



JOÃO PAULO ATAIDE MARTINS

**DESENVOLVIMENTO DE SOFTWARES, ALGORITMOS E DIFERENTES ABORDAGENS
QUIMIOMÉTRICAS EM ESTUDOS DE QSAR**

**CAMPINAS
2013**



**UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA**

JOÃO PAULO ATAIDE MARTINS

**DESENVOLVIMENTO DE SOFTWARES, ALGORITMOS E DIFERENTES
ABORDAGENS QUIMIOMÉTRICAS EM ESTUDOS DE QSAR**

ORIENTADORA: PROFA. DRA. MÁRCIA MIGUEL CASTRO FERREIRA

**TESE DE DOUTORADO APRESENTADA AO
INSTITUTO DE QUÍMICA DA UNICAMP PARA
OBTENÇÃO DO TÍTULO DE DOUTOR EM CIÊNCIAS.**

**ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA
POR JOÃO PAULO ATAIDE MARTINS, E ORIENTADA PELA
PROFA. DRA. MÁRCIA MIGUEL CASTRO FERREIRA.**

Assinatura da Orientadora

**CAMPINAS
2013**

FICHA CATALOGRÁFICA ELABORADA POR DANIELLE DANTAS DE SOUSA -
CRB8/6490 - BIBLIOTECA DO INSTITUTO DE QUÍMICA DA UNICAMP

M366d	<p>Martins, João Paulo Ataide (1980-). Desenvolvimento de softwares, algoritmos e diferentes abordagens quimiométricas em estudos de QSAR / João Paulo Ataide Martins. – Campinas, SP: [s.n.], 2013.</p> <p>Orientadora: Márcia Miguel Castro Ferreira.</p> <p>Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Química.</p> <p>1. QSAR. 2. PLS. 3. OPS. 4. LQTA-QSAR. 5. QSAR modeling. I. Ferreira, Márcia Miguel Castro. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.</p>
-------	---

Informações para Biblioteca Digital

Título em inglês: Development of softwares, algorithms and different chemometric approaches in QSAR studies

Palavras-chave em inglês:

QSAR
PLS
OPS
LQTA-QSAR
QSAR modeling

Área de concentração: Físico-Química

Titulação: Doutor em Ciências

Banca examinadora:

Márcia Miguel Castro Ferreira [Orientadora]
Ricardo Bicca de Alencastro
Milan Trsic
Rogério Custódio
Pedro Antônio Muniz Vazquez

Data de defesa: 28/06/2013

Programa de pós-graduação: Química

À minha filha, Lúcia.

Se as leis da Matemática referem-se à realidade, elas não estão corretas; e, se estiverem corretas, não se referem à realidade

Albert Einstein

AGRADECIMENTOS

À minha orientadora, Prof^a Dr^a Márcia Miguel Castro Ferreira, pela orientação no desenvolvimento deste trabalho.

À minha esposa, Juliana por todo, apoio, compreensão e por ter sempre ficado ao meu lado.

A todos os membros do Laboratório de Quimiometria Teórica e Aplicada (LQTA), pela companhia e amizade que tivemos durante todo o tempo em que estivemos no LQTA.

Agradecimentos especiais aos membros:

- Kerly Fernanda Mesquita Pasqualoto e Euzébio Guimarães Barbosa, pela parceria no desenvolvimento da metodologia LQTA-QSAR.
- Eduardo Borges de Melo, pela parceria na elaboração de um dos artigos da tese e de tantos outros trabalhos, assim como pela amizade e companheirismo.
- Samuel Anderson Alves de Sousa, pela amizade incondicional e pela ajuda em tantos momentos difíceis ao longo dessa trajetória.

Aos amigos do Piauí que estiveram junto comigo ao longo desse doutorado, em especial ao meu compadre Reginaldo Santos Silva.

Aos Professores das disciplinas que cursei no doutorado, fundamentais na elaboração dessa tese.

Ao Cnpq, pelo apoio financeiro.

Ao Instituto de Química, por todo o suporte fornecido.

À minha família por ter me proporcionado o estudo e ter tornado possível a elaboração desse trabalho.

CURRICULUM VITAE

Dados pessoais

Nome João Paulo Ataide Martins

Filiação Francisco das Chagas Eulálio Martins e Maria Carmelita Estanislau Ataide Martins

Nascimento 25/07/1980 - São Paulo/SP - Brasil

Formação acadêmica/titulação

Mestrado em Química. Universidade Federal do Piauí, UFPI, Teresina – PI.

Período: 03/2004 – 11/2005

Orientador: José Machado Moita Neto

Graduação em Bacharelado em Ciência da Computação. Universidade Federal do Piauí, UFPI, Teresina – PI.

Período: 03/1998 – 02/2002

Atuação profissional

1. Universidade Estadual de Campinas - UNICAMP

Programa de estágio a docência, Monitor.

Período: 08/2008 – 02/2010

2. Companhia de Saneamento Ambiental do Distrito Federal - CAESB

Servidor público , Analista de Sistemas.

3. Instituto de Educação Superior de Brasília - IESB

Docente universitário

Período: 07/2010 até o momento

4. Instituto Dom Barreto - IDB

Docente ensino médio

Período: 02/2004 – 02/2006

5. Colégio Sagrado Coração de Jesus - CSCJ

Docente ensino médio
Período: 02/2004 – 02/2006

6. Universidade Federal do Piauí - UFPI

Monitor
Período: 02/1999 – 07/2000

7. Colégio Diferencial - ANGLO

Docente ensino médio
Período: 02/2004 – 02/2006

8. Instituto Antoine Lavoisier - ANGLO

Docente ensino médio
Período: 02/2004 – 02/2006

Prêmios e títulos

- 1998** Medalha de bronze na Olimpíada Ibero Americana de Química.
- 1998** Medalha de ouro na Olimpíada Brasileira de Química, Associação Brasileira de Química - ABQ
- 1997** Medalha de ouro na Olimpíada Brasileira de Química, Associação Brasileira de Química - ABQ

Produção

Produção bibliográfica

Artigos completos publicados em periódicos

- Martins, João Paulo A.**, Teófilo, Reinaldo F., Ferreira, Márcia M. C. Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. Journal of Chemometrics. , p.320 - 332, 2010.
- Borges de Melo, Eduardo, **Ataide Martins, João Paulo**, Marinho Jorge, Teresa

Cristina, Friozi, Marcelo Couto, Castro Ferreira, Márcia Miguel

Multivariate QSAR study on the antimutagenic activity of flavonoids against 3-NFA on *Salmonella typhimurium* TA98. *European Journal of Medicinal Chemistry.*, v.45, p.4562 - 4569, 2010.

3. **Martins, João Paulo A.**, BARBOSA, E. G., PASQUALOTO, K. F. M., FERREIRA, M. M. C.

LQTA-QSAR: A New 4D-QSAR Methodology. *Journal of Chemical Information and Modeling.*, v.49, p.1428 - 1436, 2009.

4. Teófilo, Reinaldo F., **Martins, João Paulo A.**, Ferreira, Márcia M. C.

Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *Journal of Chemometrics.*, v.23, p.32 - 48, 2009.

Artigos aceitos para publicação

1. **MARTINS, J. P. A.**, FERREIRA, M. M. C. QSAR modeling: Um novo pacote computacional open source para gerar e validar modelos QSAR. *Química Nova* (Impresso). , 2013.

Trabalhos publicados em anais de eventos (resumo)

1. **MARTINS, J. P. A.**, BARBOSA, E. G., PASQUALOTO, K. F. M., FERREIRA, M. M. C. LQTAgrid: an open source package to generate 4D-QSAR descriptors In: 4º Simpósio Brasileiro em Química Medicinal, 2008, Porto de Galinhas, PE. **4º Simpósio Brasileiro em Química Medicinal.** , 2008.

2. **MARTINS, J. P. A.**, de Melo E. B., FERREIRA, M. M. C. 2D-QSAR study of antimutagenic flavonoids using Ordered Predictors Selection (OPS). In: 4º Simpósio Brasileiro em Química Medicinal, 2008, Porto de Galinhas - PE. **4º Simpósio Brasileiro em Química Medicinal.** , 2008.

3. **MARTINS, J. P. A.**, CARVALHO, M. S., IMAMURA, P. M., FERREIRA, M. M. C. Estudo qualitativo da relação estrutura-atividade de derivados do ácido abiético contra *Artemia salina* In: XIV Simpósio Brasileiro de Química Teórica, 2007, Poços de Caldas, MG. **XIV Simpósio Brasileiro de Química Teórica.** , 2007.

4. **MARTINS, J. P. A.**, SOUSA, S. A. A., MOITA NETO, J. M., FERREIRA, M. M. C. Estudo teórico da atividade de Tiossemicarbazonas contra *Salmonella typhimurium*

In: XIV Simpósio Brasileiro de Química Teórica, 2007, Poços de Caldas, MG. **XIV Simpósio Brasileiro de Química Teórica.** , 2007.

5. **MARTINS, J. P. A.**, MUNIZ FILHO, R. C. D., PEREIRA, F. S., FERREIRA, M. M. C. Investigação Teórica do Mecanismo de Abertura de Anéis Epoxídicos. In: XIV Simpósio Brasileiro de Química Teórica, 2007, Poços de Caldas, MG.

XIV Simpósio Brasileiro de Química Teórica. , 2007.

6. **MARTINS, J. P. A.**, de Melo E. B., PEREIRA, F. S., Nogueira, M. A., FERREIRA, M. M. C. QSAR multivariado de dibenzoilmetanos (DBMs) alfa substituídos com atividade anti câncer de mama. In: XLVII Congresso Brasileiro de Química, 2007, Natal, RN. **XLVII Congresso Brasileiro de Química.** , 2007.

7. TEOFILO, R. F., **MARTINS, J. P. A.**, FERREIRA, M. M. C. Study of the computational performance of PLS algorithms using experimental design In: 10th Scandinavian Symposium on Chemometrics, 2007, Lappeenranta, Finlândia. **10th Scandinavian Symposium on Chemometrics.** , 2007.

8. PEREIRA, F. S., **MARTINS, J. P. A.**, PASQUALOTO, K. F. M., FERREIRA, M. M. C., ARAÚJO, R. C. M. U., MONTE, E. V. ESTUDO QUIMIOMÉTRICO DAS PROPRIEDADES ESTRUTURAIS DA OXIRANA E TIRANA In: XLVI Congresso Brasileiro de Química, 2006, Salvador-BA. **XLVI Congresso Brasileiro de Química.** , 2006.

9. TEOFILO, R. F., **MARTINS, J. P. A.**, FERREIRA, M. M. C. Ordered Predictors Selection: an intuitive method to find the most relevant variables in multivariate regression In: 10th International Conference on Chemometrics in Analytical Chemistry, 2006, Águas de Lindóia - SP. **10th International Conference on Chemometrics in Analytical Chemistry.** , 2006.

10. **MARTINS, J. P. A.**, MOITA NETO, J. M. Relação estrutura-atividade de um conjunto de semicarbazonas e tiossemicarbazonas contra o micróbio Bacillus subtilis In: 29a Reunião Anual da Sociedade Brasileira de Química, 2006, Águas de Lindóia - SP. **29a Reunião Anual da Sociedade Brasileira de Química.** , 2006.

11. PEREIRA, F. S., **MARTINS, J. P. A.**, de Melo E. B., FERREIRA, M. M. C. 2D-QSAR Analysis of aziridinyl-1,4-naphtoquinone Antimalarials Using Partial Least Square (PLS) In: 3rd Brazilian Symposium on Medicinal Chemistry, 2006, São Pedro-SP. **3rd Brazilian Symposium on Medicinal Chemistry.** , 2006.

12. **MARTINS, J. P. A.**, MOITA NETO, J. M. Estudo Semi-Empírico de

Semicarbazonas e Tiossemicarbazonas In: XLV Congresso Brasileiro de Química, 2005, Belém - PA. **XLV Congresso Brasileiro de Química.** , 2005.

13. **MARTINS, J. P. A.**, COSTA JUNIOR, J. S., LUZ JUNIOR, G. E., MOITA NETO, J. M. ESTUDO MULTIVARIADO DAS ENERGIAS DE ORBITAIS DE SISTEMAS DECAELETRÔNICOS. In: XLIV Congresso Brasileiro de Química, 2004, Fortaleza. **XLIV Congresso Brasileiro de Química.** , 2004.

Trabalhos publicados em anais de eventos (resumo expandido)

1. TEOFILO, R. F., **MARTINS, J. P. A.**, FERREIRA, M. M. C. Computational performance of PLS algorithms: a comparison. In: 5th International symposium on PLS and related methods., 2007, Matforsk, Aas, Noruega. **5th International symposium on PLS and related methods..** , 2007.

Produção técnica

Programa de computador sem registro

1. **MARTINS, J. P. A.**, FERREIRA, M. M. C.
QSAR modeling: Um novo pacote computacional open source para gerar e validar modelos QSAR, 2009

2. **MARTINS, J. P. A.**, BARBOSA, E. G., PASQUALOTO, K. F. M., FERREIRA, M. M. C.
LQTAgrid, 2008

Demais produções técnicas

1. **Martins, João Paulo A.**
Química computacional aplicada a QSAR, 2010. (Extensão, Curso de curta duração ministrado)

2. **Martins, João Paulo A.**, BARBOSA, E. G., PASQUALOTO, K. F. M., FERREIRA, M. M. C.
Aplicação da metodologia QSAR-4D usando o programa LQTA-QSAR., 2009. (Extensão, Curso de curta duração ministrado)

3. **Martins, João Paulo A.**

OPS - algoritmo de seleção de variáveis– construção e validação de modelos QSAR – programa QSAR modeling, 2009. (Outra produção técnica)

4. **Martins, João Paulo A.**

Química computacional aplicada a QSAR, 2009. (Extensão, Curso de curta duração ministrado)

5. **MARTINS, J. P. A., ANDRADE, T. C.**

Introdução à internet, 2000. (Extensão, Curso de curta duração ministrado)

Orientações e supervisões concluídas

Trabalhos de conclusão de curso de graduação

1. Daniel da Silva Souza. **BioAgents: Uma ferramenta multiagente para anotação de sequências biológicas**. 2012. Curso (Ciência da Computação) - Instituto de Educação Superior de Brasília

RESUMO

O planejamento de fármacos com o auxílio do computador é uma área de pesquisa de extrema importância em química e áreas correlatas. O conjunto de ferramentas disponíveis para tal fim consiste, dentre outras, em programas para geração de descritores e construção e validação de modelos matemáticos em QSAR (do inglês, *Quantitative Structure-Activity Relationship*). Com o objetivo de tornar esse estudo mais acessível para a comunidade científica, novas metodologias e programas para geração de descritores e construção e validação de modelos QSAR foram desenvolvidos nessa tese. Uma nova metodologia de QSAR 4D, conhecida com LQTA-QSAR, foi desenvolvida com o objetivo de gerar descritores espaciais levando em conta os perfis de amostragem conformacional das moléculas em estudo obtidos a partir de simulações de dinâmica molecular. A geração desses perfis é feita com o software livre GROMACS e os descritores são gerados a partir de um novo software desenvolvido nesse trabalho, chamado de LQTAgrid. Os resultados obtidos com essa metodologia foram validados comparando-os com resultados obtidos para conjuntos de dados disponíveis na literatura. Um outro software de fácil uso, e que engloba as principais ferramentas de construção e validação de modelos em QSAR, foi desenvolvido e chamado de QSAR *modeling*. Esse software implementa o método de seleção de variáveis OPS, desenvolvido em nosso laboratório, e utiliza PLS (do inglês *Partial Least Squares*) como método de regressão. A escolha do algoritmo PLS implementado no programa foi feita com base em um estudo sobre o desempenho e a precisão no erro de validação dos principais algoritmos PLS disponíveis na literatura. Além disso, o programa QSAR *modeling* foi utilizado em um estudo de QSAR 2D para um conjunto de 20 flavonóides com atividade anti-mutagênica contra 3-nitrofluoranteno (3-NFA).

Palavras-chave: QSAR; PLS; Construção e validação de modelos; OPS; Dinâmica molecular; LQTA-QSAR; QSAR *modeling*.

ABSTRACT

Computer aided drug design is an important research field in chemistry and related areas. The available tools used in such studies involve software to generate molecular descriptors and to build and validate mathematical models in QSAR (Quantitative Structure-Activity Relationship). A new set of methodologies and software to generate molecular descriptors and to build and validate QSAR models were developed aiming to make these kind of studies more accessible to scientific community. A new 4DQSAR methodology, known as LQTA-QSAR, was developed with the purpose to generate spatial descriptors taking into account conformational ensemble profile obtained from molecular dynamics simulations. The generation of these profiles is performed by free software GROMACS and the descriptors are generated by a new software developed in this work, called LQTAgrid. The results obtained with this methodology were validated comparing them with results available in literature. Another user friendly software, which contains some of the most important tools used to build and validate QSAR models was developed and called QSAR modeling. This software implements the OPS variable selection algorithm, developed in our laboratory, and uses PLS (Partial Least Squares) as regression method. The choice of PLS algorithm implemented in the program was performed by a study about the performance and validation precision error involving the most important PLS algorithms available in literature. Further, QSAR modeling was used in a 2D QSAR study with 20 flavonoid derivatives with antimutagenic activity against 3-nitrofluoranthene (3-NFA).

Keywords: QSAR; PLS; Models building and validation; OPS; Molecular Dynamics; LQTA-QSAR; QSAR modeling.

SUMÁRIO

Lista de Abreviaturas	xxvii
Lista de tabelas	xxix
Lista de figuras	xxxi
PREFÁCIO	1
Capítulo 1. Desempenho computacional e precisão no erro de validação cruzada de cinco algoritmos PLS usando dados reais e simulados	3
1.1. Introdução	4
1.2. Notação	6
1.3. Algoritmos	7
1.3.1. O algoritmo NIPALS clássico	7
1.3.2. Algoritmo NIPALS modificado (NIPALSy)	8
1.3.3. Algoritmo Kernel	8
1.3.4. O algoritmo SIMPLS	9
1.3.5. O algoritmo de bidiagonalização	10
1.4. Experimental	12
1.4.1. Conjuntos de dados simulados	12
1.4.1.1. Planejamento fatorial	12
1.4.1.2. Planejamento quadrado latino	14
1.4.2. Conjuntos de dados reais	15
1.5. Resultados e discussão	18
1.5.1. Conjuntos de dados simulados	18
1.5.1.1. Planejamento fatorial	18
1.5.1.2. Planejamento quadrado latino	25
1.5.2. Conjunto de dados reais	30
1.6. Conclusões	32
Capítulo 2. Fundamentação teórica sobre QSAR e quimiometria	35
2.1. Uma introdução aos estudos de QSAR	35
2.2. QSAR-3D	38
2.3. QSAR-4D	39
2.4. Estudos de QSAR que resultaram em fármacos hoje no mercado	42
2.5. Quimiometria aplicada aos estudos de QSAR	43
2.5.1. Construção do modelo matemático	43
2.5.1.1. Regressão Linear Múltipla (MLR)	44
2.5.1.2. Regressão de componentes principais (PCR)	45
2.5.1.3. Regressão de quadrados mínimos parciais (PLS)	48

2.5.2.	Pré-processamento	50
2.5.3.	Validação cruzada	52
2.5.4.	Detecção de amostras anômalas	55
2.5.5.	Seleção de variáveis com o algoritmo OPS	56
2.5.6.	Validação externa	58
2.5.7.	Avaliação da robustez do modelo com o teste leave- <i>N</i> -out	60
2.5.8.	Avaliação da correlação ao acaso com o teste de aleatorização de <i>y</i>	61
Capítulo 3.	Estudo QSAR multivariado da atividade antimutagênica de flavonoides contra 3-NFA em <i>Salmonella typhimurium</i> TA98	63
3.1.	Introdução	63
3.2.	Farmacologia	65
3.3.	Química	67
3.4.	Metodologia	68
3.5.	Resultados	71
3.6.	Interpretação do modelo	78
3.7.	Conclusões	83
Capítulo 4.	LQTA-QSAR: Uma nova metodologia de QSAR 4D	85
4.1.	Introdução	86
4.2.	Metodologia	87
4.2.1.	Conjuntos de dados investigados – comparação de metodologias	90
4.2.2.	Simulações de dinâmica molecular	93
4.2.3.	Análises LQTAgrid	95
4.2.4.	Seleção de variáveis e validação do modelo	97
4.3.	Resultados e discussão	98
4.3.1.	Interpretação dos descritores	102
4.4.	Conclusões	105
Capítulo 5.	QSAR <i>modeling</i>: um pacote computacional open source para gerar e validar modelos QSAR	107
5.1.	Introdução	108
5.2.	Metodologia	110
5.3.	Resultados e discussão	110
5.3.1.	Pré-processamento dos dados	111
5.3.2.	Construção de modelos de regressão com o método PLS	112
5.3.3.	Seleção de variáveis com o algoritmo OPS	114

5.3.4.	Detecção de amostras anômalas (<i>outliers</i>)	119
5.3.5.	Validação cruzada excluindo N amostras	121
5.3.6.	Teste de aleatorização de y (y -randomization)	125
5.3.7.	Comparação com alguns dos softwares citados	127
5.4.	Conclusões	128
	Conclusão geral e perspectivas futuras	131
	Referências Bibliográficas	133

LISTA DE ABREVIATURAS

ANOVA	Analysis Of Variance
CEP	Conformational Ensemble Profile
CHELPG	Charges from Eletrostatic Potentials using a Grid based method
CoMFA	Comparative Molecular Field
CSD	Cambridge Structural Database
DFT	Density Functional Theory
GC	Gas Chromatography
GCOD	Grid Cell Occupancy Descriptors
GETAWAY	Geometry, Topology and Atom-Weights Assembly
HCA	Hierarchical Cluster Analysis
HF	Hartree-Fock
HOMO	Highest Occupied Molecular Orbital
IPE	Interaction Pharmacophore Elements
JRE	Java Runtime Environment
JVM	Java Virtual Machine
LINCS	Linear Constraint Solver
LNO	Leave-N-Out
LQTA	Laboratório de Quimiometria Teórica e Aplicada
LUMO	Lowest Unoccupied Molecular Orbital
MD	Molecular Dynamics
MIM	Matriz de Influência Molecular
MLR	Multiple Linear Regression
MS	Mean Square
NIPALS	Nonlinear Iterative Partial Least Squares
NIR	Near InfraRed Spectroscopy
NMR	Nuclear Magnetic Ressonance
NPT	Número de partículas constante, Pressão e Temperatura
OLS	Ordinary Least Squares
OPS	Ordered Predictors Selection
PAH	Hidrocarbonetos Poliaromáticos
PCA	Principal Component Analysis
PCR	Principal Component Regression
PDB	Protein Data Bank
PLS	Partial Least Squares
PME	Particle Mesh Ewald
QSAR	Quantitative Structure Activity Relationship
QSPR	Quantitative Structure Property Relationship
RMSECV	Root Mean Sqaure Error of Cross-Validation
SEC	Standard Error of Calibration

SEP	Standard Error of Prediction
SEV	Standard Error of Validation
SS	Sum of Squares
SVD	Singular Values Decomposition
UV	Ultravioleta-visível
WHIM	Weighted Holistic Invariant Molecular

LISTA DE TABELAS

Tabela 1.1.	Fatores, níveis codificados e domínio investigado em um planejamento.	13
Tabela 1.2.	Níveis estudados para cada fator no planejamento quadrado latino	15
Tabela 1.3.	Modelos fatoriais completos para os cinco algoritmos usando os conjuntos de dados SX e LX.	21
Tabela 1.4.	Comparação das diferenças de tempo de execução entre algoritmos usando teste t-pareado para os conjuntos de dados SX e LX.	22
Tabela 1.5.	Diferença nos valores de RMSECV (Equação 1.6) entre ensaios para os conjuntos de dados SX e LX.	24
Tabela 1.6.	Resultados da ANOVA usando planejamento quadrado latino para os cinco algoritmos.	26
Tabela 1.7.	Comparação das diferenças de tempos de execução entre algoritmos usando teste t pareado para o conjunto de dados usado no planejamento quadrado latino.	29
Tabela 1.8.	Diferença de valores de RMSECV entre ensaios para os conjuntos de dados do planejamento quadrado latino.	29
Tabela 1.9.	Tempo (em segundos) de cada algoritmo variando o tipo de conjunto de dados, dimensão e número de variáveis latentes.	32
Tabela 2.1.	Dez passos operacionais realizados na análise QSAR 4D.	41
Tabela 2.2.	Parâmetros estatísticos que costumam ser calculados para avaliar a qualidade de um modelo durante uma validação cruzada.	53
Tabela 2.3.	Parâmetros estatísticos usados na validação externa.	60
Tabela 3.1.	Conjunto de treinamento selecionado da literatura ² e efeitos antimutagênicos observados (no pID50) na atividade mutagênica induzida pelo 3-NFA em <i>S. typhimurium</i> TA98.	66
Tabela 3.2.	Valores preditos para o conjunto teste e resultados dos parâmetros estatísticos.	71
Tabela 3.3.	Valores dos descritores usados para a formulação do modelo e resultados da validação cruzada LOO (exceto para a amostra anômala 14).	73
Tabela 3.4.	Coeficientes de correlação individual de Pearson	76

	(modelo final sem amostras anômalas) e coeficientes padronizados do modelo.	
Tabela 4.1.	Sondas disponíveis no módulo LQTAgrid.	88
Tabela 4.2.	Estruturas e atividades experimentais dos conjuntos de dados 1 [106] e 2 [107]. Os átomos numerados foram usados para o alinhamento dos CEPs de todos os	91
Tabela 4.3.	Parâmetros estatísticos obtidos para os modelos OPS-PLS e modelos da literatura [106,107]. Os valores entre parênteses correspondem ao número de variáveis latentes usadas nos modelos PLS.	98
Tabela 4.4.	Valores de resíduos obtidos para os conjuntos teste usando os modelos OPS-PLS.	101
Tabela 5.1.	Comparativo entre as principais características do programa <i>QSAR modeling</i> e outros programas disponíveis na literatura.	108
Tabela 5.2.	Resultados da validação cruzada obtidos para um modelo com 3 LV após a seleção de variáveis feita com o programa <i>QSAR modeling</i> .	119

