação do algoritmo Q-PLS

A simulação do banco de dados foi realizada utilizando usando o programa GNU Octave 2.1.73 seguindo um planejamento de misturas para calibração de Brereton (2003) com introdução de ruído aos espectros, com coeficientes arbitrários, seguindo os seguintes passos:

- Gerar as constantes de emissão

Foram assumidos espectros com 300 comprimentos de onda, referentes a mistura de três componentes com máximos de emissão nas variáveis 120, 150 e 180, respectivamente.
As primeiras e as últimas 50 variáveis não possuem informação.

```
octave:1> x = 1:300;
octave:2> window = [zeros(1,50) ones(1,200) zeros(1,50)];
octave:3> k1 = exp(-(x-120).^2/1000).*window;
octave:4> k2 = exp(-(x-150).^2/1000).*window;
octave:5> k3 = exp(-(x-180).^2/1000).*window;
octave:6> K = [k1;k2;k3];
```

- Gerar espectros de calibração

Para construir o conjunto de amostras de calibração admitiu-se a mistura dos três componentes em cinco níveis de concentração: 1, 3, 5, 7 e 9, seguindo um planejamento de misturas para calibração de Brereton com 25 amostras [8]. Os espectros foram gerados em seguinda fazendo-se o produto da matriz de concentrações pela matriz das constantes de emissão.

```
octave:7> Ccal=[5 5 5
5 1 1
1 1 9 ... (matriz truncada para simplificar a visualização)
octave:8> Xcal=Ccal*K;
```

- Introduzir ruído

```
octave:9> nivel ruido = 0.03 + (x-150).^4/(10*150^4);
octave:10> randn('state',0); % Reseta o gerador de ruido gaussiano
octave:11> Xcal = Xcal + repmat(nivel ruido,25,1).*randn(25,300);
```

Após a introdução do ruído os espectros de calibração já estão definidos. A figura 1 reproduz os espectros simulados.

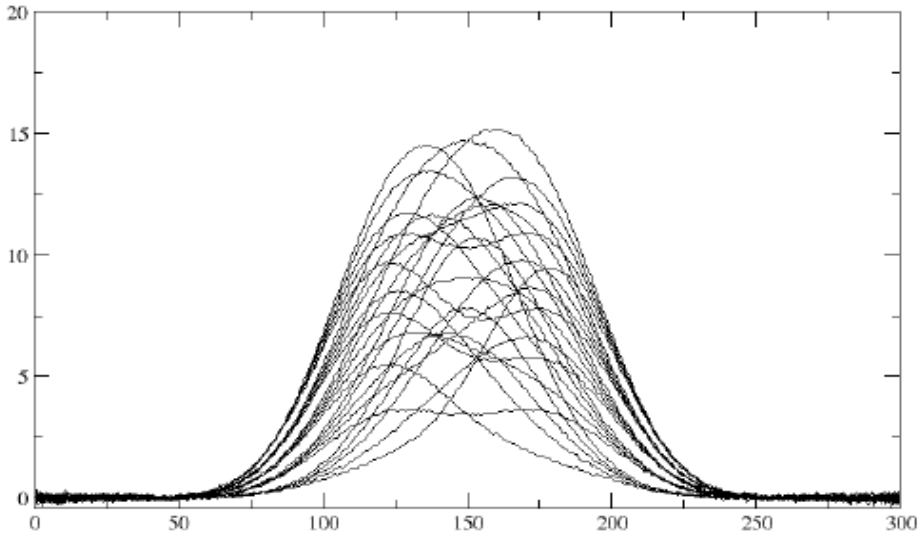


Figura 1: Espectros simulados de uma mistura ternária. Conjunto de calibração.

- Obtenção das respostas simuladas

As quatro respostas com dependência quadrática são apresentadas a baixo:

$$Y_1 = b \cdot c_2^2$$

onde c_2 é a concentração do componente 2.

$$Y_2 = b_0 + b_1 \cdot c_2 + b_2 \cdot c_2^2$$
$$Y_3 = b_1 \cdot c_2^2 + b_2 \cdot c_2^2$$

que simula o caso em que a resposta depende quadraticamente de dois componentes; e

$$Y_4 = b_2 \cdot c_1^2 + b_{23} \cdot c_2 \cdot c_3$$

onde a resposta depende quadraticamente de um componente e do produto dos outros dois. Os coeficientes usados nessas expressões foram definidos arbitrariamente

```
octave:12> y1cal = Ccal(:,2).^2;
octave:13> y2cal = Ccal(:,2).^2 + 5*Ccal(:,2)+4;
octave:14> y3cal = 2*Ccal(:,1).^2+3*Ccal(:,2).^2;
octave:15> y4cal = 3*Ccal(:,1).^2+C(:,2).*C(:,3)
```

A tabela 1 mostra das respostas simuladas para as 50 amostras do conjunto de validação.