LISTA DE FIGURAS

Figura 1.1.	Efeitos obtidos a partir do planejamento fatorial completo para os conjuntos de dados <i>SX</i> (A) e <i>LX</i> (B).	25
Figura 1.2.	Valores de quadrado médio obtidos a partir do planejamento quadrado latino.	27
Figura 1.3.	Gráficos de efeitos para o planejamento quadrado latino. PLSBi, A1, B1, C1; SIMPLS, A2, B2, C2; Kernel, A3, B3, C3; NIPALSy, A4, B4, C4; NIPALS, A5, B5, C5.	28
Figura 1.4.	Tempo de execução versus <i>nVL</i> para uma matriz 1000×10000.	28
Figura 1.5.	Conjuntos de dados de testes utilizados. A: espectros de infravermelho próximo (NIR); B: espectros raman (Raman); C: espectros de fluorescência em forma de matriz (Fluor); D: voltamogramas (Volt); E: conjunto de dados tipo UV (Tipo UV); F: cromatografia gasosa (CG).	31
Figura 2.1.	Exemplo de um CEP dentro de um grid onde podem ser calculados os descritores de ocupação em 4D-QSAR	41
Figura 2.2.	Representação das variáveis depois de cada pré-processamento.	52
Figura 2.3.	Exemplo de execução de uma validação cruzada leave-3-out	53
Figura 2.4.	Exemplo de funcionamento do algoritmo OPS. a) Matriz original. O tamanho das barras representa a importância do descritor dada pelo vetor informativo. b) Matriz rearranjada de acordo com a importância de cada descritor. c) Construção de modelos PLS para uma janela inicial igual a 5 e incremento igual a 3.	59
Figura 2.5.	Exemplo de aleatorização de <i>y</i> . Os descritores originais são mantidos enquanto que as atividades biológicas são permutadas entre as amostras.	62
Figura 3.1.	Estruturas dos nitroarenos 2-NF, 3-NFA e 1-NP.	64
Figura 3.2.	Histograma apresentando a distribuição dos compostos nas faixas de <i>pID</i> ₅₀ .	68
Figura 3.3.	Dendrograma (dados autoescalados) do conjunto de treinamento, com os compostos 6 , 14 , 13 e 18	74

	destacados.	
Figura 3.4.	Gráficos do teste de aleatorização de y (A, B e C) e validação cruzada LNO (D). No gráfico de LNO (D), cada ponto se refere ao valor médio de um teste em triplicata e as barras se referem ao desvio padrão.	75
Figura 3.5.	Dendrograma (dados autoescalados) do conjunto completo (sem a amostra anômala 14), com os compostos do conjunto teste destacados.	77
Figura 3.6.	Representação das componentes principais para os compostos 6 e 16 . $x = 1^{\text{a}}$ componente, $y = 2^{\text{a}}$ componente, $z = 3^{\text{a}}$ componente.	82
Figura 4.1.	Representação da caixa 3D virtual ou grid gerada pelo módulo LQTAgrid. A distância recomendada entre as coordenadas CEP e as bordas da rede 3D são de pelo menos 5 Å. A distância do grid entre cada ponto adjacente é de 1 Å.	88
Figura 4.2.	Comparação dos CEPs resultantes das simulações de DM para um dos compostos mais ativos e um dos mais inativos de cada conjunto de dados investigado. Os dados biológicos dos conjuntos 1 e 2 são expressos como ΔG (kcal/mol) e pIC_{50} , respectivamente.	95
Figura 4.3.	Gráficos dos resultados de LNO obtidos para os conjuntos 1 e 2, respectivamente.	99
Figura 4.4.	Gráficos de q^2 versus r^2 obtidos para 50 aleatorizações de y .	100
Figura 4.5.	Gráfico das atividades observadas (experimentais) versus preditas (calculadas) para os conjuntos de treinamento (preto) e teste (cinza claro) (conjuntos 1 e 2).	101
Figura 4.6.	Visualização dos descritores de campo obtidos pelo método LQTAgrid e selecionados pelo algoritmo OPS para a molécula mais ativa do conjunto 1 (ViewerLite 5.0, Accelrys, Inc., 2002).	102
Figura 4.7.	Visualização dos descritores de campo obtidos pelo método LQTAgrid e selecionados pelo algoritmo OPS para a molécula mais ativa do conjunto 2 (ViewerLite 5.0, Accelrys, Inc., 2002).	103
Figura 5.1.	Tela principal do programa <i>QSAR modeling</i> .	111
Figura 5.2.	Janela do programa <i>QSAR modeling</i> na qual o usuário escolhe o número máximo de variáveis latentes e o número de amostras a serem removidas durante a	114

	validação cruzada.	
Figura 5.3.	Janela do programa <i>QSAR modeling</i> na qual os resultados da validação cruzada são mostrados. Os parâmetros 1 a 9 da Tabela 2.2, os coeficientes de regressão, os valores previstos para a variável dependente na validação cruzada e os valores previstos para a variável dependente no modelo de regressão podem ser vistos nessa janela.	115
Figura 5.4.	Janela do programa <i>QSAR modeling</i> na qual o usuário escolhe as opções de execução do algoritmo OPS.	116
Figura 5.5.	O valor do parâmetro escolhido para avaliar o modelo, o número de variáveis selecionadas e o número de variáveis latentes dos dez modelos selecionados são mostrados como resultados do algoritmo OPS.	118
Figura 5.6.	Resultado da detecção de amostras anômalas mostrando os valores de influência e dos resíduos de Student para os compostos do conjunto de treinamento.	122
Figura 5.7.	Gráfico de Influência <i>versus</i> Resíduos de Student para a detecção de amostras anômalas (<i>outliers</i>). As linhas azuis indicam os limites aceitos pela literatura.	122
Figura 5.8.	Procedimento de validação <i>leave-N-out</i> para garantir a robustez de um modelo usando o programa <i>QSAR modeling</i> .	123
Figura 5.9.	Resultados obtidos com o procedimento de validação <i>leave-N-out</i> .	124
Figura 5.10.	Validação <i>leave-N-out</i> aplicada ao modelo final obtido depois da seleção de variáveis com o algoritmo OPS. Os pontos representam a média e as barras indicam o desvio padrão de uma triplicata para cada valor de <i>N</i> . O modelo mostrou-se robusto até um valor de <i>N</i> igual a 11 (30% das amostras).	124
Figura 5.11.	Procedimento de validação de aleatorização de <i>y</i> para verificar a correlação ao acaso de um modelo usando o programa <i>QSAR modeling</i> .	125
Figura 5.12.	Resultados do teste de aleatorização de <i>y</i> fornecidos pelo programa <i>QSAR modeling</i> .	126
Figura 5.13.	Valores de R^2 e Q^2 obtidos com o teste de aleatorização de <i>y</i> . O ponto distante representa os valores de R^2 e Q^2 para o modelo real.	127

Prefácio

O trabalho de tese aqui apresentado tem como foco geral o desenvolvimento de ferramentas quimiométricas com aplicações no estudo das relações quantitativas estrutura e atividade (QSAR do inglês *Quantitative Structure Activity Relationship*). Este ramo da ciência tem demonstrado grande crescimento nos últimos anos, apresentando o desenvolvimento de novas abordagens, com o objetivo de cada vez tornar mais claro as relações entre a estrutura das diversas moléculas, que presumivelmente atuam como fármacos, e as atividades farmacológicas propostas.

Tendo em vista que um modelo matemático preditivo é sempre o alvo final dos estudos em QSAR, torna-se importante a investigação dos métodos multivariados de regressão que são usados para a construção dos modelos. Diante disso, o primeiro capítulo desta tese trata de uma avaliação do desempenho computacional de cinco algoritmos para realização do método multivariado de quadrados mínimos parciais, PLS (do inglês *Partial Least Squares*) usualmente utilizados e amplamente comentados na literatura. Ainda neste capítulo, a implicação do uso de grandes conjuntos de dados no desempenho dos algoritmos será comentada, com a exemplificação através de conjuntos de dados reais e simulados.

Na sequência desta tese, o segundo capítulo traz a fundamentação teórica envolvida nos estudos de QSAR onde serão destacados QSAR 2D, 3D e 4D, além de aspectos sobre as principais ferramentas quimiométricas que normalmente são utilizadas. Neste ponto, os métodos multivariados, pré-tratamentos e as metodologias para validação dos modelos são expostos. Adicionalmente, é descrito um método desenvolvido no nosso laboratório, para a seleção de variáveis, denominado OPS (do inglês *Ordered Predictor Selection*). Alguns detalhes do algoritmo do OPS serão comentados.

Uma vez que os principais aspectos de QSAR são apresentados no capítulo anterior, o terceiro capítulo tratará de uma aplicação com o desenvolvimento de um modelo de QSAR 2D para estudo de vinte flavonóides com atividade anti-mutagênica

contra 3-nitrofluoranteno sobre *Salmonella typhimurium* TA98. A aplicação conta com o uso de PLS como método de regressão, do OPS para seleção de variáveis, abordagens para validação (*leave-n-out* e aleatorização do vetor y) que são previamente delineados no segundo capítulo. Além disso, é realizada uma discussão envolvendo o significado de cada descritor importante para o modelo e os seus possíveis papéis no mecanismo de ação anti-mutagênica.

No quarto capítulo uma nova abordagem de QSAR 4D, chamada LQTA-QSAR, é apresentada com aplicações. A abordagem usa trajetórias de dinâmica molecular e informações topológicas das moléculas em estudo e calcula energias de interações intermoleculares entre o perfil dinâmico das moléculas e pontos a distâncias específicas em uma caixa (ou *grid*), onde uma sonda (um átomo ou fragmento de uma molécula) se movimenta. Novamente, o método OPS é usado para a seleção de variáveis e modelos promissores, após extensa validação, são obtidos. O objetivo da tese é também fornecer as ferramentas computacionais (algoritmos) para a execução desta abordagem e desse modo o módulo LQTAgrid foi desenvolvido para a construção do *grid* supracitado. Outros softwares de acesso livre são necessários, por exemplo, o GROMACS. Finalmente, uma comparação com outras abordagens de QSAR 3D e 4D também são mostradas.

O quinto capítulo corresponde à apresentação de um software de uso livre, chamado *QSAR modeling*, especificamente desenvolvido para a execução das tarefas inerentes aos estudos em QSAR, tais como construção de modelos de regressão PLS, seleção de variáveis com o método OPS, abordagens de validação (*leave-n-out* e aleatorização do vetor y) e detecção de amostras anômalas (*outliers*). O software foi desenvolvido em Java versão 6 e pode ser usado em qualquer computador cujo sistema operacional suporte o *Java Runtime Environment* (JRE) versão 6. Este capítulo é escrito como um tutorial com a ilustração da construção de um modelo para 644 compostos com toxicidade contra *T. pyriformis*. Após este capítulo as considerações finais da tese são apresentadas.

Capítulo 1

Desempenho computacional e precisão no erro de validação cruzada de cinco algoritmos PLS usando dados reais e simulados

Os estudos de QSAR têm como objetivo a construção de um modelo matemático que relacione a estrutura química de um conjunto de moléculas com a atividade biológica apresentada por elas. Esse modelo matemático costuma ser obtido através de uma regressão linear realizada entre a matriz de descritores que contêm as informações a respeito da estrutura química do conjunto estudado e a atividade biológica expressa por um vetor.

Dentre os métodos de regressão existentes para a obtenção desse modelo matemático, o método PLS tem se mostrado o mais promissor e vem sendo bastante utilizado em estudos de QSAR. Diversos algoritmos já foram propostos na literatura com o intuito de melhorar cada vez mais o desempenho do método. Como hoje em dia estudos de QSAR apresentam matrizes com até dezenas de milhares de descritores, é extremamente importante escolher o algoritmo mais eficiente quando se vai implementar uma regressão PLS, pois o tempo computacional é fator fundamental para uma análise rápida e de qualidade.

Assim, com o objetivo de identificar o algoritmo PLS mais eficiente para ser implementado nos softwares e algoritmos desenvolvidos nessa tese, foi feito um estudo comparando-se o tempo computacional e a precisão dos erros observados durante a validação cruzada para os cinco mais importantes algoritmos PLS disponíveis na literatura. O resultado desse estudo foi publicado como artigo na revista *Journal of Chemometrics* e será apresentado a seguir.

1.1. Introdução

A Calibração multivariada é usada para desenvolver uma relação quantitativa entre várias variáveis preditivas e uma propriedade de interesse (a resposta ou variável dependente). O problema da regressão, isto é, como modelar uma ou várias variáveis dependentes, y , por meio de um conjunto de variáveis preditivas, x , é um dos mais comuns problemas no tratamento de dados analíticos em ciência e tecnologia. As variáveis dependentes em química são comumente concentrações, atividades biológicas, respostas de dados sensoriais, entre outras, ao passo que as variáveis preditivas são respectivamente, medidas espectrais, descritores físico-químicos e cromatogramas. A solução desse problema é obtida pela resolução da equação $\mathbf{Y} = \mathbf{X}\mathbf{B}$, onde \mathbf{Y} é uma matriz ou um vetor contendo as variáveis dependentes, \mathbf{X} é a matriz dos descritores e \mathbf{B} é a matriz ou o vetor de regressão, dada por $\mathbf{B} = \mathbf{X}^+\mathbf{Y}$, onde \mathbf{X}^+ é a inversa generalizada de Moore-Penrose [1,2].

A modelagem tradicional de \mathbf{Y} por meio de \mathbf{X} é baseada no uso da regressão linear múltipla (MLR do inglês *Multiple Linear Regression*) que funciona bem quando existem somente poucas variáveis em \mathbf{X} comparadas ao número de amostras (uma variável para cada 5 ou 6 amostras), e quando estas são pouco correlacionadas entre si (correlação inferior a 0,7), isto é, quando a matriz \mathbf{X} tem posto completo. As matrizes de dados podem ser muito grandes em diferentes aplicações de calibração multivariada, como por exemplo, nos estudos de relação quantitativa entre a estrutura e a atividade/propriedade (QSAR/QSPR do inglês *Quantitative Structure Activity/Property Relationship*) [3], mineração de dados (*data mining*) [4], espectroscopia no infravermelho próximo (NIR do inglês *Near Infrared Spectroscopy*) [5], ressonância magnética nuclear (NMR do inglês *Nuclear Magnetic Resonance*) [6], cromatografia [7] e estudos que tratam com matrizes aumentadas a partir de dados multímodos (*multi-way*) [8], entre outros [9]. Em tais casos, as variáveis preditoras são fortemente correlacionadas entre si de uma maneira natural, gerando matrizes \mathbf{X} mal condicionadas (posto deficiente). Portanto,

MLR não pode ser usada nestes casos, a menos que seja realizada uma cuidadosa seleção de variáveis. Para evitar o problema do mau condicionamento da matriz, métodos de projeção, tais como, regressão em componentes principais (PCR do inglês *Principal Component Regression*) ou quadrados mínimos parciais (PLS do inglês *Partial Least Squares*) são boas alternativas [10]. A ideia central de ambos os métodos é obter uma nova matriz a partir de \mathbf{X} contendo alguns poucos fatores, mas que contenha grande parte da informação presente em \mathbf{X} , e estabelecer a regressão entre a variável dependente contra estes. Os dois métodos diferem essencialmente no modo como os fatores são obtidos.

Entre os métodos de calibração multivariada, PLS é o mais popular na química. Este é um método multivariado desenvolvido em meados de 1975 a partir dos conceitos básicos de Herman Wold no campo da econometria. PLS consiste no cálculo de fatores ou variáveis latentes bem como de correlações canônicas por meio de uma sequência iterativa de regressão simples por quadrados mínimos ordinários (OLS do inglês *Ordinary Least Squares*). A versão quimiométrica da regressão por PLS foi originalmente desenvolvida por Svante Wold em 1983 como um algoritmo em dois blocos, consistindo de uma sequência de modelos parciais ajustados por quadrados mínimos [10].

Os fatores na regressão por PLS são definidos de modo a manter o compromisso entre ajustar \mathbf{X} e prever \mathbf{Y} . No caso mais simples de uma única propriedade modelada, a matriz \mathbf{Y} é reduzida a um vetor \mathbf{y} e o método é chamado de PLS1. Neste caso, cada fator que relaciona \mathbf{X} e \mathbf{y} é obtido levando em consideração a informação contida em \mathbf{y} ao maximizar a covariância entre os escores (\mathbf{t}) de \mathbf{X} e \mathbf{y} , tal que $\mathbf{X}\mathbf{w} = \mathbf{t}$ e $\mathbf{w} = \frac{\mathbf{X}'\mathbf{y}}{\|\mathbf{X}'\mathbf{y}\|}$ [10-13]. Devido à sua habilidade de manusear numerosas variáveis em \mathbf{X} altamente correlacionadas (colineares) e ruidosas, o método PLS permite a investigação de problemas mais complexos do que os anteriormente encontrados [14]. Nenhuma consideração *a priori* é feita sobre a estrutura dos modelos, mas estimativas da

confiabilidade podem ser feitas usando as abordagens “jack-knife” [15] ou de validação cruzada. A modelagem por PLS tem se tornado uma importante ferramenta em diversos campos da ciência, por exemplo, psicologia [16], economia [17], química [18], ciência dos alimentos [19], medicina e ciências farmacêuticas [20,21], entre outras.

Para os grandes conjuntos de dados utilizados atualmente, o tempo computacional é um fator que não pode ser desprezado [22], especialmente nas etapas de validação cruzada e seleção de variáveis, onde o algoritmo PLS é aplicado várias vezes [23]. Portanto, um algoritmo PLS rápido é necessário em tais situações, uma vez que o tempo computacional pode ser radicalmente reduzido durante a construção do modelo. Diversas variantes do algoritmo PLS foram desenvolvidas recentemente em uma tentativa de resolver este problema. Entre os algoritmos mais utilizados estão NIPALS [11,24], NIPALS modificado [25], Kernel [25,26], SIMPLS [27] e o PLS bidiagonal [22,28].

O propósito deste trabalho é comparar estes cinco algoritmos PLS, disponíveis na literatura, com respeito aos seus tempos computacionais e a precisão observada no erro da validação cruzada pela metodologia *leave-one-out*. Matrizes de diferentes tamanhos foram testadas objetivando encontrar qual dos algoritmos seria o mais apropriado para cada situação. Nestes testes, somente o método PLS1 (uma única variável dependente) foi considerado.

1.2. Notação

Nesta tese, seguindo o padrão usado em textos de quimiometria, escalares são definidos como letras minúsculas em itálico (*a*, *b*, *c*), vetores como letras minúsculas em negrito (**a**, **b**, **c**) e matrizes como letras maiúsculas em negrito (**A**, **B**, **C**). Elementos de matrizes são representados pela correspondente letra minúscula em itálico com a indicação da linha e da coluna em subscripto (x_{ij} é um elemento da linha *i* e da coluna *j* de **X**) e uma determinada coluna de uma matriz **X** pode ser representada pela

correspondente letra minúscula em negrito com número da coluna em subscrito (\mathbf{x}_j é a coluna j de \mathbf{X}). Em alguns casos as matrizes serão escritas explicitamente como $\mathbf{X}(m \times n)$ para indicar suas dimensões (m linhas e n colunas). A matriz identidade será representada por \mathbf{I} com suas dimensões indicadas adequadamente. Normalmente a matriz de descritores será chamada de \mathbf{X} e o vetor contendo as atividades biológicas será chamado de \mathbf{y} . Sobrescritos t e -1 representam as operações transposta e inversa, respectivamente.

1.3. Algoritmos

Cinco algoritmos foram testados com o objetivo de avaliar seus tempos computacionais e a precisão observada no erro da validação cruzada pela metodologia *leave-one-out*. Assume-se que as matrizes são adequadamente pré-tratadas. Os algoritmos são descritos no texto seguinte.

1.3.1. O Algoritmo NIPALS clássico

O primeiro algoritmo usado na regressão por PLS foi o NIPALS (*nonlinear iterative partial least squares*), apresentado em detalhes na literatura [11,24]. O NIPALS pode ser resumido como segue:

- (1) Nomeie a matriz \mathbf{X} e o vetor \mathbf{y} como \mathbf{X}_0 e \mathbf{y}_0 , respectivamente;
- (2) Calcule as quantidades \mathbf{w} (*weights* PLS para \mathbf{X}), \mathbf{t} (escores PLS para \mathbf{X}), q (*loading* PLS para \mathbf{y}) e \mathbf{p} (*loadings* PLS para \mathbf{X}):

$$\mathbf{w}_{a+1} = \mathbf{X}_a^t \mathbf{y}_a$$

$$\mathbf{w}_{a+1} = \frac{\mathbf{w}_{a+1}}{\|\mathbf{w}_{a+1}\|}$$

$$\mathbf{t}_{a+1} = \mathbf{X}_a \mathbf{w}_{a+1}$$

$$\mathbf{p}_{a+1} = \frac{\mathbf{X}_a^t \mathbf{t}_{a+1}}{\mathbf{t}_{a+1}^t \mathbf{t}_{a+1}}$$

$$q_{a+1} = \frac{\mathbf{y}_a^t \mathbf{t}_{a+1}}{\mathbf{t}_{a+1}^t \mathbf{t}_{a+1}}$$

(3) Atualize \mathbf{X} e \mathbf{y} pela subtração dos vetores latentes computados a partir destes:

$$\mathbf{X}_{a+1} = \mathbf{X}_a - \mathbf{t}_{a+1} \mathbf{p}_{a+1}^t$$

$$\mathbf{y}_{a+1} = \mathbf{y}_a - \mathbf{t}_{a+1} q_{a+1}$$

(4) Vá até o passo 2 para computar o próximo vetor latente, até alcançar A vetores latentes ($a = A$);

(5) Armazene \mathbf{w} , \mathbf{t} , \mathbf{p} e q em \mathbf{W} , \mathbf{T} , \mathbf{P} e \mathbf{q} respectivamente; onde \mathbf{W} ($J \times A$) e \mathbf{P} ($J \times A$) são as matrizes cujas colunas são os vetores \mathbf{w} e \mathbf{p} , respectivamente.

(6) Calcule os coeficientes de regressão finais \mathbf{b} , por meio da expressão $\mathbf{b} = \mathbf{W}^t (\mathbf{P}\mathbf{W}^t)^{-1} \mathbf{q}$ [29].

1.3.2. Algoritmo NIPALS modificado (NIPALS_y)

Dayal e Macgregor [25] mostraram que somente uma das matrizes \mathbf{X} ou \mathbf{Y} precisam ser atualizadas. Uma vez que somente o vetor \mathbf{y} ($I \times 1$) é atualizado após o cálculo de cada vetor latente, a velocidade do algoritmo NIPALS é melhorada.

1.3.3. Algoritmo kernel

O algoritmo kernel apresentado por Lindgren *et al.* [26] foi desenvolvido para matrizes com um número grande de objetos e relativamente poucas variáveis preditivas. Uma solução PLS completa pode ser obtida manipulando a matriz kernel condensada $\mathbf{X}^t \mathbf{y} \mathbf{y}^t \mathbf{X}$, normalmente computada usando o produto cruzado de $\mathbf{X}^t \mathbf{y}$, ou seja, $(\mathbf{X}^t \mathbf{y})(\mathbf{X}^t \mathbf{y})^t$. Este procedimento evita a necessidade de atualização da matriz kernel, e as duas

matrizes de covariância, $\mathbf{X}'\mathbf{X}$ e $\mathbf{X}'\mathbf{y}$ são de um tamanho consideravelmente menor que as matrizes originais \mathbf{X} e \mathbf{y} .

O algoritmo kernel é dado abaixo:

- (1) Compute as matrizes de covariância $\mathbf{X}'\mathbf{X}\mathbf{e}\mathbf{X}'\mathbf{y}$, e a matriz kernel $\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}$;
- (2) O vetor de *weights* PLS, \mathbf{w}_a é calculado como o autovetor correspondente ao maior autovalor da matriz $(\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X})_a$;
- (3) os *loadings* PLS \mathbf{p}_a e q_a são computados como:

$$\mathbf{p}_a^t = \frac{\mathbf{w}_a^t (\mathbf{X}'\mathbf{X})_a}{\mathbf{w}_a^t (\mathbf{X}'\mathbf{X})_a \mathbf{w}_a}$$

$$q_a = \frac{\mathbf{w}_a^t (\mathbf{X}'\mathbf{y})_a}{\mathbf{w}_a^t (\mathbf{X}'\mathbf{X})_a \mathbf{w}_a}$$

- (4) Após o cálculo de cada vetor latente, as matrizes de covariância $\mathbf{X}'\mathbf{X}$ e $\mathbf{X}'\mathbf{y}$ podem ser atualizadas como:

$$(\mathbf{X}'\mathbf{X})_{a+1} = (\mathbf{I} - \mathbf{w}_a \mathbf{p}_a^t) (\mathbf{X}'\mathbf{X})_a (\mathbf{I} - \mathbf{w}_a \mathbf{p}_a^t)$$

$$(\mathbf{X}'\mathbf{y})_{a+1} = (\mathbf{I} - \mathbf{w}_a \mathbf{p}_a^t) (\mathbf{X}'\mathbf{y})_a$$

- (5) Calcule o vetor de regressão da mesma forma do algoritmo NIPALS.

Baseado no fato de que somente a atualização de \mathbf{y} em $\mathbf{X}'\mathbf{y}$ é requerida, Dayal e MacGregor [24] propuseram uma modificação que melhorou o algoritmo kernel original e esta é a versão testada neste trabalho.

1.3.4. O algoritmo SIMPLS

O algoritmo SIMPLS, proposto por De Jong [27], deriva os fatores PLS diretamente como uma combinação linear das variáveis originais (centradas na média) em \mathbf{X} . Uma vantagem deste método é que não é necessário atualizar \mathbf{X} ou \mathbf{y} , o que pode resultar em mais rápidas computações e menor uso de memória.

Quando aplicado a uma simples variável dependente y , os resultados obtidos pelo método SIMPLS tornam-se essencialmente os mesmos daqueles obtidos pelo algoritmo NIPALS. O algoritmo SIMPLS para PLS1 pode ser resumido como segue:

(1) Compute \mathbf{s} como $\mathbf{s} = \mathbf{X}'\mathbf{y}$;

(2) Compute as quantidades \mathbf{r} (*weights* PLS para \mathbf{X}), \mathbf{t} (escores PLS para \mathbf{X}), q (*loadings* PLS para y) e \mathbf{p} (*loadings* PLS para \mathbf{X}) como seguem:

$$\mathbf{r}_a = \mathbf{s}$$

$$\mathbf{t}_a = \mathbf{X}\mathbf{r}_a$$

$$\mathbf{t}_a = \frac{\mathbf{t}_a}{\|\mathbf{t}_a\|}$$

$$\mathbf{r}_a = \frac{\mathbf{r}_a}{\|\mathbf{r}_a\|}$$

$$\mathbf{p}_a = \mathbf{X}'\mathbf{t}_a$$

$$q_a = \mathbf{y}'\mathbf{t}_a$$

(3) Armazene \mathbf{r} , \mathbf{t} , q e \mathbf{p} em \mathbf{R} , \mathbf{T} , \mathbf{q} e \mathbf{P} , respectivamente;

(4) Projete \mathbf{s} no subespaço ortogonal a \mathbf{P}_a :

$$\mathbf{s} = \mathbf{s} - \mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{s}$$

(5) Vá ao passo (2) para calcular o próximo vetor latente até alcançar A vetores latentes;

(6) Calcule o vetor de regressão como $\mathbf{b} = \mathbf{R}\mathbf{q}$.

1.3.5. O algoritmo de bidiagonalização (PLSBi)

Manne [28] mostrou que o método PLS1 é equivalente a um algoritmo desenvolvido por Golub e Kahan [2] para bidiagonalização de matrizes. A bidiagonalização matricial é uma decomposição útil frequentemente empregada como uma rápida inicialização de algoritmos para decomposição em valores singulares [1].