Tabela 01: Respostas com dependência quadrática com a concentração para os espectros simulados a partir do planejamento Brereton de misturas de calibração.

Mistura	y1	y2	y3	y4	Mistura	y1	y2	y3	y4
1	25	54	125	100	14	81	130	293	156
2	1	10	53	76	15	81	130	405	252
3	1	10	5	12	16	1	10	165	250
4	81	130	245	30	17	49	88	149	10
5	9	28	189	270	18	1	10	101	152
6	81	130	261	72	19	25	54	77	38
7	25	54	237	258	20	49	88	197	124
8	9	28	77	84	21	49	88	245	168
9	9	28	45	48	22	9	28	125	150
10	49	88	165	90	23	1	10	21	30
11	81	130	341	210	24	9	28	29	18
12	49	88	309	278	25	25	54	93	32
13	25	54	173	192					

- Gerar os espectros do conjunto de validação e suas respostas.

Foram gerados 50 espectros para o conjunto de validação

```
octave:16> rand('state',0); % Reseta o gerador de distribuicao uniforme;
octave:17> Cval(:,1)=8*rand(50,1)+1;
octave:18> Cval(:,2)=8*rand(50,1)+1;
octave:19> Cval(:,3)=8*rand(50,1)+1;
octave:20> Xval = Cval*K;
octave:21>Xval = Xval + repmat(nivel ruido,50,1).*randn(50,300);
```

A figura 2 reproduz os espectros simulados para o conjunto de validação.

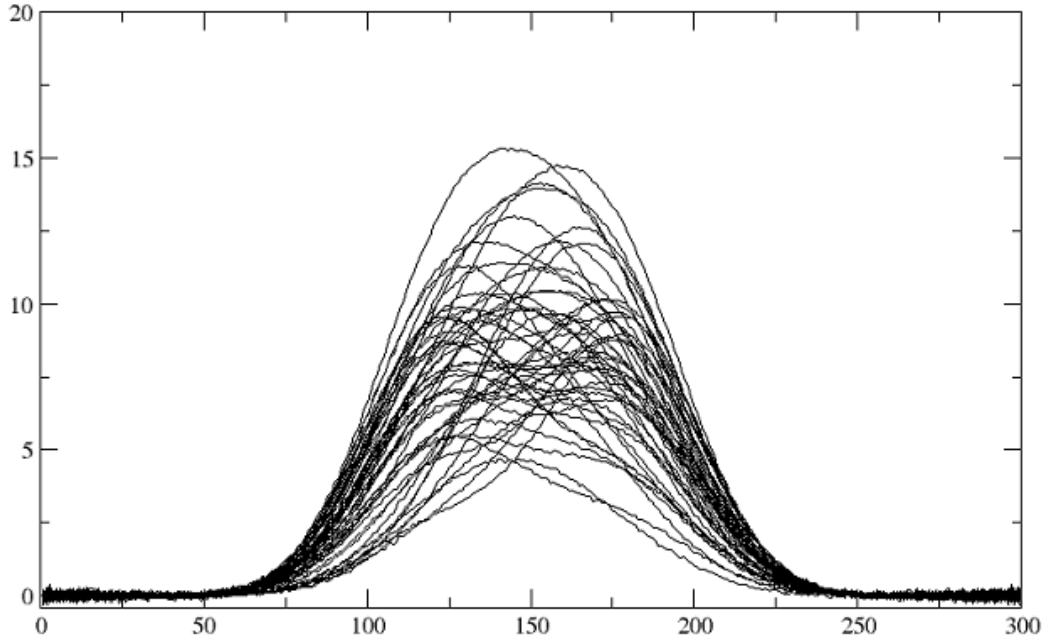


Figura 2: Espectros simulados para o conjunto de validação

```
octave:22> y1val = Cval(:,2).^2
octave:23> y2val = Cval(:,2).^2 + 5*Cval(:,2)+4
octave:24> y3val = 2*Cval(:,1).^2+3*Cval(:,2).^ 2
octave:25> y4val = 3*Cval(:,1).^2+Cval(:,2).*Cval(:,3)
```

A tabela 2 mostra os valores das respostas simuladas para as 50 amostras do conjunto de validação.

Tabela 2: Conjunto de validação, valores das quatro repostas para as misturas simuladas.