Este método considera que qualquer matriz \mathbf{X} ($I \times J$) pode ser escrita como $\mathbf{X} = \mathbf{U}\mathbf{R}\mathbf{V}^t$, onde \mathbf{U} ($I \times J$) e \mathbf{V} ($I \times J$) são matrizes com colunas ortonormais, ou seja, elas satisfazem as relações $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{I}$, e \mathbf{R} ($J \times J$) é uma matriz bidiagonal.

Vários artigos na literatura descrevem a relação entre o PLS1 e a decomposição bidiagonal [28,30-33]. O algoritmo PLSBi pode ser resumido como segue [30,32]:

(1) Inicialize o algoritmo para a primeira componente:

$$\mathbf{v}_1 = \frac{\mathbf{X}^t \mathbf{y}}{\|\mathbf{X}^t \mathbf{y}\|}$$

$$\alpha_1 \mathbf{u}_1 = \mathbf{X} \mathbf{v}_1$$

(2) Compute os seguintes valores para $a = 2, \dots, A$ variáveis latentes:

$$\gamma_{a-1} \mathbf{v}_a = \mathbf{X}^t \mathbf{u}_{a-1} - \alpha_{a-1} \mathbf{v}_{a-1}$$

$$\alpha_a \mathbf{u}_a = \mathbf{X} \mathbf{v}_a - \gamma_{a-1} \mathbf{u}_{a-1}$$

com

$$\mathbf{V}_A = (\mathbf{v}_1, \dots, \mathbf{v}_A), \mathbf{U}_A = (\mathbf{u}_1, \dots, \mathbf{u}_A) \text{ e}$$

$$\mathbf{R}_A = \begin{pmatrix} \alpha_1 & \gamma_1 & & & & \\ & \alpha_2 & \gamma_2 & & & \\ & & \ddots & \ddots & & \\ & & & \alpha_{A-1} & \alpha_{A-1} & \\ & & & & \alpha_A & \end{pmatrix}$$

Pode ser provado que $\mathbf{X}\mathbf{V}_A = \mathbf{U}_A\mathbf{R}_A$ e, portanto, $\mathbf{R}_A = \mathbf{U}_A^t \mathbf{X}\mathbf{V}_A$.

Uma vez computadas as matrizes \mathbf{U} , \mathbf{V} e \mathbf{R} com a truncagem de A componentes em \mathbf{R} , a pseudoinversa de Moore-Penrose de \mathbf{X} pode ser estimada e o problema de quadrados mínimos é resolvido como:

$$\mathbf{y} = \mathbf{X}\mathbf{b}$$

$$\mathbf{y} = \mathbf{U}_A \mathbf{R}_A \mathbf{V}_A^t \mathbf{b}$$

$$\mathbf{b} = \mathbf{V}_A \mathbf{R}_A^{-1} \mathbf{U}_A^t \mathbf{y}$$

1.4. Experimental

Esta seção está dividida em duas partes principais: a primeira delas trata dos conjuntos de dados simulados especialmente desenvolvidos para cobrir uma larga faixa de tamanhos de dados, com a ajuda de planejamentos experimentais fatorial e “quadrado latino”.

Na segunda parte, conjuntos de dados reais de diferentes naturezas e tamanhos foram investigados.

Por questões de clareza, as colunas das matrizes \mathbf{X} são referidas às *variáveis* e as variáveis estudadas nos planejamentos experimentais são designadas como *fatores*.

1.4.1. Conjuntos de dados simulados

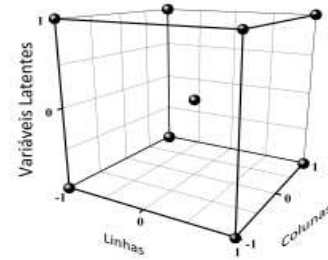
1.4.1.1 Planejamento fatorial

Dois planejamentos fatoriais completos [34,35], 2^3 , com triplicata no ponto central foram propostos para investigar dois tamanhos de conjuntos de dados: matriz \mathbf{X} pequena (SX) e matriz \mathbf{X} grande (LX). Um total de 11 experimentos foi realizado para cada planejamento, oito no nível fatorial e três no nível do ponto central. Cada algoritmo PLS foi executado para ambos os planejamentos e os experimentos no ponto central foram realizados para a estimativa do erro. As variáveis preditivas (\mathbf{X}) e dependente (\mathbf{y}) foram geradas usando um gerador de números pseudo-aleatórios. A resposta investigada no planejamento experimental foi o tempo de operação do algoritmo durante a validação cruzada pela metodologia *leave-one-out* e chamada a partir daqui como *tempo*. Três fatores foram investigados: o número de linhas, R , o número de colunas, C , de \mathbf{X} , e o número de variáveis latentes PLS, nLV . A Tabela 1.1 mostra as variáveis e o domínio explorado. As dimensões das matrizes são descritas pelos níveis de fatores linha e

coluna. Todos os dados foram centrados na média como procedimento de pré-tratamento padrão.

Tabela 1.1. Fatores, níveis codificados e domínio investigado em um planejamento fatorial completo 2^3 .

Variáveis	Níveis SX			Níveis LX		
	-1	0	1	-1	0	1
Linhas (R)	20	60	100	100	550	1000
Colunas (C)	50	275	500	500	2750	5000
Variáveis Latentes (nLV)	8	12	16	10	15	20



Assumindo que há uma relação funcional entre as variáveis experimentais e o tempo de operação dos algoritmos observado no domínio descrito, o seguinte modelo de superfície de resposta com termos lineares e de interação foi determinado:

$$\text{tempo} = \beta_0 + \beta_1 R + \beta_2 C + \beta_3 nLV + \beta_{12} R \times C + \beta_{13} R \times nLV + \beta_{23} C \times nLV + e \quad 1.1$$

O parâmetro estimado β_0 é a média de todos os valores de tempos de operação do planejamento e o parâmetro e corresponde ao erro. Os efeitos principais e de interação são os parâmetros estimados do modelo multiplicados por 2. Os efeitos podem ser também calculados pelas seguintes equações:

$$\text{Média} = \frac{\sum_{i=1}^n \text{tempo}_i}{n} \quad 1.2$$

$$\text{ef} = \frac{\sum_{i=1}^{n/2} \text{tempo}_{i(+)} - \sum_{i=1}^{n/2} \text{tempo}_{i(-)}}{n/2} \quad 1.3$$

Onde n é o número de ensaios e tempo_i é uma observação individual dada pelo tempo de operação do algoritmo PLS durante a validação cruzada pela metodologia *leave-one-out*.

A Equação 1.2 descreve o efeito médio de todas as observações, enquanto a Equação 1.3 representa os efeitos dos fatores e interações usando a diferença entre a

média das observações no nível superior (tempo_{i(+)}) e a média das observações no nível inferior (tempo_{i(-)}).

Neste trabalho, os erros padrões para os efeitos foram obtidos pela média quadrática residual, MS residual (do inglês *Mean Square* residual), de acordo com a Equação 1.4, pois, o erro puro apresentou um valor muito baixo devido à alta precisão entre as replicatas.

$$MS \text{ residual} = \frac{\sum_{i=1}^m \sum_{j=1}^r (\text{tempo}_{ij} - \hat{\text{tempo}}_i)^2}{n - q} \quad 1.4$$

Nesta equação, m é o número total de níveis (pontos do planejamento experimental); r é o número total de replicatas; $n - q$ é o número de graus de liberdade da MS residual; n é o número de ensaios, q é o número de parâmetros calculados (coeficientes ou efeitos) e $\hat{\text{tempo}}$ é o tempo de operação estimado do modelo. O erro devido ao planejamento fatorial foi obtido como descrito na Equação 1.5:

$$\text{Err} = \sqrt{\frac{MS \text{ residual}}{n}} \quad 1.5$$

1.4.1.2. Planejamento quadrado latino

Planejamentos quadrado latino são adequados quando os fatores de interesse têm mais do que dois níveis e sabe-se previamente que não existem (ou são desprezíveis) as interações entre eles. O objetivo é estimar os efeitos principais pela investigação de vários níveis para cada fator.

Um quadrado latino de ordem k é um arranjo $k \times k$ em que cada célula contém um conjunto de k símbolos, de tal modo que cada símbolo ocorra somente uma vez em cada linha e uma vez em cada coluna.

Neste trabalho, um planejamento quadrado latino 5×5 com duas replicatas foi usado para investigar a influência de vários níveis de variáveis sobre o tempo de

operação para cinco algoritmos PLS. Cinco níveis para cada fator (número de linhas, colunas e nLV) foram estudados e um total de 50 experimentos foram realizados para cada algoritmo PLS (Tabela 1.2). Todos os dados foram centrados na média como pré-tratamento padrão.

Tabela 1.2. Níveis estudados para cada fator no planejamento quadrado latino.

R^a	C^b	nLV^c
Níveis		
50	200	3
100	500	5
200	1000	10
500	5000	15
1000	10000	20

^aNúmero de linhas; ^bnúmero de colunas; ^cnúmero de variáveis latentes.

A Tabela 1.2 mostra a grande variação no número de linhas, colunas e variáveis latentes investigadas. Neste estudo, matrizes com maior número de linhas e maior número de colunas foram consideradas, cobrindo um grande número de possibilidades que podem ser encontradas no mundo real.

A avaliação estatística foi realizada usando a análise de variância (ANOVA do inglês *ANalyses Of VAriance*) bem como outras abordagens descritas na literatura [34,35].

1.4.2. Conjuntos de dados reais

Seis conjuntos de dados de aplicações reais foram explorados. Eles foram obtidos de diferentes fontes, a saber, NIR, espectroscopia Raman, espectroscopia de fluorescência, cromatografia gasosa (GC do inglês *Gas Chromatography*), voltametria e, finalmente, um conjunto de dados do tipo ultravioleta-visível (UV) simulado usando um

gerador de distribuição Gaussiana. Todos os conjuntos de dados foram investigados usando três níveis de variáveis latentes ($nLV = 3, 5$ e 10) para cada algoritmo.

Conjunto de dados NIR:

Os espectros deste conjunto de dados foram adquiridos no Southwest Research Institute (SWRI) em um projeto patrocinado pelas forças armadas dos Estados Unidos da América (US Army). Ele é formado por 231 espectros do combustível diesel medidos na faixa de comprimentos de onda entre 750 nm e 1550 nm com intervalos de 2 nm. A matriz de dados \mathbf{X} tem dimensões (231x401) e foi obtida a partir do endereço eletrônico na internet da Eigenvector Research, <http://www.eigenvector.com>. A temperatura de congelamento do combustível (°C) é a propriedade física modelada.

Conjunto de dados Raman:

Este conjunto de dados está disponível no endereço eletrônico <http://www.models.kvl.dk/research/data/> como apresentado previamente por Dyrby *et al.*[36]. Ele consiste do espalhamento Raman para 120 amostras e 3401 números de onda, na faixa de $200 - 3600 \text{ cm}^{-1}$ (intervalos de 1 cm^{-1}). A variável dependente refere-se à quantidade relativa da substância ativa em tabletes de *Escitalopram*® em unidades de porcentagem em massa (%w/w).

Conjunto de dados de fluorescência:

Este conjunto de dados foi utilizado por Bro *et al.*[37] para o estudo de vários tópicos em espectroscopia de fluorescência e pode ser encontrado no endereço eletrônico <http://www.models.kvl.dk/research/data/>. A variável dependente neste caso é a concentração de hidroquinona. Uma reorganização do arranjo de dados foi feita previamente à regressão PLS gerando uma matriz com 405 linhas e 2548 colunas.

Conjunto de dados de voltametria:

Este conjunto de dados foi obtido de Teófilo *et al.*[38] e consiste de 62 voltamogramas com correção de suas linhas de base. As variáveis para predição são as correntes de oxidação de misturas de guaiacol e cloro-guaiacol com potencial variando de 0,5 a 1,2 mV (353 variáveis). Nesta tese, o analito investigado foi o guaiacol.

Conjunto de dados do tipo UV:

Espectros com distribuição Gaussiana de quatro diferentes analitos foram usados para gerar 1000 misturas com concentrações dadas por números pseudo-aleatórios. A matriz usada neste caso foi formada por 1000 linhas e 150 colunas.

Conjuntos de dados de cromatografia gasosa:

Este conjunto de dados foi apresentado por Ribeiro *et al.*[7] e contém cromatogramas pré-tratados de 62 amostras de café torrado Brasileiro da variedade Arábica com tempos de retenção variando de 1,8 até 19 segundos com intervalo de amostragem de 0,00085 segundos (20640 variáveis). A variável dependente foi o atributo sensorial aroma das amostras de café torrado.

Todos os cálculos usando os algoritmos PLS foram realizados no MATLAB 7.0 (MathWorks, Natick, USA) com precisão dupla, instalado em um computador com Windows XP como sistema operacional, processador Intel core 2 duo com velocidade de 1,86 GHz, memória RAM de 2 GB. Os cálculos para o planejamento experimental foram realizados usando uma planilha do programa Excel de acordo com Teófilo e Ferreira [34].

A precisão dos algoritmos, considerando o erro na validação cruzada, foi definida pela diferença dos valores de raiz quadrada do erro médio da validação cruzada (RMSECV do inglês *Root Mean Square Error of Cross-Validation*) obtido em cada ensaio, de acordo com a Equação 1.6. A comparação dos resultados da validação cruzada entre os algoritmos i e j foi quantificado por:

$$\text{diff}_{ij} = |\text{RMSECV}_i - \text{RMSECV}_j| \quad 1.6$$

onde RMSECV é dada pela Equação 1.7:

$$\text{RMSECV}_k = \sqrt{\frac{\sum_{i=1}^I (y_i - \hat{y}_i)^2}{I}} \quad 1.7$$

Na Equação 1.7, y_i é a resposta medida da i -ésima amostra, \hat{y}_i é a resposta predita pela equação da calibração obtida para os dados sem a i -ésima amostra e I é o número de amostras no conjunto de calibração. A validação cruzada foi realizada utilizando um algoritmo escrito em nosso laboratório para MATLAB 7.0.

1.5. Resultados e Discussão

1.5.1. Conjuntos de dados simulados usando números aleatórios

1.5.1.1. Planejamento fatorial

Os efeitos obtidos para os cinco algoritmos considerados para os conjuntos de dados SX e LX a partir dos modelos do planejamento fatorial completo são mostrados na Tabela 1.3.

De acordo com a Equação 1.3, o efeito é a diferença entre as médias dos tempos de execução dos algoritmos obtidas para os níveis de cada fator, assim, seu valor deve estar relacionado ao tempo de execução. O efeito indica a influência de um fator ou interação entre dois fatores sobre o tempo de execução. O erro (Err) é obtido da Equação

1.5 e t é a razão entre o efeito e o erro (Efeito/Err), o parâmetro de distribuição de Student. O valor de t com graus de liberdade específicos e nível de significância, α , obtido da distribuição t disponível em tabelas estatísticas [35] ou a partir do valor de p [34,35], é usado para julgar se o efeito é estatisticamente significativo.

Como os cálculos foram realizados sob as mesmas condições para diferentes algoritmos, é possível comparar os efeitos e respostas entre os algoritmos e entre os conjuntos de dados. Portanto, observando o conjunto de dados SX da Tabela 1.3, pode ser notado que ambos os fatores (R e C) são significativos. Entretanto, C mostrou-se mais importante do que R . A interação $R \times C$ mostrou-se também importante nos cálculos, sendo até mesmo mais importante do que nLV que mostrou somente uma menor importância.

Quando analisamos os resultados para o conjunto de dados LX na Tabela 1.3, pode ser visto que somente os fatores principais R e C são significativos. Ao contrário de SX, o fator R mostrou-se mais importante do que C . Esta inversão com respeito à significância de R e C está relacionada ao procedimento de validação cruzada. O aumento no número de linhas é acompanhado do aumento do número de etapas na validação cruzada pela metodologia *leave-one-out* e, conseqüentemente, no tempo de execução. Diferentemente de SX, o fator nLV não é significativo para LX dentro dos níveis estudados. Neste caso, o efeito de nLV sobre o tempo de execução pode ter sido minimizado pela sua menor faixa relativa quando comparada a R e C . Por outro lado, a interação $R \times C$ é muito importante em todos os cálculos.

Após discutir a significância de cada fator e suas interações para os modelos, é necessário focar na diferença entre os tempos de execução para os vários algoritmos. Uma vez que o mesmo conjunto de dados foi testado por vários algoritmos, o teste t pareado é a escolha para checar se os tempos de execução dos algoritmos i e j são significativamente diferentes (Tabela 1.4).

A hipótese nula, H_0 , neste caso é que a diferença média entre os tempos de execução para os algoritmos i e j é zero, o que significa que não há evidência estatística

que os tempos computacionais são diferentes para os dois algoritmos. A hipótese alternativa é que os tempos de execução para os dois algoritmos são diferentes. Para o teste t pareado, a diferença entre os tempos de execução é calculada para cada par i, j e a média e o desvio padrão destas diferenças são calculados. Dividindo a média pelo desvio padrão entre as médias, é gerado um valor de t que segue a distribuição t com $n - 1$ graus de liberdade (df do inglês *degrees of freedom*). A hipótese nula foi rejeitada ao nível de significância de 0,05 quando t calculado é maior que o t crítico (tabelado) ou p é menor ou igual a 0,050, onde o valor de p é o menor nível de significância que levaria à rejeição de H_0 com o conjunto dado [35].

Tabela 1.3. Modelos fatoriais completos para os cinco algoritmos usando os conjuntos de dados *SX* e *LX*.

SX (Matrizes pequenas)															
	PLSBi			SIMPLS			Kernel			NIPALSy			NIPALS		
	Efeito	Err.	<i>t</i>	Efeito	Err.	<i>t</i>	Efeito	Err.	<i>t</i>	Efeito	Err.	<i>t</i>	Efeito	Err.	<i>t</i>
<i>Média</i>	<i>0,21</i>	0,01	15,96	<i>0,23</i>	0,01	16,58	<i>0,24</i>	0,02	13,90	<i>0,27</i>	0,02	13,43	<i>0,47</i>	0,05	9,40
<i>R</i>	<i>0,22</i>	0,03	7,37	<i>0,24</i>	0,03	7,26	<i>0,28</i>	0,04	6,84	<i>0,29</i>	0,05	6,26	<i>0,68</i>	0,12	5,76
<i>C</i>	<i>0,35</i>	0,03	11,50	<i>0,37</i>	0,03	11,33	<i>0,39</i>	0,04	9,74	<i>0,46</i>	0,05	9,74	<i>0,83</i>	0,12	7,09
<i>nLV</i>	<i>0,12</i>	0,03	4,01	<i>0,13</i>	0,03	3,91	<i>0,17</i>	0,04	4,14	<i>0,19</i>	0,05	4,09	<i>0,36</i>	0,12	3,03
<i>R×C</i>	<i>0,18</i>	0,03	6,06	<i>0,20</i>	0,03	6,08	<i>0,23</i>	0,04	5,70	<i>0,25</i>	0,05	5,24	<i>0,60</i>	0,12	5,09
<i>R×nLV</i>	0,05	0,03	1,68	0,07	0,03	2,02	0,08	0,04	2,02	0,09	0,05	1,92	0,23	0,12	1,97
<i>C×nLV</i>	<i>0,10</i>	0,03	3,23	0,09	0,03	2,74	<i>0,12</i>	0,04	2,99	<i>0,14</i>	0,05	3,08	0,29	0,12	2,50

LX (Matrizes grandes)															
	PLSBi			SIMPLS			Kernel			NIPALSy			NIPALS		
	Efeito	Err.	<i>t</i>	Efeito	Err.	<i>t</i>	Efeito	Err.	<i>t</i>	Efeito	Err.	<i>t</i>	Efeito	Err.	<i>t</i>
<i>Média</i>	<i>200,48</i>	29,35	6,83	<i>206,14</i>	29,66	6,95	<i>210,33</i>	30,72	6,85	<i>284,95</i>	45,05	6,32	<i>584,93</i>	93,84	6,23
<i>R</i>	<i>443,52</i>	68,83	6,44	<i>454,97</i>	69,55	6,54	<i>461,39</i>	72,05	6,40	<i>630,97</i>	105,66	5,97	<i>1294,97</i>	220,07	5,88
<i>C</i>	<i>384,65</i>	68,83	5,59	<i>399,50</i>	69,55	5,74	<i>407,76</i>	72,05	5,66	<i>561,66</i>	105,66	5,32	<i>1110,20</i>	220,07	5,04
<i>nLV</i>	108,90	68,83	1,58	107,71	69,55	1,55	117,10	72,05	1,63	175,77	105,66	1,66	396,07	220,07	1,80
<i>R×C</i>	<i>375,12</i>	68,83	5,45	<i>388,89</i>	69,55	5,59	<i>395,26</i>	72,05	5,49	<i>547,01</i>	105,66	5,18	<i>1083,66</i>	220,07	4,92
<i>R×nLV</i>	105,51	68,83	1,53	103,78	69,55	1,49	111,77	72,05	1,55	169,75	105,66	1,61	385,43	220,07	1,75
<i>C×nLV</i>	94,34	68,83	1,37	94,82	69,55	1,36	103,36	72,05	1,43	155,64	105,66	1,47	333,45	220,07	1,52

Err: resíduo MS; *t*: razão Efeito/Err.o parâmetro da distribuição de Student. *R*: número de linhas; *C*: número de colunas; *nLV*: número de variáveis latentes. Efeitos em negrito e itálico com 4 graus de liberdade são estatisticamente significativos, $\alpha = 0,05$.

Tabela 1.4. Comparação das diferenças de tempo de execução entre algoritmos usando teste t -pareado para os conjuntos de dados SX e LX .

SX (Matrizes pequenas)										
	<i>PLS_{Bi}</i>	<i>SIMPLS</i>	<i>PLS_{Bi}</i>	<i>Nipalsy</i>	<i>PLS_{Bi}</i>	<i>Nipals</i>	<i>PLS_{Bi}</i>	<i>Kernel</i>	<i>SIMPLS</i>	<i>Nipalsy</i>
Média	0,21	0,23	0,21	0,27	0,21	0,47	0,21	0,24	0,23	0,27
Variância	0,05	0,05	0,05	0,09	0,05	0,37	0,05	0,07	0,05	0,09
Correlação	1,00		1,00		0,99		1,00		1,00	
t_0	4,91		2,65		2,24		2,34		1,87	
p	0,000		0,012		0,025		0,021		0,045	
	<i>SIMPLS</i>	<i>Nipals</i>	<i>SIMPLS</i>	<i>Kernel</i>	<i>Nipalsy</i>	<i>Nipals</i>	<i>Nipalsy</i>	<i>Kernel</i>	<i>Nipals</i>	<i>Kernel</i>
Média	0,23	0,47	0,23	0,24	0,27	0,47	0,27	0,24	0,47	0,24
Variância	0,05	0,37	0,05	0,07	0,09	0,37	0,09	0,07	0,37	0,07
Correlação	0,99		1,00		0,99		1,00		0,99	
t_0	2,09		0,77		2,12		2,70		2,21	
p	0,031		0,230		0,030		0,011		0,026	
LX (Matrizes grandes)										
	<i>PLS_{Bi}</i>	<i>SIMPLS</i>	<i>PLS_{Bi}</i>	<i>Nipalsy</i>	<i>PLS_{Bi}</i>	<i>Nipals</i>	<i>PLS_{Bi}</i>	<i>Kernel</i>	<i>SIMPLS</i>	<i>Nipalsy</i>
Média	200,48	206,14	200,48	284,95	200,48	584,93	200,48	210,33	206,14	284,95
Var $\times 10^5$	1,07	1,14	1,07	2,28	1,07	9,39	1,07	1,19	1,14	2,28
Correlação	1,00		1,00		1,00		1,00		1,00	
t_0	1,79		1,86		1,98		1,88		1,85	
P	0,052		0,047		0,038		0,045		0,047	
	<i>SIMPLS</i>	<i>Nipals</i>	<i>SIMPLS</i>	<i>Kernel</i>	<i>Nipalsy</i>	<i>Nipals</i>	<i>Nipalsy</i>	<i>Kernel</i>	<i>Nipals</i>	<i>Kernel</i>
Média	206,14	584,93	206,14	210,33	284,95	584,93	284,95	210,33	584,93	210,33
Var $\times 10^5$	1,14	9,39	1,14	1,19	2,28	9,39	2,28	1,19	9,39	1,19
Correlação	1,00		1,00		1,00		1,00		1,00	
t_0	1,98		-1,50		2,02		1,85		1,99	
p	0,038		0,082		0,035		0,047		0,038	

Graus de liberdade: 10; Nível de significância: 0,05; t crítico: 1,81; Os números em negrito e itálico indicam que a hipótese nula foi aceita. Valores negativos significam que o tempo de execução do primeiro algoritmo é menor que o do segundo.

A Tabela 1.4 apresenta os resultados obtidos. Nota-se que para o conjunto de dados *SX*, o algoritmo SIMPLS foi estatisticamente igual ao kernel e para o conjunto de dados *LX*, os pares PLSBi – SIMPLS e SIMPLS – kernel foram estatisticamente iguais. Em outras comparações os tempos foram estatisticamente diferentes indicando a necessidade de avaliar que algoritmo deve ser usado.

O desempenho dos cinco algoritmos dado em termos dos tempos de execução pode ser observado na Figura 1.1(A) onde os efeitos dão uma medida do tempo computacional. Nota-se que os algoritmos PLSBi, SIMPLS e kernel mostram desempenho equivalente; NIPALS_y é ligeiramente inferior e o NIPALS tem o pior desempenho. Fica evidente que pelo uso da atualização somente em *y*, o NIPALS é significativamente melhorado com respeito ao tempo de execução.

Os efeitos para os algoritmos em *LX* são mostrados na Figura 1.1(B), onde uma tendência similar àquela do conjunto *SX* pode ser observada.

A Tabela 1.5 mostra a precisão relativa, considerando os resultados da validação cruzada, calculada como mostrado na Equação 1.6 para os algoritmos testados usando os conjuntos de dados *SX* e *LX*. Diferenças significativas entre os algoritmos são na maioria das vezes observadas, para grandes valores de *nLV*. O SIMPLS mostrou resultados notavelmente diferentes dos outros para algumas matrizes de dimensões específicas e grandes valores de *nLV*. Outros resultados indicam diferenças desprezíveis entre os algoritmos, isto é, resultados iguais para RMSECV.

Tabela 1.5. Diferença nos valores de RMSECV (Equação 1.6) entre ensaios para os conjuntos de dados SX e LX^a.

Planejamento fatorial SX			Diferença no RMSECV						
R^b	C^c	nLV^d	$Bi-Si$	$Bi-Niy$	$Bi-Ni$	$Bi-K$	$Si-Niy$	$Si-Ni$	$Si-K$
20	50	8	$1,91 \times 10^{-14}$	0	0	0	$-1,90 \times 10^{-14}$	$-1,90 \times 10^{-14}$	$-1,90 \times 10^{-14}$
100	50	8	$-7,52 \times 10^{-13}$	0	0	0	$7,51 \times 10^{-13}$	$7,51 \times 10^{-13}$	$7,51 \times 10^{-13}$
20	500	8	$2,33 \times 10^{-15}$	0	0	0	$-2,50 \times 10^{-15}$	$-2,55 \times 10^{-15}$	$-2,55 \times 10^{-15}$
100	500	8	$5,27 \times 10^{-15}$	0	0	0	$-5,27 \times 10^{-15}$	$-5,27 \times 10^{-15}$	$-5,27 \times 10^{-15}$
20	50	16	$-6,20 \times 10^{-12}$	0	0	0	$6,20 \times 10^{-12}$	$6,20 \times 10^{-12}$	$6,20 \times 10^{-12}$
100	50	16	$-1,84 \times 10^{-4}$	0	0	0	$1,84 \times 10^{-4}$	$1,84 \times 10^{-4}$	$1,84 \times 10^{-4}$
20	500	16	$1,24 \times 10^{-14}$	$5,00 \times 10^{-16}$	$6,11 \times 10^{-16}$	$5,55 \times 10^{-16}$	$-1,19 \times 10^{-14}$	$-1,18 \times 10^{-14}$	$-1,18 \times 10^{-14}$
100	500	16	$-1,77 \times 10^{-11}$	0	0	0	$1,77 \times 10^{-11}$	$1,77 \times 10^{-11}$	$1,77 \times 10^{-11}$
60	275	12	$-7,32 \times 10^{-14}$	0	0	0	$7,32 \times 10^{-14}$	$7,32 \times 10^{-14}$	$7,32 \times 10^{-14}$
60	275	12	$-3,61 \times 10^{-13}$	0	0	0	$3,61 \times 10^{-13}$	$3,61 \times 10^{-13}$	$3,61 \times 10^{-13}$
60	275	12	$-1,08 \times 10^{-13}$	0	0	0	$1,08 \times 10^{-13}$	$1,08 \times 10^{-13}$	$1,08 \times 10^{-13}$
Planejamento fatorial LX			Diferença no RMSECV						
R	C	nLV	$Bi-Si$	$Bi-Niy$	$Bi-Ni$	$Bi-K$	$Si-Niy$	$Si-Ni$	$Si-K$
100	500	10	$1,81 \times 10^{-14}$	0	0	0	$-1,82 \times 10^{-14}$	$-1,82 \times 10^{-14}$	$-1,83 \times 10^{-14}$
1000	500	10	$-1,17 \times 10^{-15}$	$-4,44 \times 10^{-16}$	0	0	$7,22 \times 10^{-16}$	$8,88 \times 10^{-16}$	$8,33 \times 10^{-16}$
100	5000	10	$1,59 \times 10^{-09}$	0	0	0	$-1,59 \times 10^{-09}$	$-1,59 \times 10^{-09}$	$-1,59 \times 10^{-09}$
1000	5000	10	$-4,33 \times 10^{-15}$	0	0	0	$4,50 \times 10^{-15}$	$4,55 \times 10^{-15}$	$4,66 \times 10^{-15}$
100	500	20	$-4,17 \times 10^{-10}$	0	0	0	$4,17 \times 10^{-10}$	$4,17 \times 10^{-10}$	$4,17 \times 10^{-10}$
1000	500	20	$3,00 \times 10^{-15}$	$-9,99 \times 10^{-16}$	$-1,05 \times 10^{-15}$	$-9,99 \times 10^{-16}$	$-4,00 \times 10^{-15}$	$-4,05 \times 10^{-15}$	$-4,00 \times 10^{-15}$
100	5000	20	$-0,10$	$-7,22 \times 10^{-16}$	$-7,22 \times 10^{-16}$	$-1,44 \times 10^{-15}$	$0,10$	$0,10$	$0,10$
1000	5000	20	$-1,25 \times 10^{-10}$	0	0	0	$1,25 \times 10^{-10}$	$1,25 \times 10^{-10}$	$1,25 \times 10^{-10}$
550	2750	15	$-3,27 \times 10^{-12}$	0	0	0	$3,27 \times 10^{-12}$	$3,27 \times 10^{-12}$	$3,27 \times 10^{-12}$
550	2750	15	$-4,52 \times 10^{-13}$	0	0	0	$4,52 \times 10^{-13}$	$4,52 \times 10^{-13}$	$4,52 \times 10^{-13}$
550	2750	15	$-3,87 \times 10^{-12}$	0	0	0	$3,87 \times 10^{-12}$	$3,87 \times 10^{-12}$	$3,87 \times 10^{-12}$

^aPLS_{Bi} (Bi), SIMPLS(Si), Kernel (K), NIPALS (Ni), NIPALS_{Sy} (Niy); ^bNúmero de linhas; ^cnúmero de colunas; ^dnúmero de variáveis latentes. Os valores para $Niy-K$, $Ni-K$ e $Niy-Ni$ são iguais a zero.

1.5.1.2. Planejamento quadrado latino

A Tabela 1.6 mostra os resultados da ANOVA para os cinco algoritmos. As somas dos quadrados (SS do inglês *Sum of Squares*) na Tabela 1.6 estão relacionadas à variância em cada fator. Quanto mais alta é a variância, mais alta é a influência de um fator no tempo de execução. A média quadrática (MS do inglês *Mean Square*) é dada pela razão entre SS e df, e explica melhor os resultados. F é o parâmetro com distribuição F para testes de variância, e é obtido como a razão de MS e MS residual. Usando o valor de F para específicos df e nível de significância, α , é possível determinar se SS é estatisticamente significativo.

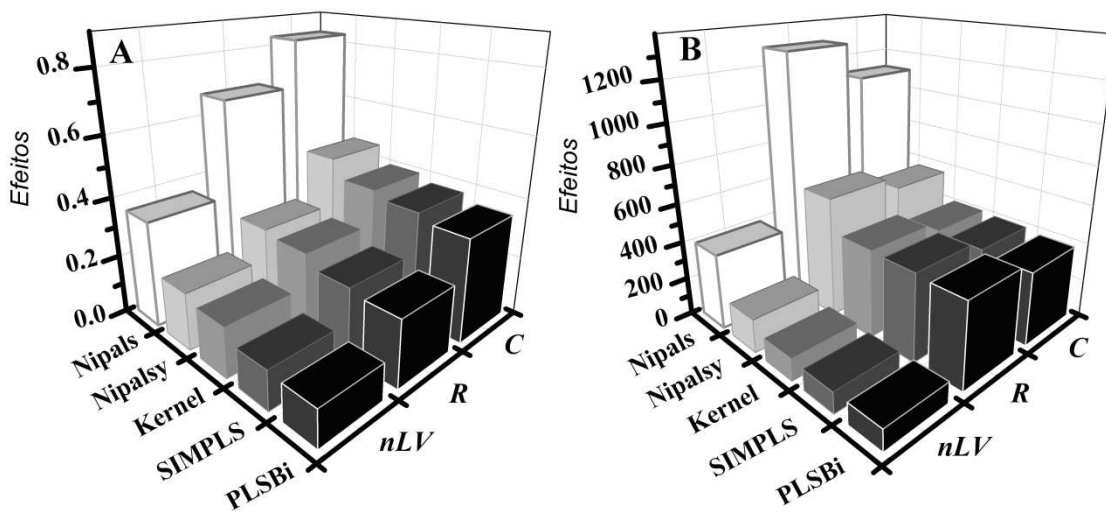


Figura 1.1. Efeitos obtidos a partir do planejamento fatorial completo para os conjuntos de dados SX (A) e LX (B).

Pela análise dos valores de SS e MS da Tabela 1.6, as similaridades entre os algoritmos PLSBi, SIMPLS e kernel, e os altos valores para NIPALSy e especialmente NIPALS, podem ser observados como antes. Neste caso, o número de linhas é mais importante do que o número de colunas, e nLV é ligeiramente menos importante quando comparado a R e C .

Quando a MS residual é usada para representar o tempo de execução, é possível observar na Figura 1.2, o comportamento dos algoritmos e a influência dos fatores R , C e nLV . O alto tempo de execução para o algoritmo NIPALS, bem como, o melhor desempenho dos algoritmos PLSBi, SIMPLS e kernel podem também ser vistos.

Tabela 1.6. Resultados da ANOVA usando planejamento quadrado latino para os cinco algoritmos^a.

PLS<i>B</i>i	SS	df	MS	F
<i>R</i>	1824949	4	456237,3	7,48
<i>C</i>	1250554	4	312638,5	5,13
<i>nLV</i>	866337	4	216584,3	3,55
Resíduo	2255381	37	60956,3	
SIMPLS	SS	df	MS	F
<i>R</i>	2001938	4	500484,6	7,74
<i>C</i>	1336032	4	334008	5,16
<i>nLV</i>	919261	4	229815,1	3,55
Resíduo	2393615	37	64692,3	
Kernel	SS	df	MS	F
<i>R</i>	1982460	4	495615	7,43
<i>C</i>	1375964	4	343991	5,15
<i>nLV</i>	947643	4	236910,8	3,55
Resíduo	2469374	37	66739,8	
NIPALS<i>y</i>	SS	df	MS	F
<i>R</i>	3618584	4	904646	7,08
<i>C</i>	2550944	4	637735,9	4,99
<i>nLV</i>	1796580	4	449145	3,52
Resíduo	4725097	37	127705,3	
NIPALS	SS	df	MS	F
<i>R</i>	14190920	4	3547730	7,09
<i>C</i>	9703320	4	2425830	4,84
<i>nLV</i>	6977808	4	1744452	3,48
Resíduo	18525600	37	500692	

^aSS: Soma de quadrados; df: graus de liberdade; MS : Resíduo quadrático médio; F: Razão estatística; α : 0,05; R : Número de linhas; C : Número de colunas; nLV : número de variáveis latentes. Valores em negrito e itálico são estatisticamente significativos.

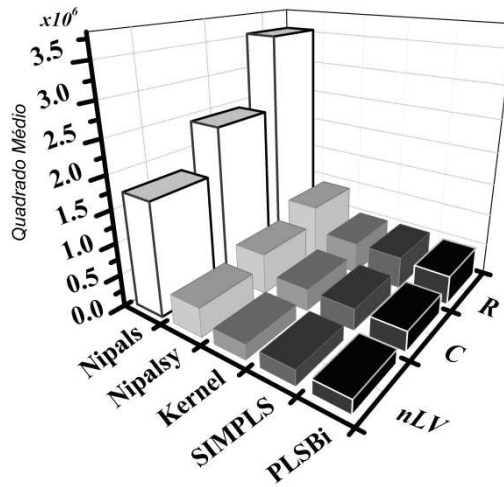


Figura 1.2. Valores de quadrado médio obtidos a partir do planejamento quadrado latino.

O gráfico para os efeitos principais é mostrado na Figura 1.3, indicando a influência do nível de cada fator no tempo computacional. Um crescimento exponencial é observado em todos os casos devido ao aumento do número de linhas e colunas. Entretanto, uma diminuição no tempo de execução é observada para o nLV máximo estudado. Esta tendência é devido à ausência de investigação para o nível máximo de nLV com o nível máximo de R e C . O nível máximo de nLV foi estudado somente para os mais baixos níveis de R e C . Como o nLV tem pouca influência no tempo de execução, o resultado obtido para o nível máximo de nLV , como notado na Figura 1.3, não é real. A influência real de nLV sobre o tempo pode ser observada na Figura 1.4 para uma dimensão de matrizes fixa, onde um aumento linear é observado.

As diferenças entre os tempos computacionais dos algoritmos resultantes do planejamento quadrado latino foram calculados e estatisticamente avaliados usando o teste t pareado como anteriormente.

A Tabela 1.7 apresenta os resultados obtidos. Nota-se que o algoritmo SIMPLS foi estatisticamente igual ao kernel, em acordo com os resultados obtidos previamente. Em outras comparações, os tempos foram estatisticamente diferentes indicando a necessidade de avaliar que algoritmo deve ser usado.

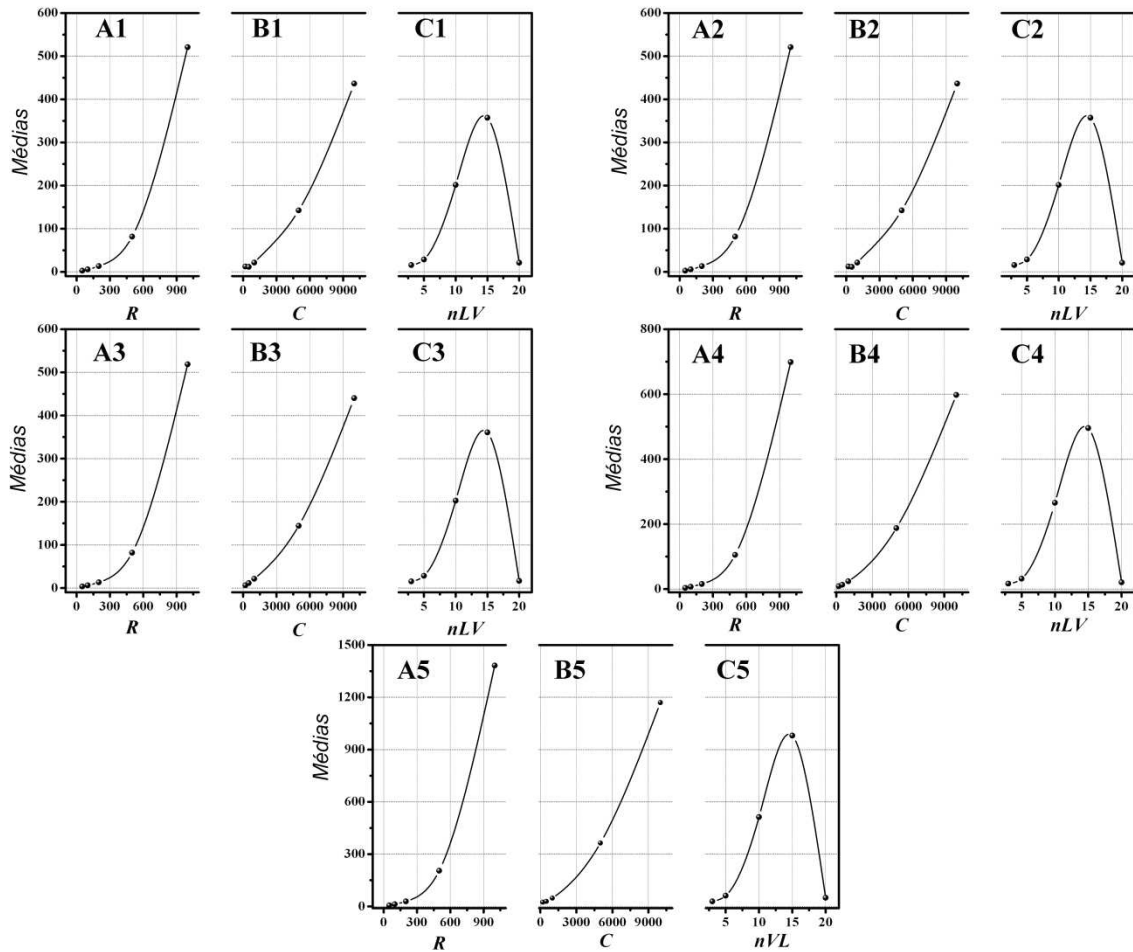


Figura 1.3. Gráficos de efeitos para o planejamento quadrado latino. PLSBi, A1, B1, C1; SIMPLS, A2, B2, C2; Kernel, A3, B3, C3; NIPALSy, A4, B4, C4; NIPALS, A5, B5, C5.

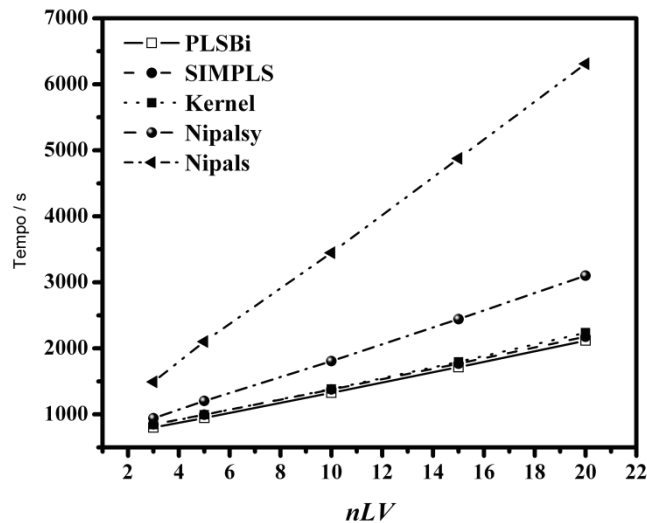


Figura 1.4. Tempo de execução versus nLV para uma matriz 1000×10000 .

Tabela 1.7. Comparação das diferenças de tempos de execução entre algoritmos usando teste *t* pareado para o conjunto de dados usado no planejamento quadrado latino.

	<i>PLS_{Bi}</i>	<i>SIMPLS</i>	<i>PLS_{Bi}</i>	<i>Nipalsy</i>	<i>PLS_{Bi}</i>	<i>Nipals</i>	<i>PLS_{Bi}</i>	<i>Kernel</i>	<i>SIMPLS</i>	<i>Nipalsy</i>
Mean	119,114	124,767	119,114	166,189	119,114	326,695	119,114	124,749	124,767	166,189
Var×10 ⁵	1,26	1,36	1,26	2,59	1,26	10,08	1,26	1,38	1,36	2,59
Corr	0,9999		0,9998		0,9997		1,0000		0,9997	
<i>t</i> ₀	-2,898		-2,169		-2,263		-2,442		-2,079	
<i>p</i>	0,0028		0,0175		0,0140		0,0091		0,0214	
	<i>SIMPLS</i>	<i>Nipals</i>	<i>SIMPLS</i>	<i>Kernel</i>	<i>Nipalsy</i>	<i>Nipals</i>	<i>Nipalsy</i>	<i>Kernel</i>	<i>Nipals</i>	<i>Kernel</i>
Mean	124,767	326,695	124,767	124,749	166,189	326,695	166,189	124,749	326,695	124,749
Var×10 ⁵	1,36	10,08	1,36	1,38	2,59	10,08	2,59	1,38	10,08	1,38
Corr	0,9996		0,9998		0,9999		0,9998		0,9996	
<i>t</i> ₀	-2,245		0,018		-2,292		2,133		2,258	
<i>p</i>	0,0146		0,4929		0,0131		0,0190		0,0142	

Grau de liberdade: 49; Nível de significância: 0,05; *t* crítico: 1,68. Os números em negrito e itálico indicam que a hipótese nula foi aceita ($p > 0,05$).

Tabela 1.8. Diferença de valores de RMSECV entre ensaios para os conjuntos de dados do planejamento quadrado latino^a.

Planejamento quadrado latino			Diferença RMSECV								
<i>R</i>	<i>C</i>	<i>nLV</i>	<i>Bi-Si</i>	<i>Bi-Niy</i>	<i>Bi-Ni</i>	<i>Bi-K</i>	<i>Si-Niy</i>	<i>Si-Ni</i>	<i>Si-K</i>	<i>Niy-K</i>	<i>Ni-K</i>
50	5000	15	-6,85×10⁻³	0	0	0	6,85×10⁻³	6,85×10⁻³	6,85×10⁻³	0	0
50	10000	20	-0,45	3,44×10⁻³	3,44×10⁻³	2,28×10⁻³	0,45	0,45	0,45	-1,16×10⁻³	-1,16×10⁻³
100	5000	20	-0,53	4,33×10 ⁻¹⁵	4,27×10 ⁻¹⁵	8,94×10 ⁻¹⁵	0,53	0,53	0,53	4,61×10 ⁻¹⁵	4,66×10 ⁻¹⁵

^aPLS_{Bi} (*Bi*), SIMPLS(*Si*), Kernel (*K*), NIPALS (*Ni*), NIPALSy (*Niy*) ; *R*: Número de linhas; *C*: número de colunas; *nLV*: número de variáveis latentes. Os valores para *Niy-Ni* são iguais a zero. Outros experimentos foram menores que 1×10⁻⁹.

A Tabela 1.8 mostra a análise das precisões, considerando os resultados da validação cruzada, para o planejamento quadrado latino. Três ensaios indicam uma grande diferença nos valores de RMSECV. Observando estes valores pode ser concluído que o número de variáveis latentes é crítico para matrizes onde o número de amostras é aproximadamente 2% (ou menor) do que o número de variáveis. Com um alto número de variáveis latentes (maior que 10), grandes diferenças entre os resultados com os algoritmos PLS, principalmente para o algoritmo SIMPLS, podem ser observadas.

Entretanto, as diferenças entre os RMSECV para os outros ensaios são muito pequenas (inferior a 10^{-9}), indicando que os algoritmos são bastante similares considerando a precisão na maioria dos casos.

1.5.2. Conjunto de dados reais

Os espectros, voltamogramas e cromatogramas são mostrados na Figura 1.5. Nota-se que cada um dos conjuntos de dados mostra uma característica diferente. Estes dados foram estudados variando o número de variáveis latentes para cada algoritmo aplicado.

A Tabela 1.9 contém os tempos de execução obtidos para cada conjunto de dados real e algoritmo. Nota-se que o tempo computacional aumenta linearmente com nLV para todos os algoritmos e conjuntos de dados. O melhor desempenho foi obtido para o algoritmo PLSBi, enquanto o pior desempenho foi, na maioria dos casos, obtido para o NIPALS. O algoritmo kernel foi ligeiramente melhor do que o SIMPLS para todos os ensaios.

Foi observado, na maioria dos casos, que o tipo de dado não afetou o comportamento das diferenças dos tempos computacionais entre os algoritmos com respeito aos dados aleatórios. Entretanto, o conjunto de dados do tipo UV apresentou um resultado inesperado onde o algoritmo SIMPLS mostrou o pior desempenho (Tabela 1.9). Este resultado anormal pode ser justificado pelas dimensões das matrizes utilizadas

neste conjunto de dados. O número de linhas (1000) é muito maior do que o número de colunas (150). Assim, não se recomenda o uso do SIMPLS para matrizes com tais dimensões.

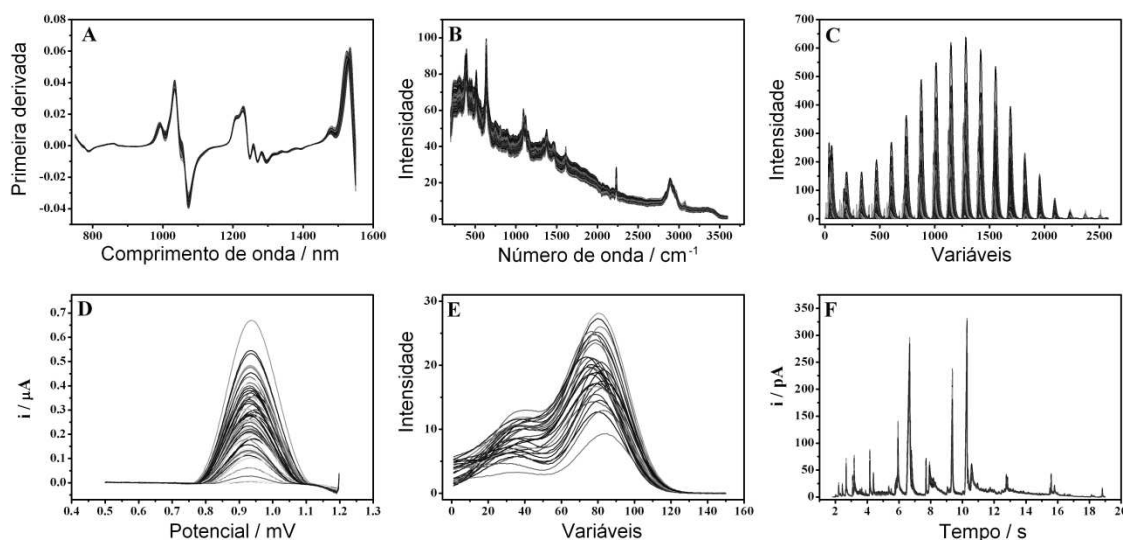


Figura 1.5. Conjuntos de dados de testes utilizados. A: espectros de infravermelho próximo (NIR); B: espectros raman (Raman); C: espectros de fluorescência em forma de matriz (Fluor); D: voltamogramas (Volt); E: conjunto de dados tipo UV (Tipo UV); F: cromatografia gasosa (CG).

A diferença entre os valores de RMSECV (precisão relativa dos erros da validação cruzada) para os seis conjuntos de dados reais situam-se entre $1,45 \times 10^{-5}$ e $7,37 \times 10^{-18}$, indicando que não existem diferenças significativas entre os algoritmos, com respeito à precisão dos erros da validação cruzada.

Tabela 1.9. Tempo (em segundos) de cada algoritmo variando o tipo de conjunto de dados, dimensão e número de variáveis latentes.

Conjunto de dados	Dimensão	nLV^a	PLSBi	SIMPLS	Kernel	NIPALSy	NIPALS
NIR	231×401	3	1,16	1,24	1,17	1,27	2,33
		5	1,39	1,48	1,41	1,56	3,17
		10	1,86	1,94	1,91	2,23	5,25
Raman	120×3401	3	3,30	3,61	3,58	3,98	6,78
		5	4,17	4,64	4,53	5,34	9,72
		10	6,59	7,20	7,66	9,03	17,49
Fluor	405×2584	3	27,98	29,58	29,50	32,38	55,77
		5	33,52	35,17	35,14	40,80	77,38
		10	47,88	49,81	50,63	63,33	132,88
Volt	62×353	3	0,11	0,11	0,11	0,11	0,19
		5	0,14	0,14	0,13	0,14	0,28
		10	0,20	0,23	0,22	0,25	0,50
Tipo UV	1000×150	3	8,55	27,38	8,39	9,36	17,27
		5	9,88	28,25	9,41	10,92	23,38
		10	13,41	30,72	12,38	15,63	38,84
CG	58×20640	3	67,24	73,95	67,36	78,17	110,06
		5	74,72	78,89	78,92	91,09	144,00
		10	106,63	113,23	118,39	144,00	249,64

^a: Número de variáveis latentes

1.6. Conclusões

A escolha do algoritmo PLS para a construção de modelos multivariados de regressão é uma importante questão quando se trabalha com grandes conjuntos de dados, devido às diferenças significativas nos tempos de execução dos algoritmos apresentados na literatura e, em alguns casos, devido à importância na diferença entre os valores de RMSECV.

Para as matrizes analisadas neste trabalho, foi mostrado que a dimensão das mesmas é o principal fator responsável pelo tempo computacional, enquanto o número de variáveis latentes no modelo tem uma influência secundária. Adicionalmente, na maioria dos casos o número de linhas tem maior influência do que o número de colunas

para os algoritmos. O número de variáveis latentes exibe uma influência linear com o aumento no tempo, mas é menos importante do que a influência do número de linhas e colunas.