Mistura	y1	y2	y3	y4	Mistura	y1	y2	y3	y4
1	30,93	62,74	150,91	122,89	26	1,73	12,29	80,07	120,03
2	20,33	46,87	151,35	149,78	27	10,64	30,96	41,15	17,59
3	79,33	127,87	305,80	162,98	28	3,85	17,66	158,00	226,38
4	3,30	16,38	67,34	101,96	29	11,35	32,20	87,61	101,50
5	7,13	24,49	59,93	65,78	30	3,80	17,55	48,68	62,39
6	5,25	20,70	91,81	126,95	31	12,56	34,28	57,10	50,20
7	38,75	73,87	156,76	96,48	32	18,61	44,18	159,34	174,37
8	9,16	28,29	159,81	215,38	33	2,29	13,86	50,10	68,00
9	22,38	50,03	218,84	240,73	34	42,77	79,47	189,85	114,47
10	8,73	27,51	59,29	75,12	35	30,61	62,28	94,48	34,73
11	5,16	20,52	123,05	171,75	36	9,75	29,37	99,85	123,78
12	3,55	16,96	65,37	96,73	37	26,89	56,82	150,23	133,33
13	39,07	74,32	178,69	133,45	38	3,07	15,83	79,66	116,59
14	4,43	18,96	154,58	219,04	39	31,44	63,48	240,54	254,17
15	6,62	23,48	24,77	26,70	40	71,14	117,31	296,74	162,53
16	15,60	39,35	52,56	25,12	41	12,59	34,33	67,82	74,07
17	57,27	99,11	174,52	64,96	42	40,19	75,88	160,99	85,63
18	3,16	16,04	126,85	186,11	43	4,22	18,49	99,28	139,15
19	59,34	101,86	282,44	218,66	44	45,30	82,95	140,30	61,34
20	3,13	15,97	136,11	201,66	45	10,99	31,57	113,22	145,06
21	77,65	125,70	388,84	293,79	46	6,08	22,41	99,27	137,89
22	22,55	50,30	176,99	187,78	47	32,40	64,86	111,60	31,85
23	77,69	125,76	277,09	142,27	48	1,35	11,15	12,30	22,08
24	34,09	67,28	207,23	193,36	49	58,24	100,40	199,55	88,48
25	47,80	86,38	150,99	41,72	50	1,08	10,26	33,80	55,19

ANEXO II – Trabalhos apresentados em congressos

Como resultado, também, foi apresentado (Figura 12) um trabalho no XV Simpósio Brasileiro de Química teórica, que ocorreu em Poços de Caldas no período de 18 a 21 de outubro de 2009, mostrando o estudo comparativo entre os três conjuntos de dados e os modelos de regressão construídos com os algoritmos PLS e Q-PLS que foi apresentado acima. Este trabalho foi apresentado na forma de pôster.

Certificado de apresentação de trabalho no XV Simpósio Brasileiro de Química Teórica

Anexo III – Manuscrito submetido para publicação na revista Química Nova, com carta de confirmação da submissão on-line.

Programa MultiMOL: uma nova ferramenta quimiométrica para regressão multivariada

Jorge Ferraz de Oliveira Filho, Bruno Feitosa Marques, Marcelo Zaldini Hernandes*

Laboratório de Química Teórica Medicinal – LQTM, Departamento de Ciências Farmacêuticas, Universidade Federal de Pernambuco, Rua Prof. Arthur de Sá, s/n, Cidade Universitária CEP: 50740-521, Recife-PE, Brasil.