Entre os cinco algoritmos analisados neste trabalho, PLSBi foi considerado como o melhor com respeito ao tempo computacional, seguido pelos algoritmos kernel e SIMPLS, sendo que as diferenças no tempo de execução, embora relativamente pequenas, são estatisticamente diferentes. Comparando NIPALS com o NIPALSy, este último foi estatisticamente mais rápido, por que somente os valores de y precisam ser atualizados. Os valores de RMSECV para todos os algoritmos testados foram essencialmente os mesmos na maioria dos casos. Entretanto, diferenças pronunciadas entre os valores de RMSECV foram observadas em alguns ensaios específicos (com um alto número de variáveis latentes), especialmente para o algoritmo SIMPLS. Investigações futuras são necessárias para uma explicação teórica deste comportamento.

Capítulo 2

Fundamentação teórica sobre QSAR e quimiometria aplicada a QSAR

2.1. Uma introdução aos estudos de QSAR

Um ramo da Química de grande interesse é o planejamento de fármacos com o auxílio do computador. A possibilidade de projetar compostos com propriedades bem definidas evitando os custos da síntese experimental exploratória de um grande número de substâncias tem impulsionado muita pesquisa nessa área.

Grande parte dos fundamentos necessários para um projeto efetivo no planejamento de um fármaco com o auxílio do computador estão na relação quantitativa estrutura atividade, QSAR (do inglês *Quantitative Structure-Activity Relationship*). Nas técnicas utilizadas em QSAR considera-se que existe uma relação entre as propriedades macroscópicas de um composto e sua estrutura molecular (a nível microscópico). Relações matemáticas simples são estabelecidas para tentar descrever e, em seguida, prever uma dada propriedade para um conjunto de compostos, geralmente pertencentes a uma mesma família química. O estudo de QSAR compreende também a definição dos descritores moleculares capazes de caracterizar satisfatoriamente conjuntos moleculares diferentes e o tratamento estatístico que pode ser aplicado a esses descritores a fim de melhorar sua capacidade preditiva.

O objeto de um estudo de QSAR é a busca por relações quantitativas entre a estrutura química, isto é, propriedades físico-químicas, estruturais e conformacionais, e a resposta biológica. Estas relações ajudam a entender e explicar o mecanismo de ação dos

fármacos, atualmente servindo de base para o desenvolvimento de novos compostos que exibam propriedades biológicas desejáveis [20].

Os princípios das técnicas que hoje são utilizadas em QSAR surgiram em 1863 quando Crois, da universidade de Estrasburgo, observou que a toxicidade de álcoois em mamíferos aumentava quando suas solubilidades em água diminuían. Crum-Brown e Fraser postularam em 1868 a existência de uma relação entre as atividades fisiológicas e as estruturas químicas. Mais tarde, Richet propôs que a toxicidade de alguns álcoois e éteres era inversamente proporcional à suas solubilidades em água. Por volta de 1900, Meyer e Overton, trabalhando independentemente, estabeleceram relações lineares entre a ação narcótica de alguns compostos orgânicos e seus respectivos coeficientes de solubilidade em água e em lipídios, descrevendo um parâmetro que pode ser considerado como um precursor do atual $\log P$, o coeficiente de partição octanol-água. Em 1939, Ferguson estudou o comportamento de propriedades diversas (solubilidade em água, partição, capilaridade, e pressão de vapor) em relação à atividade tóxica de diferentes séries homólogas de compostos [39].

Mesmo considerando estes procedimentos como as raízes do atual QSAR, Hammett propôs no final da década de 30 o primeiro procedimento metodológico de propósito geral [40]. Hammett verificou que as constantes de equilíbrio de ionização dos ácidos benzóicos meta e para substituídos estavam relacionadas aos substituintes nessas posições. Esta relação levou à definição da chamada constante de Hammett, σ . Este parâmetro tornou-se um descritor capaz de caracterizar a atividade biológica de muitos conjuntos de moléculas.

Em 1964, Free e Wilson postularam que para uma série de compostos similares, diferindo entre si apenas pela presença de certos substituintes, a contribuição destes substituintes para a atividade biológica seria aditiva e dependeria apenas do tipo e da posição do substituinte [41].

A sistematização das análises em QSAR, no entanto, deve ser associada ao trabalho de Hansch e Fujita publicado em 1964 [42]. As bases para o modelo de Hansch-

Fujita é considerar que a atividade biológica observada é o resultado da contribuição de diversos fatores que se comportam de maneira diferente. Cada contribuição para a atividade é representada por um descritor estrutural, e a atividade biológica de um conjunto de compostos é ajustada em um modelo multilinear. Os descritores mais utilizados nas primeiras análises de QSAR foram o coeficiente de partição octanol/água ($\log P$), a constante de Hammett σ agindo como um descritor eletrônico e o parâmetro de lipofilicidade π , definido em analogia ao descritor eletrônico.

Com o avanço da química quântica e, principalmente, dos computadores, foi possível incluir juntamente com esses descritores empíricos outras propriedades físico-químicas, algumas das quais derivadas de cálculos mecânico-quânticos. Dentre elas podemos citar, por exemplo, as cargas atômicas parciais, o momento de dipolo, as energias do orbital molecular mais alto ocupado (HOMO) e do orbital mais baixo desocupado (LUMO), a polarizabilidade e a refratividade molar, etc.

Outra classe de descritores bastante usada em estudos de QSAR é baseada nos conceitos de topologia molecular. Esta perspectiva, desenvolvida principalmente por Wiener [43], Kier e Hall [44] e Randic [45], representa numericamente as características topológicas das moléculas através dos chamados índices de conectividade e de distância baseados na teoria dos grafos. Hoje em dia inúmeros descritores topológicos podem ser usados em QSAR, grande parte deles implementada nos programas DRAGON [46] e MARWIN [47].

Outros descritores usados em estudos de QSAR clássico são: i) descritores constitucionais, que levam em conta elementos constituintes da molécula, como número de ligações, massa molecular, número de átomos aromáticos, etc; ii) descritores geométricos, que dependem do arranjo espacial dos átomos constituintes da molécula, como volume molecular e área superficial molecular; iii) descritores eletrotológicos, que levam em conta as características eletrônicas e topológicas em conjunto, como os índices e-state [3].

Estes são apenas alguns exemplos dos diversos tipos de descritores que podem ser empregados para a construção de modelos de QSAR 2D ou clássico.

2.2. QSAR-3D

Em 1988, as técnicas de QSAR sofreram uma grande transformação devido à introdução dos chamados parâmetros moleculares tridimensionais, que levam em conta a influência de diferentes conformêros, estereoisômeros ou enantiômeros. Este tipo de método, conhecido como QSAR-3D, também implica no alinhamento das estruturas moleculares de acordo com um farmacóforo comum, derivado do conhecimento da interação fármaco-receptor. O primeiro trabalho publicado utilizando essa metodologia foi a análise comparativa de campo molecular (CoMFA¹), proposta por Cramer [48], difundida e muito utilizada pelos químicos e cientistas de áreas correlatas, tornando-se uma ferramenta fundamental em estudos QSAR-3D [49,50]. No formalismo CoMFA, PLS [10,12,20,27,28,51] é o método de regressão usado para modelar a relação entre a atividade biológica de um conjunto de compostos com um alinhamento específico e seus campos de energia 3D. Estes campos são determinados em uma caixa virtual, chamada de *grid*, que contém todas as estruturas químicas alinhadas. A etapa de um planejamento racional de um fármaco que utiliza QSAR-3D pode ser dividida em três partes: alinhamento das moléculas, geração de campos moleculares e regressão com um ou mais parâmetros de atividades biológicas como resposta [51].

Em primeiro lugar, as conformações obtidas a partir da geometria otimizadas moléculas são alinhadas por superposição de pontos de possíveis interações, átomos em moléculas, por exemplo, com uma proteína que seria um receptor alvo.

Um campo molecular é uma abstração para o conjunto formado pelo ligante e um receptor rígido hipotético dado por uma caixa ou *grid* tridimensional suficientemente

1 do inglês *Comparative Molecular Field Analysis*

grande para conter todas as moléculas alinhadas onde, em cada ponto do grid, as interações entre uma sonda e cada molécula são calculadas. Assim, as interações eletrostáticas e estéricas, calculadas em cada ponto no grid com base em potenciais de Coulomb e Lennard-Jones respectivamente, correspondem às variáveis ou descritores em CoMFA. Esses potenciais são dados pelas equações 2.1 e 2.2.

$$E_c = \sum_{i=1}^{i=n} \frac{q_i q_j}{D r_{ij}} \quad 2.1$$

$$E_{vdW} = \sum_{i=1}^{i=n} A_{ij} r_{ij}^{-12} + C_{ij} r_{ij}^{-6} \quad 2.2$$

Nas equações acima E_c é o potencial de Coulomb, n é o número de átomos na molécula, q_i é a carga atômica parcial do átomo i da molécula, q_j é a carga da sonda, D é a constante dielétrica, r_{ij} é a distância entre o átomo i e a sonda, E_{vdW} é a energia de interação de van der Waals e A_{ij} e C_{ij} são constantes que dependem dos raios de van der Waals do átomo i e da sonda.

A metodologia CoMFA vem sendo bastante usada em estudos de QSAR desde que foi proposta em 1988. Uma busca na base de dados *Web of Science* com a palavra chave CoMFA em 22/01/2013 trouxe 2189 resultados desde o primeiro artigo de Cramer *et al* [48].

2.3. QSAR-4D

Em 1997, Hopfinger e colaboradores propuseram uma nova metodologia de QSAR chamada de QSAR-4D [52]. A análise na metodologia QSAR-4D incorpora a liberdade conformacional ao desenvolvimento de modelos usando a metodologia QSAR-3D fazendo com que a mudança de estado molecular, observada a partir de

simulações de dinâmica molecular, constitua a quarta dimensão. Os descritores em QSAR-4D são representados pelas medidas de ocupação de cada célula de uma caixa virtual, chamada de *grid*, pelos átomos que formam as moléculas do conjunto de treinamento. Os descritores de ocupação das células do grid, GCODs (*grid cell occupancy descriptors*), podem ser gerados a partir de diferentes tipos de átomos (polar positivo, polar negativo, apolar, aromático, doador de ligação de hidrogênio, acceptor de ligação de hidrogênio), que em QSAR 4D são chamados de IPEs (*interaction pharmacophore elements*). A equação 2.3 mostra como são calculados os descritores na metodologia QSAR-4D.

$$A_i(x, y, z, a) = \frac{1}{N} \times \sum_{n=1}^{n=N} O_i(x, y, z, a, n) \quad 2.3$$

Na equação acima $A_i(x, y, z, a)$ representa a ocupação absoluta da célula localizada na coordenada (x, y, z) para um IPE do tipo a em relação ao composto i . N representa o número de conformações (passos) recuperadas da dinâmica molecular e é dado pelo tempo total da dinâmica dividido pelo intervalo de tempo determinado para cada passo ($N = T/\Delta t$). $O_i(x, y, z, a, n)$ representa a ocupação da célula do grid localizada na coordenada (x, y, z) por um IPE do tipo a na conformação n . Se nenhum átomo do tipo a ocupa a célula (x, y, z) na conformação n , então $O_i(x, y, z, a, n) = 0$. Se m átomos do tipo a ocuparem a célula (x, y, z) na conformação n , então $O_i(x, y, z, a, n) = m$. O somatório é dividido pelo número de conformações (N) para que os dados sejam normalizados.

Assume-se em uma análise de QSAR-4D que as diferenças nos dados de atividades biológicas estão relacionadas às diferenças existentes na distribuição espacial média de Boltzmann da forma molecular em relação aos IPEs. Uma única conformação ativa pode ser postulada para cada composto no conjunto de treinamento e, quando

combinada com o alinhamento ótimo, pode ser usada posteriormente em aplicações de planejamento molecular incluindo outros métodos de QSAR-3D.

A análise QSAR-4D, através do uso dos IPEs, permite que cada um dos compostos em um conjunto de treinamento possa ser particionado em conjuntos de classes com respeito a possíveis interações com um receptor comum. Os GCODs, definidos pelos IPEs, são simultaneamente mapeados em um grid comum (Figura 2.1). Os dez passos operacionais necessários para uma análise QSAR-4D são mostrados na Tabela 2.1.

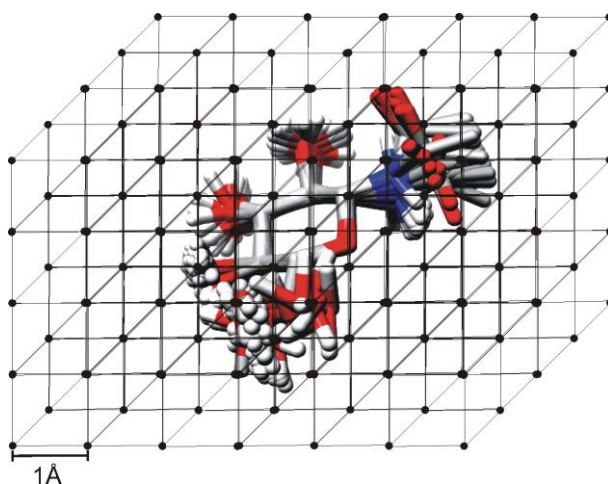


Figura 2.1. Exemplo de um CEP dentro de um grid onde podem ser calculados os descritores de ocupação em 4D-QSAR

Tabela 2.1. Dez passos operacionais realizados na análise QSAR 4D

Passo	Descrição da operação
1	Gerar o grid de referência e os modelos 3D iniciais para todos os compostos no conjunto de treinamento.
2	Selecionar os IPEs.
3	Geração dos perfis conformacionais para cada composto (CEP).

- 4 Selecionar um alinhamento
 - 5 Colocar cada conformação de cada composto no grid de referência de acordo com o alinhamento, salvar o GCOD para cada IPE e escolher uma medida de ocupação para o CEP
 - 6 Realizar uma regressão PLS para reduzir a matriz de descritores formada pelos GCODs para uma matriz formada por um conjunto menor de fatores.
 - 7 Utilizar os GCODs com maiores pesos obtidos durante a regressão PLS e quaisquer outros descritores escolhidos pelo usuário para um conjunto inicial em uma otimização do modelo QSAR-4D com algoritmo genético.
 - 8 Retornar ao passo 4 e repetir os passos 4 – 7 até que todos os alinhamentos desejados tenham sido incluídos na análise.
 - 9 Selecionar o modelo QSAR-4D ótimo com respeito ao alinhamento e quaisquer outros parâmetros da metodologia.
 - 10 Selecionar o estado conformacional de baixa energia, a partir do conjunto CEP, para cada composto que prevê a atividade máxima utilizando o modelo QSAR-4D ótimo como a conformação ativa.
-

2.4. Estudos de QSAR que resultaram em fármacos hoje no mercado

Atualmente, existem diversos exemplos de fármacos já disponíveis no mercado que foram planejados a partir de estudos de QSAR. Alguns exemplos são mostrados a seguir.

Em sua revisão, Fujita [53] reporta alguns exemplos de estudos de QSAR bem sucedidos. Dentre esses exemplos estão uma benzidrilbenzilpiperazina agente contra enxaqueca, antifúngicos agrícolas do tipo azola e um agente anti-inflamatório obtido a partir de uma série de ácidos 4-bifenilil-4-oxobutanóico, que levou ao anti-inflamatório flobufen. Krohn *et al.* [54] utilizaram modelagem molecular e estudos de QSAR para chegarem a um inibidor da protease HIV-1 contendo hidróxi-etilamina, que tornou o composto protótipo do fármaco squinavir, primeiro inibidor de protease aprovado pelo FDA (do inglês *Food and Drug Administration*) em 1995.

O inibidor de acetilcolinesterase (AChE) cloridrato de donepezila, conhecido como Aricept, usado no tratamento do mal de Alzheimer e liberado pelo FDA em 1996, é membro de uma família de inibidores da AChE baseados na N-benzilpiperidina sintetizados e avaliados pela companhia Eisai do Japão [55] com base em estudos de QSAR realizados por Cardozo *et al.* [56].

2.5. Quimiometria aplicada aos estudos de QSAR

Conforme mencionado anteriormente, em um estudo de QSAR o principal objetivo é encontrar um modelo matemático que relacione as propriedades de um conjunto de compostos e as atividades biológicas medidas para esses compostos. Este modelo matemático é obtido com o auxílio da quimiometria.

2.5.1. Construção do modelo matemático

A relação entre os descritores moleculares e as propriedades físico-químicas ou biológicas pode ser feita de maneira linear. Desse modo, a equação obtida é:

2.4

$$\mathbf{y} = b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \dots + b_j\mathbf{x}_j + \mathbf{e}$$

onde \mathbf{y} é um vetor I -dimensional contendo as propriedades ou atividades da família molecular estudada, \mathbf{X} ($I \times J$) é a matriz de descritores, \mathbf{e} é um vetor de erros normalmente distribuídos. Os estimadores b_i são chamados de coeficientes de regressão e o objetivo da análise de regressão é encontrar esses coeficientes. Quando se usa a matriz de descritores \mathbf{X} diretamente na equação 2.4, o método de regressão é conhecido como regressão linear múltipla, MLR (do inglês *Multiple Linear Regression*), ou quadrados mínimos ordinários, OLS (do inglês *Ordinary Least Squares*). No entanto, pode-se usar no

lugar da matriz \mathbf{X} uma matriz derivada dela contendo combinações lineares das variáveis em \mathbf{X} . Os principais métodos que usam desse expediente são a regressão de componentes principais, PCR do inglês *Principal Component Regression*, e a regressão de quadrados mínimos parciais, PLS (do inglês *Principal Component Regression*).

2.5.1.1. Regressão Linear Múltipla (MLR)

A regressão linear múltipla foi o primeiro método de regressão multivariada usado em QSAR e consiste na resolução da equação 2.4 utilizando diretamente a matriz de descritores \mathbf{X} . Colocada na forma matricial e considerando-se que a matriz \mathbf{X} e o vetor \mathbf{y} estejam centrados na média, a equação 2.5 pode ser escrita como:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \quad 2.5$$

onde \mathbf{b} é o vetor que contém os coeficientes de regressão $b_j (j = 1, 2, \dots, J)$. O objetivo da regressão linear é encontrar o vetor \mathbf{b} de modo a minimizar o erro \mathbf{e}

$$\min \|\mathbf{e}\|_2 = \min \|\mathbf{y} - \mathbf{Xb}\|_2 \quad 2.6$$

Em matemática esse problema é conhecido como problema de quadrados mínimos $\|\mathbf{e}\|_2$ e é a norma-2 do vetor \mathbf{e} . A solução para esse problema é encontrada projetando-se o vetor \mathbf{y} no espaço gerado pelas colunas de \mathbf{X} , que equivale a dizer que \mathbf{e} está no núcleo de \mathbf{X}^t . Assim temos:

$$\begin{aligned} \mathbf{X}^t \mathbf{e} &= \mathbf{X}^t (\mathbf{y} - \mathbf{Xb}) = 0 \\ \mathbf{X}^t \mathbf{y} &= \mathbf{X}^t \mathbf{Xb} \\ \mathbf{b} &= (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y} \end{aligned} \quad 2.7$$

Observando a equação 2.7, pode-se perceber que ela só tem solução se a matriz $\mathbf{X}^t\mathbf{X}$ possuir inversa. Isso só acontece se o número de colunas da matriz \mathbf{X} (descritores) for menor que o número de linhas (compostos) e se todas as colunas de \mathbf{X} forem linearmente independentes, o que equivale a dizer que os descritores não podem ser correlacionados. No entanto, em estudos de QSAR, normalmente o número de descritores é maior que o número de amostras e muitos deles são correlacionados entre si. Assim, o método MLR não pode ser usado nesses casos, a menos que uma cuidadosa seleção de variáveis seja feita.

Para contornar esse problema costuma-se usar métodos de projeção como PCR e PLS. A idéia central desses métodos é substituir os descritores originais por variáveis latentes, que são combinações lineares dos descritores originais e carregam grande parte da informação contida neles, e fazer a regressão com essas novas variáveis.

2.5.1.2. Regressão de componentes principais (PCR)

A ideia principal na regressão de componentes principais é substituir os descritores originais por um subconjunto de componentes principais de \mathbf{X} . Essas componentes são sucessivas combinações lineares das colunas de \mathbf{X} (descritores) que levam em conta a máxima variação possível sujeita a restrições de ortogonalidade e de tamanho do vetor de pesos. Assim, cada componente principal é dada por:

$$\mathbf{t}_i = \mathbf{X}\mathbf{p}_i \quad \text{para } i = 1, 2, \dots, A \quad 2.8$$

onde A é o número de componentes principais extraídas de \mathbf{X} , cujo valor máximo é o menor valor entre I e J , e \mathbf{p} é o vetor de pesos. Este vetor tem norma-2 igual a 1 e corresponde a um autovetor da matriz de variância $\mathbf{X}^t\mathbf{X}$

$$\mathbf{X}^t \mathbf{X} \mathbf{p}_i = \lambda_i \mathbf{p}_i \quad 2.9$$

onde λ_i é o autovalor correspondente. Estes autovetores formam os eixos no novo sistema de coordenadas no qual as variáveis originais são projetadas. Convenciona-se que os autovalores estão em ordem decrescente, ou seja, $\lambda_1 > \lambda_2 > \dots > \lambda_A$. Multiplicando-se à esquerda ambos os lados da equação 2.10 por \mathbf{p}_i^t , pode-se notar facilmente que

$$\begin{aligned} \mathbf{p}_i^t \mathbf{X}^t \mathbf{X} \mathbf{p}_i &= \lambda_i \mathbf{p}_i^t \mathbf{p}_i \\ \mathbf{t}_i^t \mathbf{t}_i &= \lambda_i \end{aligned} \quad 2.10$$

e, portanto, que a variância de uma componente principal é proporcional ao seu autovalor correspondente. Além disso, devido à restrição de ortogonalidade entre os vetores \mathbf{p}_i pode-se perceber que:

$$\begin{aligned} \mathbf{p}_j^t \mathbf{X}^t \mathbf{X} \mathbf{p}_i &= \lambda_i \mathbf{p}_j^t \mathbf{p}_i \\ \mathbf{t}_j^t \mathbf{t}_i &= 0 \end{aligned} \quad 2.11$$

ou seja, uma dada componente principal é ortogonal a todas as outras. A decomposição bilinear de \mathbf{X} , conhecida como análise de componentes principais (PCA do inglês *principal component analysis*), é expressa algebricamente como:

$$\hat{\mathbf{X}} = \mathbf{t}_1 \mathbf{p}_1^t + \mathbf{t}_2 \mathbf{p}_2^t + \dots + \mathbf{t}_A \mathbf{p}_A^t = \sum_{i=1}^A \mathbf{t}_i \mathbf{p}_i^t = \mathbf{T}_A \mathbf{P}_A^t \quad 2.12$$

onde a matriz \mathbf{T} , chamada de matriz de escores, tem como colunas os vetores \mathbf{t} e a matriz de pesos \mathbf{P} tem como colunas os vetores \mathbf{p} .

Com essa decomposição, pode-se considerar que com apenas as primeiras componentes principais tem-se uma boa representação de \mathbf{X} , já que as últimas componentes representam pouca variação em \mathbf{X} , o que pode ser insignificante ou apenas ruído.

Se inserirmos $\lambda_i^{-1/2}\lambda_i^{1/2}$ entre \mathbf{t}_i e \mathbf{p}_i^t na equação 2.12 temos:

$$\hat{\mathbf{X}} = \sum_{i=1}^A (\mathbf{t}_i \lambda_i^{-1/2}) \lambda_i^{1/2} \mathbf{p}_i^t = \sum_{i=1}^A \mathbf{u}_i \sigma_i \mathbf{v}_i^t = \mathbf{USV}^t \quad 2.13$$

que mostra a equivalência entre a análise de componentes principais e a decomposição de valores singulares (SVD²). Na equação 2.13, \mathbf{u}_i representa um vetor singular à esquerda, σ_i representa um valor singular e $\mathbf{v}_i = \mathbf{p}_i$ representa um vetor singular à direita.

A matriz \mathbf{X} é então projetada em um novo sistema de coordenadas em que os novos eixos são representados pelos vetores \mathbf{p}_i e os vetores \mathbf{t}_i , para i variando de 1 até A , são as coordenadas das amostras nesse novo sistema. Como grande parte da variação em \mathbf{X} pode ser expressa em poucas componentes principais, a matriz \mathbf{T} pode ser usada agora na resolução do problema de quadrados mínimos, pois o número de colunas é menor que o número de linhas e essas colunas são ortogonais entre si. Assim, de maneira análoga ao que foi feito em MLR, temos:

$$\mathbf{y} = \mathbf{Tq} + \mathbf{e} \quad 2.14$$

com

$$\mathbf{q} = (\mathbf{T}^t \mathbf{T})^{-1} \mathbf{T}^t \mathbf{y} \quad 2.15$$

É interessante, no entanto, termos uma equação que relacione diretamente \mathbf{X} e \mathbf{y} , isto é $\hat{\mathbf{y}} = \mathbf{Xb}$, já que é fundamental em QSAR a interpretação da equação obtida em

2 do inglês *Singular Values Decomposition*

termos dos descritores originais. Como $\hat{\mathbf{y}} = \mathbf{T}\mathbf{q} = \mathbf{X}\mathbf{b}$, substituindo \mathbf{T} por $\mathbf{X}\mathbf{P}$ percebe-se facilmente que:

$$\mathbf{b} = \mathbf{P}\mathbf{q} \quad 2.16$$

Assim, uma equação de regressão pode ser obtida em termos dos descritores originais mesmo que a matriz \mathbf{X} tenha mais descritores do que compostos e que existam descritores correlacionados entre si, pois a matriz usada na regressão (matriz de escores \mathbf{T}) é uma matriz que atende os requisitos para a resolução do problema de quadrados mínimos.

2.5.1.3. Regressão de quadrados mínimos parciais (PLS)

As componentes principais descrevem a estrutura latente de \mathbf{X} e podem ser usadas na regressão em \mathbf{y} . A regressão por quadrados mínimos parciais é similar à regressão de componentes principais, pois a matriz original de descritores \mathbf{X} também é substituída por um conjunto reduzido de combinações lineares, as variáveis latentes. No entanto, uma diferença importante reside na forma como essas combinações lineares são obtidas. Em PCR as combinações lineares são derivadas sem qualquer referência à variável dependente \mathbf{y} , enquanto que em PLS a variável dependente tem papel fundamental. Além disso, as componentes principais são ótimas no sentido de maximizar a quantidade de variância explicada em \mathbf{X} , enquanto que em PLS as variáveis latentes são ótimas no sentido de maximizar a covariância entre as variáveis independentes em \mathbf{X} com a variável dependente \mathbf{y} . O processo para a obtenção dessas variáveis latentes pode ser feito de diversas maneiras diferentes e as principais delas foram discutidas nos algoritmos explicados no capítulo 1. Se o número de variáveis latentes for igual ao número de descritores então a regressão PLS leva ao mesmo resultado que MLR. Os vetores \mathbf{t}_i são ortogonais entre si e formam a matriz de escores \mathbf{T} , os vetores \mathbf{w}_i são

ortonormais entre si e formam a matriz de “weights” \mathbf{W} enquanto que os vetores \mathbf{p}_i formam a matriz de “loadings” \mathbf{P} . Os coeficientes de regressão q_i formam o vetor de regressão \mathbf{q} e $\hat{\mathbf{y}} = \mathbf{T}\mathbf{q}$.

Do mesmo modo que ocorre em PCA, PLS também leva a uma decomposição bilinear de \mathbf{X} de acordo com a seguinte equação:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^t + \mathbf{t}_2\mathbf{p}_2^t + \dots + \mathbf{t}_A\mathbf{p}_A^t = \sum_{i=1}^A \mathbf{t}_i\mathbf{p}_i^t = \mathbf{TP}^t \quad 2.22$$

como acontece em PCR, as variáveis latentes em PLS também são combinações lineares dos descritores em \mathbf{X} , ou seja, podem ser expressas diretamente a partir de \mathbf{X} como:

$$\mathbf{t}_i = \mathbf{X}\mathbf{r}_i \quad \text{ou} \quad \mathbf{T} = \mathbf{X}\mathbf{R} \quad 2.23$$

A relação entre as matrizes \mathbf{R} e \mathbf{W} é dada por [29]

$$\mathbf{R} = \mathbf{W}(\mathbf{P}^t\mathbf{W})^{-1} \quad 2.24$$

e essas matrizes geram o mesmo espaço [57].

Assim como foi feito para o PCR, podemos expressar a regressão entre \mathbf{X} e \mathbf{y} diretamente através de um vetor de regressão \mathbf{b} . Como $\hat{\mathbf{y}} = \mathbf{T}\mathbf{q} = \mathbf{X}\mathbf{b}$ e $\mathbf{T} = \mathbf{X}\mathbf{R}$, então $\mathbf{b} = \mathbf{R}\mathbf{q}$. Logo

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^t\mathbf{W})^{-1}\mathbf{q} \quad 2.25$$

Dessa maneira pode-se relacionar diretamente \mathbf{X} e \mathbf{y} através dos coeficientes de regressão em \mathbf{b} , o que facilita a interpretação do modelo obtido, além da previsão para a atividade de novos compostos.

O algoritmo NIPALS proposto por Wold, e sua versão não ortogonal proposta por Martens [10], foram os primeiros algoritmos usados em PLS. No entanto, diversos outros algoritmos, tomando como base o algoritmo NIPALS, foram propostos posteriormente com o intuito de melhorar o tempo de execução e a compreensão do método. Os principais algoritmos usados em PLS e seus tempos de execução foram discutidos no capítulo 1.

Apesar de os métodos PCR e PLS serem parecidos e de ambos resolverem os problemas enfrentados quando se usa MLR, o método PLS é o mais usado em calibração multivariada, em especial em QSAR. Isso acontece porque, ao levar em conta a variável dependente y na derivação das variáveis latentes, o método PLS, em geral, com menos variáveis latentes chega a um resultado que exigiria mais componentes principais se o método PCR fosse utilizado [11].

2.5.2. Pré-processamento

Antes de se aplicar qualquer método matemático à tabela de dados (matriz X) que contém os descritores ou ao vetor que contém as atividades biológicas (y) é necessário aplicar um pré-processamento adequado, pois os dados em QSAR costumam ter descritores com diferentes faixas de valores, unidades de medida, variações e tamanhos. Os principais métodos de pré-processamento usados em QSAR são:

- Centrar na média
- Autoescalar

Centrar uma matriz na média consiste em calcular a média de cada coluna da matriz e, em seguida, subtrair esse valor de todos os elementos da coluna (equação 2.26):

$$x_{ij}(cm) = x_{ij} - \bar{x}_j \quad 2.26$$

onde x_{ij} é o valor do descritor j para o composto i e \bar{x}_j é a média dos valores para o descritor j . Assim, todas as colunas passam a ter \bar{x}_j média igual a zero.

Costuma-se centrar os dados na média quando os descritores são de mesma natureza ou apresentam faixas de valores semelhantes (Ex: QSAR-4D [52]).

Autoescalar os dados consiste em, além de centrar na média, dividir todos os elementos de uma coluna pelo desvio padrão dessa coluna (equação 2.27).

$$x_{ij}(a) = \frac{x_{ij} - \bar{x}_j}{s_j} \quad 2.27$$

onde s_j é o desvio padrão dos valores para o descritor j . Assim, todas as colunas passam a ter média igual a zero e desvio padrão igual a um.

Costuma-se autoescalar os dados quando os descritores são de natureza diferente ou apresentam faixas de valores bem diferentes. Em geral, o autoescalamamento é o pré-processamento utilizado em QSAR.

A Figura 2.2 ilustra o que acontece com as variáveis depois de cada pré-processamento.

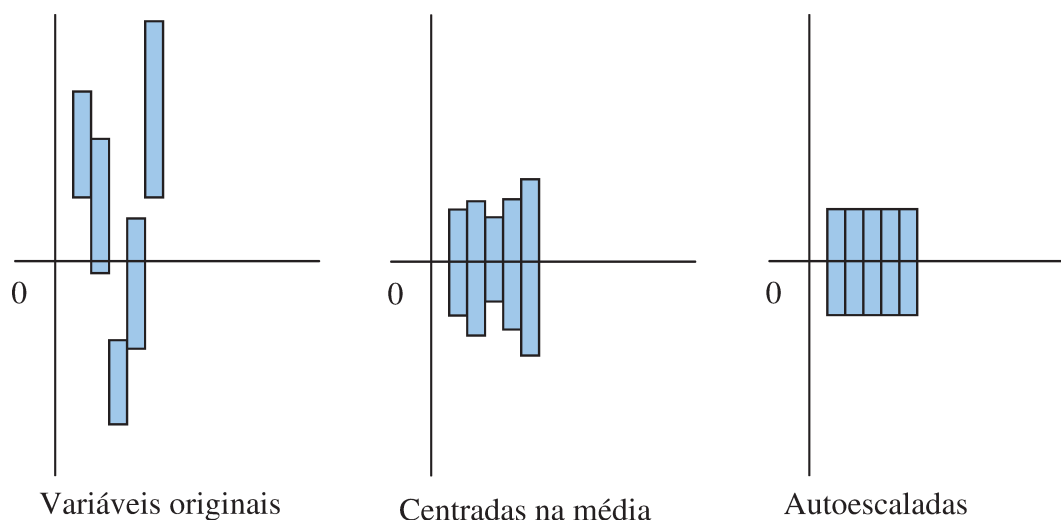


Figura 2.2. Representação das variáveis depois de cada pré-processamento.

2.5.3. Validação cruzada

Em estudos de QSAR é comum se utilizar um processo de validação interna, chamado de validação cruzada, para se determinar o número de variáveis latentes no modelo PLS. Na validação cruzada, o conjunto de treinamento é dividido em certo número de grupos e diversos modelos, com o mesmo número de variáveis latentes, são construídos sempre deixando um dos grupos de fora da análise. A variável dependente é então prevista pelo modelo construído para as amostras que foram deixadas de fora do modelo e esse processo é repetido até que todas as amostras tenham ficado de fora da análise uma vez. Esse procedimento é bastante importante para que se tenha uma idéia da capacidade preditiva e da robustez do modelo construído. Na validação cruzada pode-se utilizar da estratégia leave- N -out onde diversos números de amostras podem ser retirados durante o processo de construção de modelos. A Figura 2.3 ilustra um exemplo de execução para $N = 3$ (leave-3-out). No entanto, em QSAR costuma-se empregar a estratégia leave-one-out.

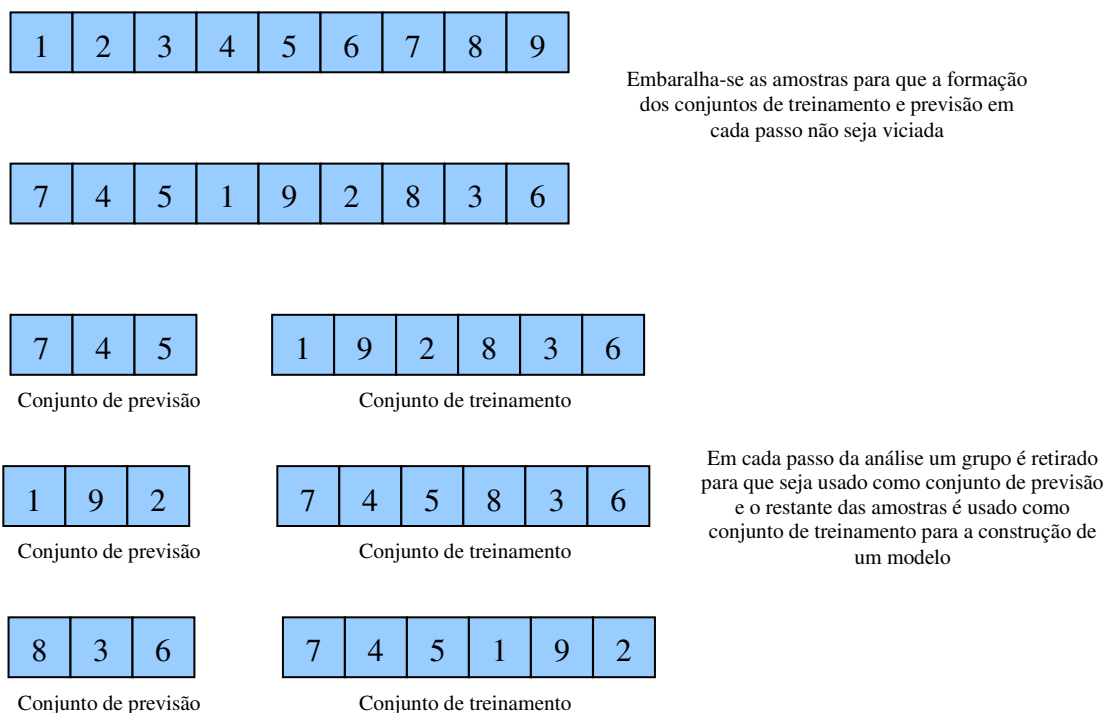


Figura 2.3. Exemplo de execução de uma validação cruzada leave-3-out

Na validação cruzada calcula-se os parâmetros estatísticos mostrados na Tabela 2.2 para avaliar a qualidade do modelo obtido. O processo de validação cruzada é repetido com a construção de modelos com diferentes números de variáveis latentes. Assim, o número de variáveis latentes ótimo para o modelo é aquele que resultou no melhor modelo de acordo com os parâmetros calculados.

Tabela 2.2. Parâmetros estatísticos que costumam ser calculados para avaliar a qualidade de um modelo durante uma validação cruzada

Parâmetro	Símbolo	Equação ^a
1 – Soma dos quadrados dos erros de predição da validação cruzada	$PRESS_{cv}$	$\sum_{i=1}^I [y(i) - \hat{y}_{cv}(i)]^2$
2 – Soma dos quadrados dos erros de predição da calibração	$PRESS_{cal}$	$\sum_{i=1}^I [y(i) - \hat{y}_{cal}(i)]^2$
3 – Coeficiente de correlação de Pearson da validação cruzada	r_{cv}	$\frac{\sum_{i=1}^I [y(i) - \bar{y}] \times [\hat{y}_{cv}(i) - \hat{\bar{y}}_{cv}]}{\sigma_y \sigma_{\hat{y}_{cv}}}$
4 – Coeficiente de correlação de Pearson da calibração	r_{cal}	$\frac{\sum_{i=1}^I [y(i) - \bar{y}] \times [\hat{y}_{cal}(i) - \hat{\bar{y}}_{cal}]}{\sigma_y \sigma_{\hat{y}_{cal}}}$
5 – Coeficiente de correlação da validação cruzada	Q^2	$1 - \frac{PRESS_{cv}}{\sum_{i=1}^I [y(i) - \bar{y}]^2}$
6 – Coeficiente de determinação múltipla	R^2	$1 - \frac{PRESS_{cal}}{\sum_{i=1}^I [y(i) - \bar{y}]^2}$
7 – Raíz quadrada do erro da validação cruzada	$RMSECV_{ou}$ SEV	$\sqrt{\frac{PRESS_{cv}}{I}}$

8 – Raíz quadrada do erro da calibração

$$\frac{RMSEC_{ou}}{SEC} = \sqrt{\frac{PRESS_{cal}}{I}}$$

9 – SPRESS

$$\frac{\sqrt{PRESS_{cv}}}{I - A - 1}$$

10 – Teste F (com intervalo de confiança de 95%), F

$$\frac{\sqrt{\frac{\sum_i (y(i) - y_{cal}(i))^2}{A}}}{\sqrt{\frac{\sum_i (y(i) - \bar{y})^2}{I - A - 1}}}$$

^a I é o número de amostras do conjunto de treinamento. A é o número de variáveis latentes no modelo. $\hat{y}_{cv}(i)$ e $\hat{y}_{cal}(i)$ são valores previstos para $y(i)$ na validação cruzada e no modelo final, respectivamente. \bar{y} , $\hat{\bar{y}}_{cv}$ e $\hat{\bar{y}}_{cal}$ são os valores médios de $y(i)$, $\hat{y}_{cv}(i)$ e de $\hat{y}_{cal}(i)$, respectivamente.

Os parâmetros mais usados em QSAR são os valores de Q^2 e R^2 . Bons modelos de QSAR devem apresentar valor de Q^2 superior a 0,5 e de R^2 superior a 0,6 [58]. No entanto, quanto mais próximos de 1 forem esses valores, melhor a qualidade do modelo obtido. Além disso, um modelo robusto não pode apresentar uma diferença entre os valores R^2 e Q^2 superior a 0,3 [59].

2.5.4. Detecção de amostras anômalas

A qualidade das amostras presentes em um conjunto de treinamento pode ser avaliada calculando-se o erro no cálculo da atividade prevista pelo modelo construído. É comum usar a influência e os resíduos de Student para detectar amostras anômalas (*outliers*) em um conjunto de treinamento [10,60,61]. A influência (*leverage*) mede a influência de uma amostra em um modelo de regressão (equação 2.28), enquanto que os resíduos de Student representam uma medida da diferença entre o valor experimental da atividade biológica e o valor predito pelo modelo (equação 2.29):

$$h_i = \mathbf{t}_i^t (\mathbf{T}_A^t \mathbf{T}_A)^{-1} \mathbf{t}_i \quad 2.28$$

$$r_i = \left[\frac{1}{I-A} \times (1 - h_i) \sum_{i=1}^m (y_i - \hat{y}_i)^2 \right]^{-1/2} (y_i - \hat{y}_i) \quad 2.29$$

Nas equações acima h_i é a influência da amostra i , I é o número de amostras, \mathbf{T}_A é a matriz de escores (obtida com PCR ou PLS) depois de extraídas A variáveis latentes, \mathbf{t}_i é a linha da matriz de escores referente à amostra i , y_i é o valor experimental da atividade biológica medida para a amostra i , \hat{y}_i é o valor predito para a atividade biológica da amostra i e r_i é o resíduo de Student para a amostra i e $I-A$ é número de graus de liberdade para o cálculo do resíduo de Student.

Costuma-se classificar como amostras anômalas as amostras com valor de influência superior a $3A/I$, onde A é o número de variáveis latentes no modelo. Já em relação aos resíduos de Student, assumindo que eles são normalmente distribuídos, um teste t pode ser aplicado para determinar se uma amostra pertence à mesma população das demais a um nível de confiança de 95%. Caso ela não pertença a esta população ela é considerada anômala. Como os resíduos de Student são medidos em unidades de desvio padrão, valores superiores a dois desvios podem ser considerados significativamente diferentes [61].

Outra maneira usada em QSAR de se avaliar se determinado composto é uma amostra anômala é através da diferença entre o valor real da atividade biológica e o valor previsto pelo modelo (resíduo). Se o resíduo de uma amostra for superior a duas vezes o desvio padrão dos resíduos da atividade biológica, provavelmente essa amostra será anômala [52,62].

A remoção de uma amostra anômala pode melhorar a qualidade estatística de um modelo. No entanto, deve-se evitar ao máximo a remoção de uma amostra anômala, pois em estudos de QSAR geralmente a quantidade de amostras é muito pequena quando comparada ao que se tem disponível em outros estudos envolvendo análise multivariada.

Caso isso seja inevitável, é importante justificar química ou biologicamente o fato de o composto ser classificado como uma amostra anômala. Por exemplo, compostos estruturalmente diferentes do restante do conjunto ou com valores suspeitos para a atividade biológica medida devem ser removidos do conjunto de treinamento antes da construção do modelo.

2.5.5. Seleção de variáveis com o algoritmo OPS

Em QSAR, normalmente o número total de variáveis disponíveis é muito maior do que o número que será efetivamente incluído nos modelos. Portanto existe a necessidade de lançar-se mão de algum tipo de procedimento de seleção para a composição dos modelos de QSAR. O processo de seleção consiste em encontrar combinações de k variáveis, dentre as J disponíveis, capazes de produzir modelos matemáticos que descrevam adequadamente os valores observados da atividade biológica. Existem diversos algoritmos de seleção de variáveis disponíveis na literatura. Dentre eles, os mais usados em QSAR são a busca sistemática e os algoritmos genéticos.

A busca sistemática consiste em combinar as J variáveis disponíveis de forma a construir e analisar todas as possíveis equações de regressão com k variáveis e, a partir daí, selecionar as melhores. Este é o único método de seleção que pode assegurar que a melhor combinação será encontrada. No entanto, com o número de descritores usados em QSAR hoje em dia este método é computacionalmente inviável. Para ter uma idéia do número de regressões necessárias para se encontrar o melhor modelo utilizando-se uma busca sistemática, para obter o modelo encontrado por Martins *et al.* [63], onde foram selecionados 12 descritores dentre 21120, seria necessário investigar $\binom{21120}{12}$, ou seja, aproximadamente $3,92 \times 10^{51}$ regressões.

Os algoritmos genéticos são baseados nos conceitos de evolução estudados em biologia. Considera-se que o subconjunto de variáveis selecionadas corresponde a um

cromossomo e o modelo construído com esse subconjunto representa um indivíduo. Diferentes indivíduos (modelos) formam uma população e aplicam-se sobre essa população processos de crossover e mutação de modo que a população evolua, ou seja, melhores indivíduos (modelos) sejam formados. O processo continua até que o melhor modelo seja encontrado. Os algoritmos genéticos são bastante usados em QSAR [62,64], mas apresentam o inconveniente de não serem reprodutíveis. Como a população inicial normalmente é escolhida de maneira aleatória, dificilmente se consegue chegar a um mesmo modelo com diferentes execuções de um mesmo algoritmo genético.

Um algoritmo de seleção de variáveis de propósito geral, chamado de OPS (do inglês, *Ordered Predictors Selection*), foi recentemente desenvolvido e vem sendo usado com sucesso em estudos de QSAR [23,65,66] assim como em dados espectrográficos e cromatográficos [7,38]. É um procedimento bastante comum em quimiometria, selecionar as melhores variáveis para a construção de um modelo analisando-se vetores que contêm alguma informação a respeito da importância de cada variável para o modelo, como o vetor de correlação e o vetor de regressão.

O algoritmo OPS traz os seguintes vetores informativos: (i) Vetor de regressão é o vetor formado pelos coeficientes de regressão obtidos quando uma regressão PLS é feita considerando todo o conjunto de variáveis. Cada coeficiente traz em si a importância de cada descritor para o modelo. (ii) Vetor de correlação é o vetor formado pelos coeficientes de correlação entre cada descritor x_j e a variável dependente y . Assim, quanto maior a correlação do descritor com a variável dependente, maior a importância desse descritor. (iii) Vetor produto é o vetor obtido a partir do produto do valor absoluto de cada elemento presente no vetor de correlação com o elemento correspondente no vetor de regressão. Assim cada elemento desse vetor traz a informação obtida nos dois vetores anteriores para a importância de um determinado descritor.

Sendo assim, este algoritmo atribui uma importância a cada descritor de acordo com um dos vetores informativos citados acima. Em seguida a matriz de descritores é rearranjada de modo que os descritores mais importantes sejam representados pelas

primeiras colunas da matriz. Finalmente, uma quantidade inicial de descritores, chamada de janela, é escolhida e diversos modelos PLS são construídos aumentando-se a quantidade de descritores de acordo com um incremento fixo pré-determinado. Dentre os modelos construídos escolhe-se aquele que apresentar melhor qualidade segundo algum dos parâmetros da Tabela 2.2. A Figura 2.4 ilustra através de um exemplo o funcionamento do algoritmo OPS. O procedimento pode ser repetido até que se encontre a matriz com os descritores que levem ao melhor modelo.

2.5.6. Validação externa

A validação externa consiste em escolher um conjunto de amostras que não fará parte da construção do modelo. Esse conjunto é chamado de conjunto teste. Assim, constrói-se um modelo com as moléculas do conjunto de treinamento e a atividade biológica das amostras do conjunto teste é calculada pelo modelo construído.

Como a atividade biológica real das amostras do conjunto teste é conhecida, pode-se fazer uma comparação entre o valor previsto pelo modelo e o valor real utilizando-se parâmetros estatísticos similares aos utilizados na validação cruzada. No entanto, o processo de validação externa é muito mais confiável para assegurar a capacidade preditiva do modelo quando comparado com a validação cruzada, pois em nenhum momento as amostras do conjunto teste participam da construção do modelo. Atualmente é exigido que se faça uma validação externa em trabalhos de QSAR [58]. A Tabela 2.3 mostra alguns parâmetros estatísticos usados para avaliar uma validação externa.

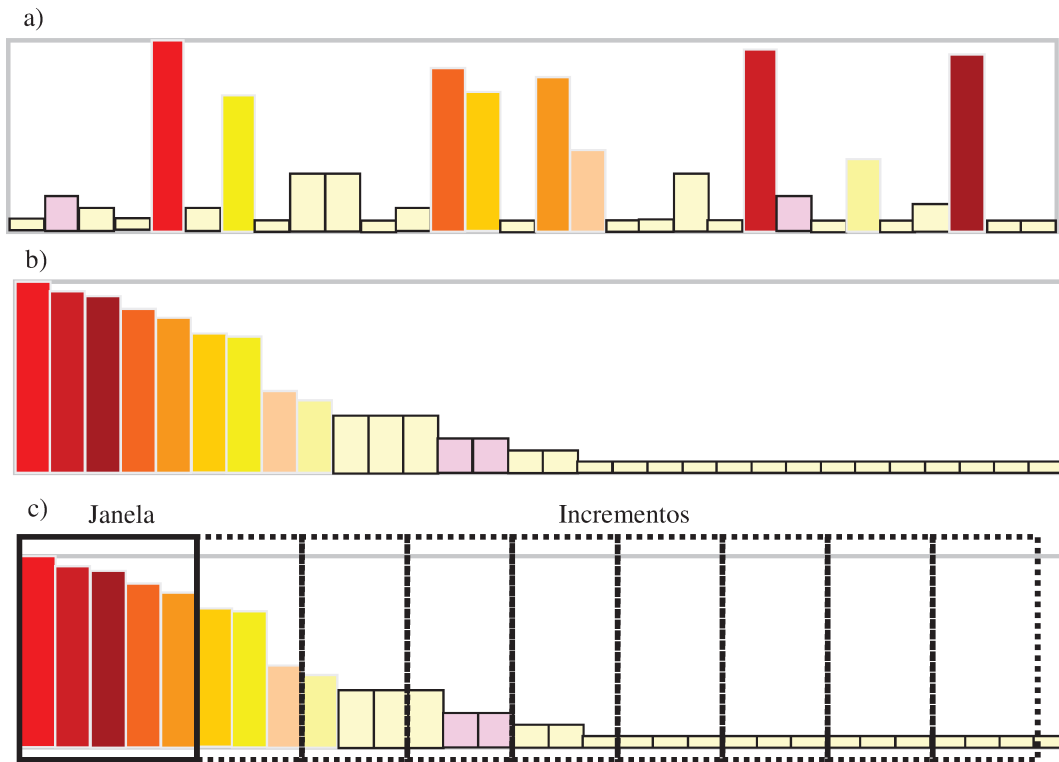


Figura 2.4. Exemplo de funcionamento do algoritmo OPS. a) Matriz original. O tamanho das barras representa a importância do descritor dada pelo vetor informativo. b) Matriz rearranjada de acordo com a importância de cada descritor. c) Construção de modelos PLS para uma janela inicial igual a 5 e incremento igual a 3.

Tabela 2.3. Parâmetros estatísticos usados na validação externa.

Parâmetro	Definição
1 – Coeficiente de determinação múltipla da predição, $R_{pred}^{2(a)}$	$1 - \frac{\sum_i (y_i - \hat{y}_{ei})^2}{\sum_i (y_i - \bar{y})^2}$
2 – Erro relativo médio da predição, ARE_{pred}	$\frac{\sum_i y_i - \hat{y}_{ei} }{y_i} 100$
3 – Erro padrão da predição, $RMSEP$ ou SEP	$\sqrt{\frac{\sum_i (y_i - \hat{y}_{ei})^2}{I_{ev}}}$

4 – Inclinação das retas de regressão linear, k e k'

$$k = \frac{\sum_i y_i \hat{y}_{ei}}{\sum_i \hat{y}_{ei}^2}; \quad k' = \frac{\sum_i y_i \hat{y}_{ei}}{\sum_i y_i^2}$$

y : atividade biológica observada; \bar{y} : média das atividades biológicas observadas para o conjunto de treinamento; \hat{y}_{ei} : atividade estimada na validação externa; I : número de amostras no conjunto de treinamento; I_{ev} : número de amostras no conjunto teste; ^(a)para R^2_{pred} , \bar{y} é a média do valor observado das atividades do conjunto de treinamento sem o conjunto teste.

2.5.7. Avaliação da robustez do modelo com o teste leave- N -out

Se o processo de validação cruzada leave- N -out for repetido várias vezes para diferentes valores de N , diferentes modelos serão construídos mesmo mantendo o número de variáveis latentes constante. Além disso, ainda que para um mesmo valor de N (desde que esse valor não seja igual a 1), diferentes execuções do procedimento leave- N -out também levarão a diferentes modelos, pois a formação dos grupos no processo de validação cruzada é feita de maneira aleatória.

A construção de diferentes modelos faz com que diferentes valores para os parâmetros estatísticos da Tabela 2.2 sejam obtidos, em especial para o valor de Q^2 . No entanto, esses valores não podem ser muito diferentes entre si (devem apresentar pouca oscilação), pois, como o modelo é construído com objetivo de prever a atividade de novas amostras, ele não pode ser muito sensível às amostras que são retiradas no processo de validação cruzada.

Assim, para avaliar se um modelo é robusto, recomenda-se fortemente que se faça um teste com repetições da validação cruzada leave- N -out. Modelos robustos não devem apresentar oscilação no valor de Q^2 superior a $\pm 0,05$ para valores de N que representem até 20% a 30% do número de amostras [59].