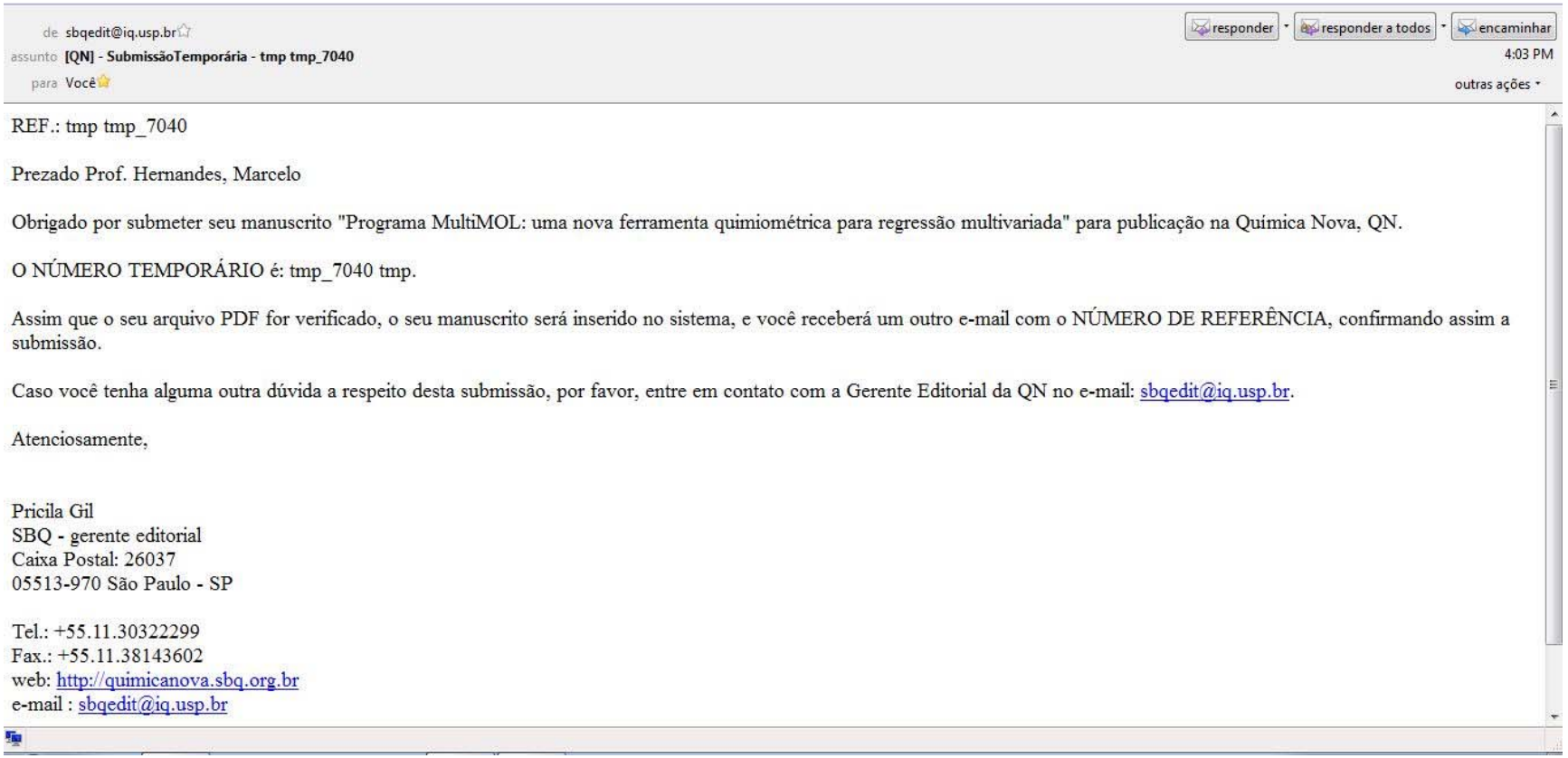
Wallace Duarte Fragoso

Departamento de Química, Universidade Federal da Paraíba, CEP: 58051-970, João Pessoa-PB, Brasil.

Abstract

The challenge to build regression models with chemometric multivariate techniques for medicinal chemistry (QSAR) or analytical chemistry is well known. In this work, we present a new software named MultiMOL, written in C/C++ and with a QT-based GUI, that provides the tools typically used in multivariate regressions (MLR, PCR and PLS). It also provides an algorithm for quadratic PLS, named QPLS. The main advantage of MultiMOL is its high performance because of the optimized implementation. The program was validated with typical examples, showing optimistic results concerning the robustness and the accuracy of the regression models.

Keywords: MultiMOL, multivariate regression, software



de sbqedit@iq.usp.br
assunto [QN] - SubmissãoTemporária - tmp tmp_7040
para Você

responder | responder a todos | encaminhar
4:03 PM
outras ações

REF.: tmp tmp_7040

Prezado Prof. Hernandes, Marcelo

Obrigado por submeter seu manuscrito "Programa MultiMOL: uma nova ferramenta quimiométrica para regressão multivariada" para publicação na Química Nova, QN.

O NÚMERO TEMPORÁRIO é: tmp_7040 tmp.

Assim que o seu arquivo PDF for verificado, o seu manuscrito será inserido no sistema, e você receberá um outro e-mail com o NÚMERO DE REFERÊNCIA, confirmando assim a submissão.

Caso você tenha alguma outra dúvida a respeito desta submissão, por favor, entre em contato com a Gerente Editorial da QN no e-mail: sbqedit@iq.usp.br.

Atenciosamente,

Pricila Gil
SBQ - gerente editorial
Caixa Postal: 26037
05513-970 São Paulo - SP

Tel.: +55.11.30322299
Fax.: +55.11.38143602
web: http://quimicanova.sbq.org.br
e-mail : sbqedit@iq.usp.br