2.5.8. Avaliação de correlação ao acaso com o teste de aleatorização de y

O objetivo do teste de aleatorização de y, (y-randomization) é detectar e quantificar correlações ao acaso entre a variável dependente (atividade biológica) e os descritores [59,67,68]. Para obter uma estimativa da significância de um valor de Q^2 ou R^2 obtido para um dado modelo, deve-se desenvolver modelos paralelos com os valores dos descritores originais mantidos (matriz **X**) e os valores da variável dependente (vetor **y**) permutados entre as amostras (Figura 2.5).

X	y
1	12
2	6
3	2
4	10
5	5
6	1
7	7
8	4
9	8
10	3
11	9
12	11

Figura 2.5. Exemplo de aleatorização de y. Os descritores originais são mantidos enquanto que as atividades biológicas são permutadas entre as amostras.

Assim, os valores reais de Q^2 e R^2 devem ser bem maiores do que os valores obtidos para os modelos paralelos (Q^2_{yal} e R^2_{yal}). Esse procedimento é extremamente útil para assegurar que o modelo de QSAR não foi obtido ao acaso. A maneira mais simples de se avaliar se ocorre uma correlação ao acaso é observar as seguintes faixas de valores para Q^2_{yal} e R^2_{yal} de acordo com Kiralj e Ferreira [59]:

- $Q^2_{yal} < 0,2$ e $R^2_{yal} < 0,2$: não há correlação ao acaso
- Qualquer valor para Q^2_{yal} e $0,2 < R^2_{yal} < 0,3$: correlação ao acaso é desprezível
- Qualquer valor para Q^2_{yal} e $0,3 < R^2_{yal} < 0,4$: correlação ao acaso é tolerável
- Qualquer valor para Q^2_{yal} e $R^2_{yal} > 0,3$: existe correlação ao acaso

No entanto, na permutação dos valores da variável dependente para a formação de \mathbf{y}_{al} , é possível que se forme um vetor muito parecido com o vetor real \mathbf{y} . Isso poderia levar à construção de um bom modelo não por acaso, mas pelo fato de as variáveis dependentes aleatória e real (\mathbf{y}_{al} e \mathbf{y}) serem muito parecidas. Assim, outra maneira de se avaliar se houve uma correlação por chance é levando-se em conta também o coeficiente de correlação de Pearson (r) entre \mathbf{y}_{al} e \mathbf{y} . Eriksson *et al.* [67] propuseram que devem ser construídos gráficos para a regressão linear entre Q^2_{yal} e r ($Q^2_{yal} = a_{Q^2} + b_{Q^2}r$) e para a regressão linear entre R^2_{yal} e r ($R^2_{yal} = a_{R^2} + b_{R^2}r$) e que sejam observados os valores dos interceptos para essas regressões. Um modelo é considerado livre de correlação ao acaso se $a_{Q^2} < 0,05$ e $a_{R^2} < 0,3$.

Capítulo 3

Estudo QSAR multivariado da atividade antimutagênica de flavonóides contra 3-NFA em *Salmonella typhimurium* TA98

3.1. Introdução

As propriedades farmacológicas e toxicológicas dos nitroarenos têm sido objeto de vários estudos por muitos anos. Estes compostos são gerados quando os hidrocarbonetos policíclico aromáticos reagem com os óxidos de nitrogênio (NO_x) em condições encontradas no ar poluído ou quando ocorre a combustão incompleta de matéria orgânica. Como resultado, os compostos nitroaromáticos estão presentes em grande número de misturas tais como a fumaça do cigarro, cinzas flutuantes de carvão, exaustão da queima de diesel e em alimentos grelhados. Além disso, os compostos nitroaromáticos também são encontrados na indústria química, e alguns nitrofuranos e nitroimidazóis são utilizados como fármacos. Portanto, a exposição humana a um ou mais compostos nitroaromáticos pode ocorrer por uma grande variedade de vias [69,70].

O 2-Nitrofluoreno (2-NF) é geralmente o nitoareno atmosférico dominante, seguido por nitrofluorantenos e nitropirenos, como por exemplo, 3-nitrofluoranteno (3-NFA) e 1-nitropireno (1-NP) (Figura 3.1). Muitos nitroarenos mostraram ser capazes de exercer atividade mutagênica em sistema de testes bacterianos e de mamíferos. Assim, estes e outros nitroarenos podem estar envolvidos na etiologia de alguns cânceres humanos, em especial pulmão e mama [69,70]. A atividade cancerígena de compostos nitroaromáticos é normalmente iniciada por uma nitroredução enzimática. Variações consideráveis nas enzimas responsáveis pela nitroredução foram observadas em diferentes organismos. Nos seres humanos, a xantina oxidase e a NADPH-citocromica microsossomial c foram identificadas como as enzimas envolvidas neste processo. Na

Salmonella typhimurium TA98, a cepa de teste utilizada no Teste de Ames (um ensaio biológico para avaliar o potencial mutagênico de compostos químicos, desenvolvida pelo biólogo americano Bruce Ames), a nitroredução é realizada por nitrorredutases bacterianas "clássicas" [69,71]. Foi proposto então que a mutagenicidade dos compostos nitroaromáticos envolve um ciclo redox que cria espécies reativas, causando lesões do DNA ou a formação de adutos de DNA derivados de formas ativadas [69].

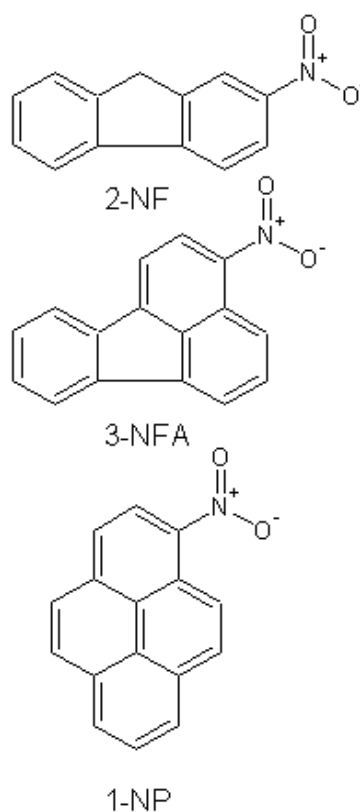


Figura 3.1. Estruturas dos nitroarenos 2-NF, 3-NFA e 1-NP.

A carcinogenicidade e mutagenicidade de alguns produtos químicos podem ser moduladas por outros produtos químicos. É bem conhecido que certos ingredientes em alimentos, e também em frutos e sementes, ou alguns compostos sintéticos, podem exercer efeitos anticarcinogênicos e antimutagênicos [70]. Estudos epidemiológicos indicaram que a ingestão de determinadas quantidades de antioxidantes, como vitaminas C, E e carotenóides, pode retardar ou prevenir o aparecimento de cânceres [72]. Esta é a

ideia central da abordagem terapêutica definida como quimioprevenção, a utilização de agentes químicos naturais ou sintéticos para reverter, suprimir ou até mesmo impedir a progressão de cânceres invasivos [73]. Os compostos com essa propriedade podem atuar por mecanismos diferentes [74], embora em alguns casos, o mecanismo específico do efeito antimutagênico de um composto (ou compostos) não é bem conhecido.

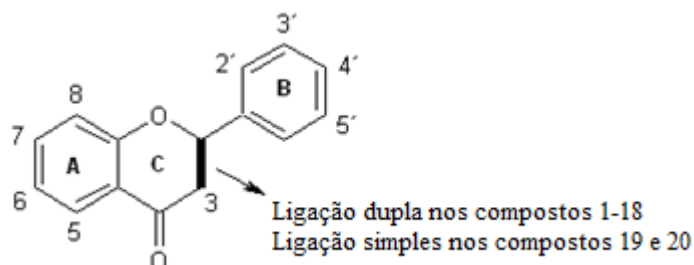
Fenóis e polifenóis estão entre os agentes quimiopreventivos potentes. Em relação a estes compostos, os flavonóides encontrados em plantas são de importância excepcional. Estas substâncias não tóxicas, encontradas em vários alimentos, demonstraram possuir propriedades protetoras, por exemplo, antioxidantes, anticarcinogênicas, antimutagênicas antialérgicas, anti-inflamatórias e atividades antivirais [70,75].

Considerando o crescente interesse quanto a anticarcinogenicidade e a antimutagenicidade de compostos fenólicos naturais e sintéticos, especialmente flavonóides, um estudo das relações quantitativas entre a estrutura e a atividade biológica (QSAR) foi realizado neste trabalho com o objetivo de obter modelos matemáticos que poderiam ajudar na compreensão e serem utilizados para a predição da atividade antimutagênica de flavonóides contra o nitrofluoranteno (3-NFA).

3.2. Farmacologia

O conjunto de treinamento selecionado foi ensaiado quanto a seu efeito antimutagênico sobre *S. typhimurium* TA98 por meio do teste de Ames. A atividade biológica, ID₅₀ (a dose de um composto em µmol/placa necessária para inibir a atividade de um agente mutagênico necessária em 50%, calculada a partir das correspondentes curvas dose-resposta), foi quantitativamente determinada em relação à 3-NFA [70]. Os valores de ID₅₀ foram convertidos em -log ID₅₀, ou pID₅₀, e estão listadas na Tabela 3.1.

Tabela 3.1. Conjunto selecionado da literatura [70] e efeitos antimutagênicos observados (no pID₅₀) na atividade mutagênica induzida pelo 3-NFA em *S. typhimurium* TA98.



Composto	Nome	3	5	6	7	8	2'	3'	4'	5'	pID ₅₀
1	5-Hydroxyflavona	H	OH	H	H	H	H	H	H	H	5,357
2	6-Hydroxyflavona	H	H	OH	H	H	H	H	H	H	6,699
3	7-Hydroxyflavona	H	H	H	OH	H	H	H	H	H	5,456
4	2'-Methoxyflavona	H	H	H	H	H	OCH ₃	H	H	H	5,046
5	Chrisina	H	OH	H	OH	H	H	H	H	H	6,000
6	Apigenina	H	OH	H	OH	H	H	H	OH	H	7,000
7	Apigenina-7-glucosídeo	H	OH	H	O-Glc ^(a)	H	H	H	OH	H	5,620
8	Luteolina	H	OH	H	OH	H	H	OH	OH	H	6,523
9	Luteolina-7-glucosídeo	H	OH	H	O-Glc ^(a)	H	H	OH	OH	H	5,092
10	Tangeretina	H	OCH ₃	OCH ₃	OCH ₃	OCH ₃	H	H	OCH ₃	H	4,967
11	Flavonol	OH	H	H	H	H	H	H	H	H	6,538
12	6-Metoxiflavonol	OH	H	OCH ₃	H	H	H	H	H	H	5,620
13	Kaempferol	OH	OH	H	OH	H	H	H	OH	H	6,538
14	Quercetina	OH	OH	H	OH	H	H	OH	OH	H	5,143
15	Isorhamnetina	OH	OH	H	OH	H	H	OCH ₃	OH	H	6,097
16	Rutina	O-Rut ^(b)	OH	H	OH	H	H	OH	OH	H	5,022
17	Morina	OH	OH	H	OH	H	OH	H	OH	H	6,155
18	Miricetina	OH	OH	H	OH	H	H	OH	OH	OH	6,222
19	Naringenina	H	OH	H	OH	H	H	H	OH	H	5,886
20	Hesperetina	H	OH	H	OH	H	H	OH	OCH ₃	H	6,097

^(a)O-Glc: O-glucose; ^(b)O-Rut: O-rutinose.

3.3. Química

Os flavonóides de interesse foram selecionados a partir de um estudo realizado por Edenharder e Tang [70] sobre a atividade antimutagênica de vários compostos em relação à mutagenicidade induzidas por 2-NF, 3-NFA e 1-NP. No teste esses três compostos foram dissolvidos em dimetil sulfóxido (DMSO) e doses adequadas foram escolhidas a partir da curva dose-resposta com cepa de *Salmonella Typhimurium* TA98. Em seguida, as curvas dose resposta foram construídas a partir de medidas de seis diferentes doses de flavonóides feitas em duplicata. Nesse estudo, 41 compostos são flavonóides, mas apenas subconjuntos de 12, 20 e 15 compostos apresentaram atividade antimutagênica determinada quantitativamente contra, respectivamente, os três mutagênicos citados.

Para este estudo, os 20 flavonóides (o maior subgrupo, contendo 10 flavonas, 8 flavonóis e 2 flavanonas), listados na Tabela 3.1, que inibiram a atividade mutagênica induzida por 3-NFA, foram selecionados. Os outros compostos (21 compostos) foram descritos como inativos (ID_{50} não fornecido) e eles não são apropriados para um estudo quantitativo. O histograma na Figura 3.2 mostra que a distribuição de atividades biológicas (pID_{50}) dos 20 compostos seguem uma distribuição razoavelmente bem normal, indicando que as atividades biológicas são bem espalhadas pelo intervalo considerado (pID_{50} 4,967-7,000). A partir dos valores de pID_{50} apresentados para estes compostos, pode ser visto que quatro deles apresentam atividades em torno de 5,000, nove deles no intervalo de 5,100 – 6,100, seis compostos têm suas atividades entre 6,100 e 6,990, e um composto tem atividade em torno de 7,000.

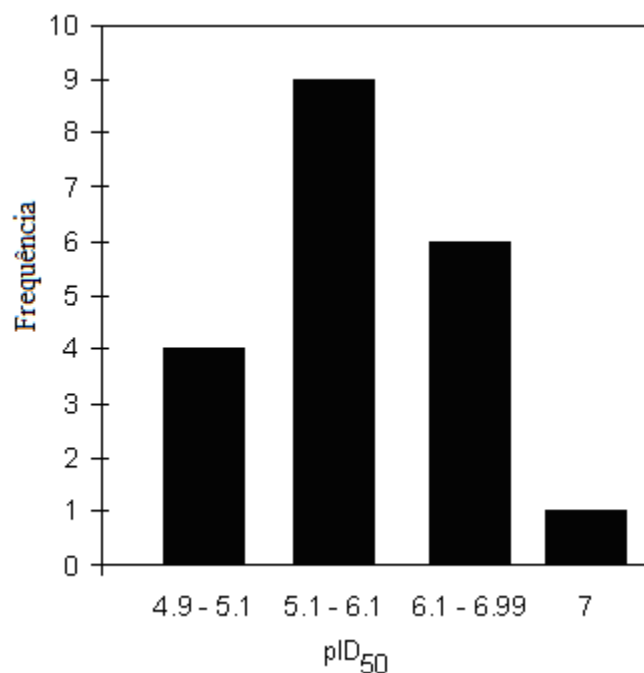


Figura 3.2. Histograma apresentando a distribuição dos compostos nas faixas de pID₅₀.

3.4. Metodologia

Estruturas tridimensionais foram construídas com base em estruturas cristalográficas similares (códigos DUMFAS, KEJBUW e WADRAV) obtidas do banco de dados *Cambridge Structural Database* [76]. As modificações necessárias dessas estruturas e as otimizações de geometria por métodos de mecânica molecular (MM+) e semi-empíricos (AM1) foram realizados utilizando o programa HyperChem 7 [77]. Através da opção *potential*, uma pesquisa conformacional no nível AM1 foi realizada para todos os compostos, com incrementos de 10 graus no ângulo diedro entre os anéis B e C. A pesquisa conformacional foi realizada entre estes anéis devido ao impedimento estérico presente em compostos com substituinte nas posições 20 (**4**) e 3 (**11 a 18**), e entre a estrutura básica dos flavonóides e do açúcar presente cadeia lateral de três compostos (**7**, **9** e **16**). As geometrias mais estáveis obtidas por este processo foram otimizadas em nível Hatree-Fock (HF/6-31G) seguido pela Teoria do Funcional de

Densidade (B3LYP/6-31G), utilizando o programa Gaussian 03 [78]. O funcional DFT/B3LYP foi escolhido porque, segundo a literatura, este método conduz a resultados muito satisfatórios para a análise de geometrias e energias [75,79]. Os descritores eletrônicos foram obtidos após a otimização final. Outros descritores (estéricos, solubilidade, topológicos) foram obtidos a partir das interfaces *Parameter Client* [80] e ALOGPS 2.1 [81], e do programa DRAGON versão 3.0 [46], levando a um total de 1221 descritores moleculares.

A fim de obter um modelo QSAR estatisticamente confiável, um procedimento de três etapas foi empregado. Na primeira, os 1221 descritores originais foram reduzidos a 840, eliminando aqueles que apresentaram o valor absoluto do coeficiente de correlação de Pearson ($|r|$) com pID_{50} inferior a 0,3.

Na segunda etapa, o algoritmo *Ordered Predictors Selection* (OPS) [23] foi usado para a seleção de variáveis. Este algoritmo constrói modelos utilizando o método PLS [10,12,20,27,28] com base em descritores autoescalados (pré-processamento recomendado para este trabalho) rearranjando as colunas da matriz de dados de tal maneira que os descritores mais importantes, classificados de acordo com um vetor informativo, são colocados nas primeiras colunas. Então, regressões PLS sucessivas são realizadas com um número crescente de descritores a fim de encontrar o melhor modelo PLS. Neste trabalho, o vetor de regressão foi usado como o vetor informativo, e o erro de predição da validação cruzada (*SPRESS* Tabela 2.2) [46] foi utilizado como um critério para classificar os modelos gerados pelo OPS.

No terceiro passo, o conjunto de nove descritores selecionados pelo algoritmo OPS (que apresentou um *SPRESS* = 0,493) foi refinado utilizando o software Pirouette 4 [82], com a remoção de amostras anômalas e cinco descritores, para obter um modelo otimizado que cumprisse os critérios estatisticamente significante, robusto e interpretativo. A regressão PLS foi escolhida como método de regressão [83].

O modelo final foi completamente validado usando um conjunto de procedimentos apresentados no capítulo 2 [59]. Alguns dos parâmetros estatísticos

listados nas Tabelas 2.2 e 2.3 foram utilizados para avaliar a qualidade do modelo. Para a qualidade interna, os limites recomendados são $R^2 > 0,6$ e $Q^2_{\text{LOO}} > 0,5$ [77,84]. Os *SEC* e *SEV* devem ser os menores possíveis. Os valores de $PRESS_{cv}$ devem ser menores do que a soma dos quadrados dos valores de resposta (SS_Y) [85]. O valor do teste F deve ser maior do que o correspondente F -crítico ($F_{A,I-A-1}$, onde I é o número de compostos e A é o número de variáveis latentes no modelo final), e quanto maior a diferença entre eles, estatisticamente mais significativo é o modelo [86].

A robustez do modelo otimizado foi examinada através da validação cruzada leave- N -out, ($N = 1, \dots, 5$). Este teste foi repetido três vezes para cada valor de " N ", sendo que foram realizadas aleatorizações de todas as linhas da matriz de dados e de seus respectivos valores de y em cada etapa do processo leave- N -out. Espera-se que o valor médio de cada Q^2_{LNO} se seja próximo ao Q^2_{LOO} (coeficiente de múltipla determinação da validação cruzada LNO) com desvios padrão próximos de zero [87]. A possibilidade de correlação ao acaso foi testada usando o teste de randomização de y [84], onde o vetor y foi aleatorizado 50 vezes [85]. A abordagem sugerida por Eriksson e colaboradores [67], com base no valor absoluto do coeficiente de correlação de Pearson entre o vetor y original e os vetores y aleatorizados, foi utilizada para quantificar a correlação ao acaso. Nesta abordagem, duas linhas de regressão são construídas usando estes coeficientes de correlação (eixo x) e os valores de R^2 e Q^2_{LOO} (eixo y). Os interceptos das equações obtidas na regressão linear devem ser inferiores a 0,3 para R^2 e 0,05 para Q^2_{LOO} .

Uma vez que o modelo foi validado internamente, o conjunto completo de dados foi dividido em conjunto de treinamento e de teste e o estudo QSAR foi efetuado integralmente. O conjunto de teste foi selecionado utilizando análise de agrupamentos por métodos hierárquicos (HCA) de tal maneira que todo o intervalo de pID_{50} e as variações estruturais fossem bem representados. O parâmetro R^2_{pred} foi utilizado como uma medida da capacidade de previsão de um modelo QSAR. Para este trabalho, foi utilizado o limite recomendado de $R^2_{\text{pred}} > 0,5$ [88,89]. No entanto, esta não é uma condição suficiente para garantir que o modelo é realmente preditivo. Recomenda-se

também verificar se: (i) as inclinações k ou k' das linhas das regressões lineares entre a atividade observada (\hat{y}_{ei}) e a atividade prevista na validação externa (y_{ei}) (Tabela 3.3), onde as inclinações devem ser $0,85 \leq k$ ou $k' \leq 1,15$; e (ii) o valor absoluto da diferença entre os coeficientes de determinação múltipla, R^2_0 e R'^2_0 , menor do que 0,3 [90,91]. Também foi considerado adequado verificar o SEP e os valores ARE_{pred} , onde os valores mínimos possíveis são desejáveis.

Tabela 3.2. Valores preditos para o conjunto teste (Tabela 3.1) e resultados dos parâmetros estatísticos.

Composto	pIC _{50obs}	pIC _{50pred}	Resíduos
4	5,046	5,517	-0,471
5	6,000	6,194	-0,194
9	5,092	5,091	0,001
13	6,538	6,411	0,127
20	6,097	5,388	0,708
R^2_{pred}			0,591
SEP			0,394
ARE_{pred}			5,230%
K			1,005
k'			0,990
$ R^2_0 - R'^2_0 $			0,109

3.5. Resultados

Quatro descritores (PJI2, R4u+, G1e e Mor27m) (Tabela 3.3) foram selecionados dentre os 1221 descritores, através da aplicação de uma pré-seleção seguida do uso do algoritmo OPS [23] e de um refinamento no programa Pirouette [82]. Uma amostra anômala foi detectada (**14**) através da análise do gráfico de *leverages versus* os resíduos de Student. A quercetina (**14**) é estruturalmente semelhante aos derivados **6** (apigenina), **13** (kaempferol) e **18** (miricetina), o que significa que estes compostos têm valores semelhantes para os descritores selecionados, o que pode ser visto no dendrograma

apresentado na Figura 3.3. No entanto, uma razoável diferença nos valores de pID_{50} são observadas entre a quercetina e os outros análogos ($pID_{50} = 5,153$ para **14**, e 7,000, 6,538 e 6,222 para **6**, **13** e **18**, respectivamente), o que pode ser causada por um erro na medida experimental. No artigo original, Edenharder e Tang [70] comentam que a quercetina (**14**) foi o único a exibir atividade mutagênica na ausência de mutagênicos (2-NF, 1-NP e 3-NFA) e a sua atividade antimutagênica contra os agentes mutagênicos teve de ser corrigida. Este fato pôde levar a um erro na atividade antimutagênica apresentada, o que é uma indicação de que o composto **14** seja realmente uma amostra anômala.

O conjunto de treinamento utilizado neste trabalho apresenta uma razoável variabilidade estrutural, mostrando substituições em quase todos os átomos de carbono dos anéis A, B e C, também incluindo açúcar entre eles. No entanto, seu tamanho ainda é pequeno quando o universo de compostos derivados de flavonóides existentes é considerado, especialmente com a remoção do composto **14**, detectado como uma amostra anômala. Assim, um rigoroso processo de validação estatística é necessário para assegurar a confiabilidade do modelo.

O melhor modelo PLS (1) foi obtido com duas variáveis latentes descrevendo 83,410% da informação original (61,150% na primeira variável latente e 22,260% na segunda). Os descritores do modelo são capazes de explicar 74,670% e de prever 59,050% da variância. O valor F , obtido a partir do teste F , foi maior do que seu valor crítico ($A= 2$ e $I-A-1= 16$) com um intervalo de confiança a 95% ($\alpha = 0,05$), e os valores de $PRESS_{val}$ foram menores do que SS_Y , o que confirma a significância estatística do modelo.

$$pID_{50} = 1,039 + 17,516(PJI2) + 0,932(Mor27m) + 3,028(G1e) + 8,218(R4u+) \quad 3.1$$

$$R^2 = 0,747; SEC = 0,332; PRESS_{cal} = 1,768; F_{(2,16)} = 23,584 (cF = 3,634);$$

$$Q^2_{LOO} = 0,590; SEV = 0,388; PRESS_{cv} = 2,858 (SS_Y = 6,979).$$

Tabela 3.3. Valores dos descritores usados para a formulação do modelo e resultados da validação cruzada LOO (exceto para a amostra anômala **14**).

Composto	PJI2^(a)	Mor27m^(b)	G1e^(c)	R4u+^(d)	pIC₅₀ obs	pIC₅₀ pred	Resíduos
1	0,800	-0,375	0,172	0,060	5,357	5,629	-0,272
2	1,000	-0,288	0,193	0,065	6,699	6,445	0,254
3	0,800	-0,339	0,172	0,067	5,456	5,812	-0,356
4	0,800	-0,431	0,168	0,061	5,046	5,589	-0,543
5	1,000	-0,384	0,171	0,060	6,000	6,190	-0,190
6	1,000	-0,382	0,169	0,077	7,000	6,370	0,630
7	0,875	-0,434	0,150	0,057	5,620	5,499	0,121
8	1,000	-0,348	0,168	0,075	6,523	6,422	0,101
9	0,875	-0,561	0,149	0,044	5,092	5,179	-0,087
10	0,857	-0,550	0,153	0,028	4,967	4,770	0,196
11	0,800	-0,332	0,172	0,079	6,538	5,753	0,785
12	0,833	-0,403	0,167	0,064	5,620	5,686	-0,066
13	1,000	-0,390	0,168	0,078	6,538	6,436	0,101
14	1,000	-0,420	0,167	0,077	5,143	-	-
15	1,000	-0,419	0,163	0,064	6,097	6,142	-0,045
16	0,889	-0,575	0,139	0,040	5,022	5,041	-0,019
17	1,000	-0,435	0,167	0,089	6,155	6,768	-0,614
18	1,000	-0,386	0,165	0,075	6,222	6,397	-0,176
19	1,000	-0,393	0,167	0,067	5,886	6,287	-0,401
20	0,833	-0,563	0,162	0,066	6,097	5,327	0,770

^(a)2D Petitjean shape index;

^(b)R maximal autocorrelation of lag 4/uniweighted;

^(c)first symmetry directional component of the Weighted Holistic Invariant Molecular;

^(d)3D-MoRSE — signal 27/ weighted by atomic masses;

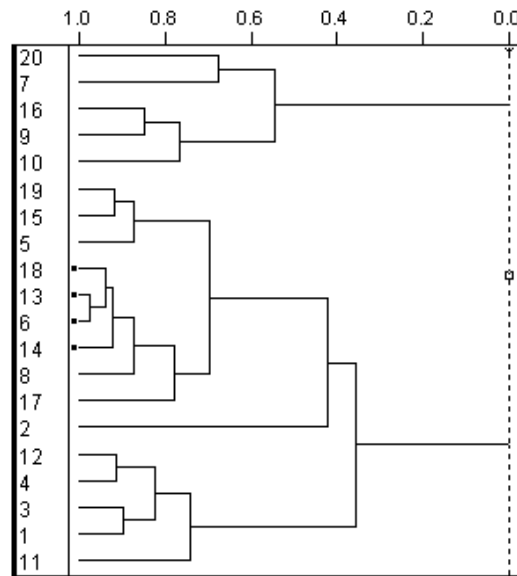


Figura 3.3. Dendrograma (dados autoescalados) do conjunto de treinamento, com os compostos **6, 14, 13 e 18** destacados.

Os resultados obtidos a partir de validação leave- N -out e da análise de aleatorização do vetor y são mostrados na Figura 3.4. O teste de aleatorização do vetor y é útil para verificar a possibilidade de que as variâncias explicadas e previstas pelo modelo obtido podem sofrer de correlação por acaso [90]. Pode-se observar que os resultados obtidos para todos os modelos aleatorizados são de má qualidade, quando comparados com o modelo de real, e os interceptos (Figuras 3.4A e 3.4B) estão dentro dos valores aceitáveis recomendados na literatura, ou seja, abaixo dos limites de 0,3 e 0,05, respectivamente [67]. Uma dispersão de pontos de dados é observada nas regiões em torno dos interceptos, o que é situação razoável para pequenos conjuntos de dados. Todos os valores obtidos para os testes R^2 e Q^2 estão abaixo de 0,4 e 0,05, respectivamente (Figura 3.4C). Estes resultados indicam que a variância explicada pelo modelo não é decorrente de correlação acaso.

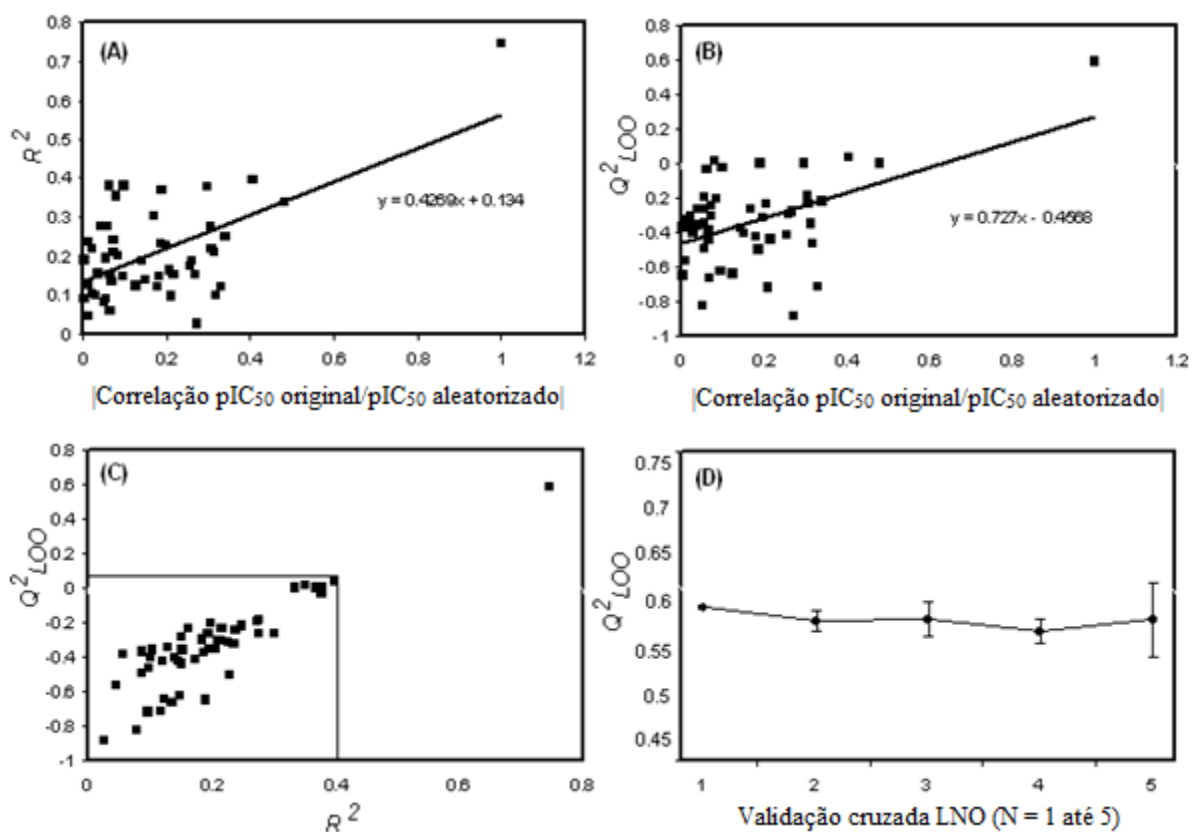


Figura 3.4. Gráficos do teste de aleatorização de y (A, B e C) e validação cruzada LNO (D). No gráfico de LNO (D), cada ponto se refere ao valor médio de um teste em triplicata e as barras se referem ao desvio padrão.

A validação cruzada LNO emprega conjuntos menores do que o procedimento LOO e pode ser repetida várias vezes, devido ao grande número de combinações quando muitos compostos são retirados do conjunto de treinamento de uma só vez. Um modelo QSAR pode ser considerado robusto quando os seus valores médios de Q^2_{LNO} são relativamente elevados e próximos do valor de Q^2_{LOO} [91]. O modelo obtido neste estudo tem Q^2_{LNO} médio relativamente alto (0,578), com pequenas variações para cada Q^2_{LNO} em relação ao Q^2_{LOO} . Os valores de desvio padrão para cada " N " são pequenos, com o máximo de $\pm 0,040$ para L5O.

Outro fator que pode ser avaliado nesse modelo é a coincidência entre os sinais de r (coeficiente de correlação de Pearson) de cada descritor com pID_{50} e os sinais de

coeficientes do modelo. De acordo com Kiralj e Ferreira [59], a incompatibilidade entre as contribuições desses dois fatores é um indício de falta de auto-consistência do modelo. Como pode ser visto na Tabela 3.4, o modelo apresenta descritores onde os sinais dos seus coeficientes coincidem com as informações fornecidas pela correlação com a atividade biológica, confirmando a auto-consistência do modelo.

Tabela 3.4. Coeficientes de correlação individual de Pearson (modelo final sem amostras anômalas) e coeficientes padronizados do modelo.

Descritor	<i>r</i>	Coefficientes Padronizados
R4u+	0,773	0,411
Mor27m	0,606	0,125
PJI2	0,617	0,427
G1e	0,597	0,150

O conjunto de dados foi dividido em um conjunto de treinamento formado por 14 compostos e um conjunto de teste formado pelos compostos **4**, **5**, **9**, **13** e **20**. Este conjunto foi utilizado para o teste de validação externa. Estes compostos cobrem bem todo o intervalo de valores de pID_{50} do conjunto completo, como pode ser visto a partir do dendrograma na Figura 3.5. O modelo construído durante a validação externa tem parâmetros estatísticos semelhantes aos encontrados para o modelo apresentado na equação 3.1 ($R^2 = 0,780$, $SEC = 0,320$, $PRESS_{cal} = 1,128$, $F_{(2,11)} = 19,467$, $Q^2_{LOO} = 0,612$, $SEV = 0,376$, e $PRESS_{cv} = 1,984$). Por conseguinte, eles podem ser considerados equivalentes.

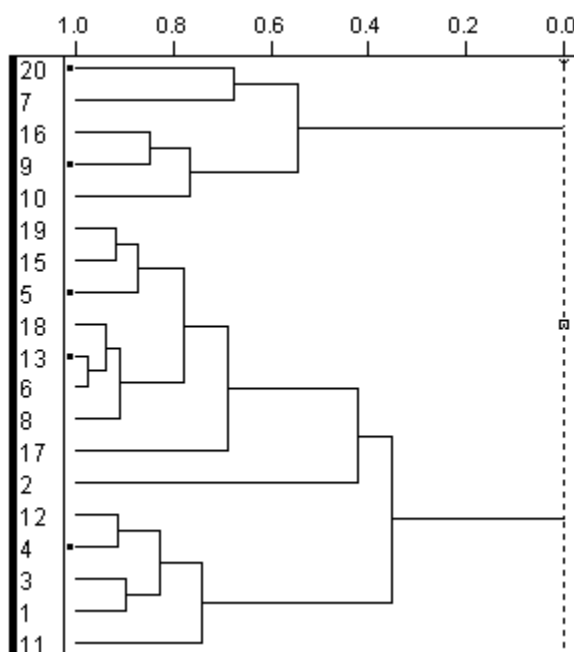


Figura 3.5. Dendrograma (dados autoescalados) do conjunto completo (sem a amostra anômala **14**), com os compostos do conjunto teste destacados.

Muitos autores argumentam que apenas modelos validados externamente (após completa a validação interna) podem ser considerados realistas e aplicáveis para o desenho de fármacos [87,91]. Alguns estudos da literatura apoiam esta conclusão [58,92]. Os resultados da validação externa (Tabela 3.3) mostram que o modelo tem alto poder de previsão externa considerando os limites propostos. Ambos os valores de k e k' e a relação $|R^2_0 - R'^2_0|$ estão dentro dos limites aceitáveis ($0,85 \leq k$ ou $k' \leq 1,15$, e $|R^2_0 - R'^2_0| < 0,3$). Os valores de SEP e de ARE_{pred} também são considerados baixos, o que é um indicativo de baixos erros de predição (desvios baixos quando comparados com o valor real) de um derivado sintetizado com base neste modelo.

Utilizando o modelo obtido, é possível sustentar a hipótese de que o composto **14** pode ser classificado como uma amostra anômala. Seu valor de pID_{50} previsto é 6,137, o que é 1,006 unidades logarítmicas acima do valor experimental do trabalho de Edenharder e Tang [70]. Além disso, os valores previstos para os compostos **6**, **13** e **18**

são próximos do composto **14**, o que concorda com a tendência de agrupamento observado no dendrograma da Figura 3.3.

Assim, os resultados da etapa de validação mostram que o modelo pode ser classificado como um bom modelo, uma vez que, de acordo com os critérios utilizados, tem boa qualidade interna, é robusto, não sofre de correlação acaso, é auto-consistente, e apresenta uma boa capacidade de previsões externas.

3.6. Interpretação do modelo

Todos os descritores selecionados foram obtidos a partir do programa Dragon 3.0 [46]. O PJI2 é um descritor topológico baseado na teoria dos grafos, enquanto os demais (R4u+, G1e e Mor27m) são descritores dependentes de geometrias tridimensionais otimizadas (neste caso, obtidas na etapa de modelagem molecular utilizando o nível de teoria B3LYP/6-31G). Quatro descritores influenciam positivamente o pID₅₀. Através dos coeficientes (0.411 para R4u+, 0.125 para Mor27m, 0.427 para PJI2 e 0.150 para G1e) obtidos no modelo PLS com dados autoescalados, é possível ver que dois deles, R4u+ e PJI2, são os mais significativos para o modelo. É interessante observar que os descritores relacionados às características estruturais, geralmente aceitos como importantes para a atividade de flavonóides (por exemplo, número de grupos OH na molécula, Log P, ou o ângulo diedro entre os anéis B e C) [92,94] não foram selecionados, mas algumas características relacionadas podem estar codificadas nos quatro descritores selecionados.

Pode ser observado que o modelo obtido neste estudo possui qualidade estatística razoável, alta capacidade de predição e robustez nos limites desejados. No entanto, em um estudo QSAR, é sempre desejável obter um modelo interpretativo capaz de relacionar as propriedades físico-químicas representadas pelos descritores moleculares selecionados com o mecanismo de ação do sistema em estudo [87]. No entanto, neste caso o mecanismo de ação não é conhecido com exatidão. De acordo com Edenharder e

Tang [70], flavonóides antimutagênicos podem modular a resposta mutagênica de nitroarenos por: (i) modificação da permeabilidade das membranas bacterianas; (ii) interações física, química ou enzimaticamente catalisadas entre flavonóides e mutagênicos extracelulares; ou (iv) efeitos dos flavonóides na fixação, expressão ou reparação do DNA danificado pelos nitroarenos.

Assim, as informações sobre o mecanismo de ação desse conjunto específico é baseada apenas nas informações codificadas nos descritores selecionados per se e em outros estudos semelhantes sobre estrutura-atividade de flavonóides [46,93-97]. Um ponto importante a ser considerado é a dificuldade na interpretação dos descritores selecionados. Em geral, a literatura relaciona descritores topológicos e geométricos com informações sobre a forma, o tamanho e ramificação [96]. Para uma melhor compreensão dos descritores selecionados e uma possível relação com mecanismo de atividade antimutagênica, informações sobre a definição de cada descritor selecionado foram consultadas na literatura e são apresentadas no texto que se segue.

Dentre os descritores selecionados, o mais importante é o índice de forma 2D Petitjean (PJI2), também designado como coeficiente de grafo teórico de forma. Este descritor de forma molecular descreve o grau de desvio de uma topologia em relação a um ciclo perfeito [98]. Os descritores de forma molecular são relacionados a vários processos físico-químicos, como fenômenos de transporte, bem como contribuições de entropia e capacidade de interação entre ligante e receptor [96]. Os valores de PJI2 variam no intervalo de 0 (uma não circunferência) a 1 (circunferência perfeita). No conjunto de treinamento, é evidente que a maioria dos compostos ativos (**6**, **2**, **13**, **8**, **18** e **17**) possuem valores de PJI2 iguais a 1, e todos eles são hidroxilados, não tendo grupos metoxi ou de açúcar. Este fato indica que os flavonóides mais compactos tendem a ter uma maior atividade antimutagênica, talvez porque pode ligar-se a um sítio de ligação pequeno ou penetrar mais facilmente através da membrana celular de *S. typhimurium* que protege o DNA bacteriano. Este descritor foi selecionado em outro estudo com flavonóides, realizado pela Rasulev e colaboradores [99], que estudou as relações

estrutura-atividade relativas à inibição da peroxidação de lipídios, e também apresentou contribuição positiva para a atividade.

O descritor R_{4u+} é denominado o *R maximal autocorrelation of lag 4/uniweighted*, um descritor R-GETAWAY. Descritores GETAWAY (*Geometry, Topology and Atom-Weights Assembly*) são baseados em uma matriz denominada "matriz de influência molecular" (MIM), proposta como uma representação molecular facilmente calculada a partir das coordenadas espaciais dos átomos de moléculas numa conformação escolhida. A magnitude da influência de uma molécula depende de seu tamanho e forma, e informações sobre as relações entre dois átomos na mesma molécula também podem ser obtidas. Esta classe de descritores tenta coincidir a geometria molecular tridimensional, fornecida pela matriz de influência molecular e pelas relações atômicas através de topologia molecular, com informações químicas utilizando diferentes ponderações atômicas (massas atômicas, polarizabilidade, volume de van der Waals, eletronegatividade e não-ponderação). Descritores GETAWAY descritores são divididos em dois conjuntos: H-GETAWAY, derivados a partir de informações fornecidas pela MIM, e R-GETAWAY, que combinam estas informações com distâncias interatômicas na molécula obtida em uma matriz de geometria [46,100,101]. Na definição de R_{4u+} , *lag* é a distância topológica, ou todas as contribuições de cada diferente ligação no grafo molecular. Termos que apresentam valores baixos, como R_1 e R_2 , representam moléculas pequenas onde se espera que as informações codificadas possuam baixa dependência das mudanças conformacionais, já que os pares de átomos estão muito próximos uns dos outros. Quanto maior o *lag*, maior é a distância entre dois átomos [98]. Como R_{4u+} não é ponderado por quaisquer propriedades químicas, este descritor provavelmente codifica apenas informações geométricas relacionadas à forma, sendo relativamente dependente das mudanças conformacionais. Semelhante ao PJI2, a tendência de altos valores também é observada para os compostos mais ativos, e menores valores para os compostos menos ativos. No entanto, o composto **14**, um dos compostos menos ativos, tem um valor elevado para este descritor, o que reforça o fato

de ele ser uma amostra anômala. Também de modo semelhante ao PJI2, a informação da forma dependente da geometria 3D pode ser relacionada com uma conformação preferida que deve ser adotada para se ligar ao seu sítio específico, ou com a facilidade de penetrar através da membrana bacteriana.

O descritor G1 é o *first symmetry directional component of the Weighted Holistic Invariant Molecular (WHIM) index weighted by atomic Sanderson electronegativities*. Os descritores WHIM são descritores tridimensionais baseados no cálculo dos eixos das componentes principais calculados a partir de uma matriz de covariância ponderada (a mesma utilizada pelos descritores GETAWAY, acrescidos dos estados eletrotopológicos atômicos), obtida a partir das coordenadas atômicas tridimensionais. Esta classe de descritores contém informações químicas sobre tamanho molecular, simetria e forma, e distribuição dos átomos constituintes [99]. Assim, G1 indica que a forma das moléculas determina fundamentalmente a distribuição eletrônica e pode ser relacionada com a importância do comportamento da eletronegatividade (comportamento em processos redox, liberação e retirada de elétrons, etc) na atividade antimutagênica. Por exemplo, no composto mais ativo (6), os dois primeiros eixos principais são paralelos em relação ao plano do esqueleto flavonóide (Figura 3.6). Em geral, a literatura descreve que os compostos têm um esqueleto flavonóidico plano, um sistema aromático com hiperconjugação e responsável por propriedades antioxidantes dos flavonóides. Este sistema de conjugação- π pode ligar radicais livres e outras espécies que danificam o DNA e outras estruturas celulares [93-95]. No entanto, estas características dos flavonóides podem ser relacionadas com a forma espacial descritores, como PJI2 e R4u+. Por exemplo, um dos compostos menos ativos menos, **16**, também é um sistema planar flavonóidico, mas com o primeiro eixo principal movido para fora do sistema devido ao açúcar em C3 (Figura 3.6).

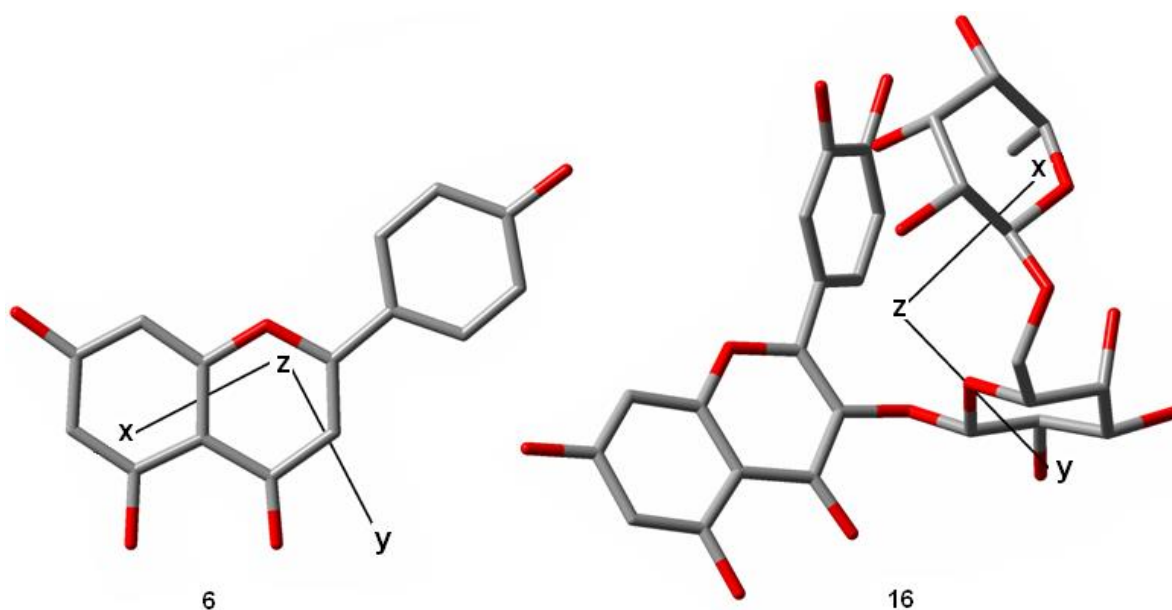


Figura 3.6. Representação das componentes principais para os compostos **6** e **16**.
 $x = 1^{\text{a}}$ componente, $y = 2^{\text{a}}$ componente, $z = 3^{\text{a}}$ componente.

Finalmente, o descritor Mor27m é uma representação tridimensional de estruturas moleculares baseadas em difração de elétrons (3D-MoRSE). Estes descritores baseiam-se na ideia da aquisição de informações a partir das coordenadas tridimensionais atômicas obtidas através da transformação usada em estudos de difração de elétrons para a preparação de curvas de dispersão teóricas [99]. A função generalizada de espalhamento, chamada de transformada molecular, pode ser calculada usando coordenadas atômicas tridimensionais. A função leva em conta o arranjo tridimensional dos átomos sem as ambiguidades como aqueles que aparecem quando o uso de grafos moleculares [97]. Neste caso, Mor27m é o "*3D-Morse signal 27/weighted by atomic mass*", calculado pela soma dos pesos atômicos visto por funções de espalhamento angular (27 \AA^{-1}) e ponderada pelas massas atômicas. Tal fato indica a importância da massa atômica, uma propriedade estérica, e dá a ideia básica que, quanto maior for a molécula, menor a atividade, porque os valores de Mor27m também diminuem quando a atividade diminui. Na verdade, os compostos menos ativos, **10** e **16**, têm valores de Mor27m -0,550 e -0,575, e os compostos mais ativos, **6** e **2**, os valores de -0,382 e -

0,288. Parece claro que os compostos pequenos penetram facilmente as bactérias através da membrana celular, contribuindo assim para o efeito antimutagênico.

Com base na discussão acima, a atividade antimutagênica do presente conjunto de treinamento de flavonóides contra 3-NFA é dependente principalmente do tamanho e forma das moléculas. Esta hipótese pode ser relacionada com características estéricas (flexibilidade e tamanho) importantes no processo de ligação. Tendo em conta que grandes moléculas são mais difíceis de difundirem-se através das membranas celulares, propriedades estéricas também podem estar relacionadas com a penetração nas bactérias. Propriedades eletrônicas, talvez relacionadas com a ligação em uma estrutura celular específica ou a capacidade de capturar mutagênicos reativos, são representadas pela eletronegatividade de Sanderson, utilizada para a ponderação do descritor WHIM.

3.7. Conclusões

Neste estudo, um modelo de QSAR multivariado para um conjunto de vinte derivados de flavonóides (10 flavonas, 8 flavonóis e 2 flavanonas) com capacidade para inibir a mutagenicidade causada por 3-NFA em *S. typhimurium TA98*, foi proposto. As estatísticas do modelo básico, seu poder de predição interna e externa, o desempenho na validação cruzada LNO e na randomização do y mostraram que o modelo é estatisticamente significativo, robusto e pode ser usado para fins de predição. A atividade inibidora destes compostos é descrita com base nos descritores PJI2, R4u+, Mor27m e G1e, indicando que a atividade antimutagênica do conjunto de treinamento é dependente principalmente do tamanho e da forma molecular, o que concorda com a literatura sobre a atividade de flavonóides. A interpretação feita para o significado dos descritores selecionados levou à sugestão de uma hipótese que eles estão relacionados com a interação dos flavonóides em um sítio de ligação e/ou com a penetração através da membrana bacteriana. Portanto, este estudo fornece conhecimento mais profundo sobre as características importantes quanto à atividade antimutagênica dos flavonóides

(neste caso, considerando-se especificamente a 3-NFA como mutagênico). Assim, pode ser útil para uma melhor compreensão da atividade desta classe de compostos e na proposta de novos agentes quimiopreventivos.

Capítulo 4

LQTA-QSAR: Uma nova metodologia de QSAR 4D

Como foi visto no capítulo 2, as metodologias de QSAR 3D e 4D vêm sendo utilizadas com bastante sucesso nos estudos de QSAR ao longo dos anos. Representadas principalmente pelos métodos CoMFA (3D) e de Hopfinger e colaboradores (4D), um enorme número de trabalhos foi desenvolvido apoiado nessas duas metodologias.

No entanto, alguns problemas podem ser observados em ambas as metodologias. Em CoMFA, o fato de a conformação bioativa não ser conhecida, transfere ao usuário a tarefa de escolher uma única conformação para cada molécula no conjunto de treinamento, geralmente a de menor energia. Em QSAR 4D, a ausência de descritores de campo dificulta a obtenção e interpretação dos modelos obtidos.

Além disso, não existem programas livres que possibilitem a geração de descritores utilizando nenhuma das duas metodologias. No caso do método CoMFA, é necessário adquirir o programa pago Sybyl. Em relação ao QSAR 4D, é necessário adquirir uma licença colaborativa do software 4D-QSAR do professor Anton J. Hopfinger.

Assim, com o objetivo de se reunir descritores de campo, como os utilizados em CoMFA, e perfis de amostragem conformacional, como os utilizados em QSAR 4D, uma nova metodologia de QSAR 4D, chamada de LQTA-QSAR, foi desenvolvida. Um software livre, chamado de LQTAgrid, foi desenvolvido para a geração de descritores de campo a partir de perfis de amostragem conformacional obtidos em simulações de dinâmica molecular feitas com o software livre GROMACS. O resultado desse estudo foi publicado na revista *Journal of Chemical Information and Modeling* e será apresentado a seguir.

4.1. Introdução

A relação quantitativa estrutura-atividade (QSAR) é um importante campo de pesquisa em química medicinal que trata da predição da atividade biológica de novos compostos utilizando relações matemáticas baseadas em propriedades estruturais, físico-químicas e conformacionais de agentes em potencial testados previamente. Os modelos de QSAR também são úteis no entendimento e explicação de mecanismos de ação de drogas em nível molecular e permite o planejamento e desenvolvimento de novos compostos com propriedades biológicas desejáveis [20].

Após Cramer e colaboradores [48] terem proposto em 1988 a Análise Comparativa de Campo Molecular (CoMFA), tais metodologias difundiram-se rapidamente na química medicinal e áreas correlatas, tornando-se um alicerce para os estudos de QSAR 3D [49,102]. No formalismo CoMFA, descritores de campo ou propriedades tri-dimensionais (eletrônico, estérico, hidrofóbico e ligações de hidrogênio) são determinados em uma rede virtual 3D. A rede equivalente a um receptor hipotético rígido e deve ser grande o suficiente para conter todas as moléculas alinhadas. As energias de interação (descritores) entre uma sonda e todos os átomos de cada molécula do conjunto investigado são calculadas em cada ponto da rede. Em tal abordagem, o método de regressão PLS [10,12,20,27,28] foi utilizado para modelar a relação entre a atividade biológica de um conjunto de compostos alinhados e seus descritores 3D calculados. A análise de QSAR 4D, originalmente proposta por Hopfinger e colaboradores em 1997 [52], incorpora a liberdade conformacional e de alinhamento para o desenvolvimento de modelos QSAR 3D criando um conjunto de estados moleculares médios, ou seja, a quarta dimensão. Nesta abordagem, os valores dos descritores de cada célula da rede cúbica são as ocupações medidas para os átomos que compõem as moléculas do conjunto investigado a partir da amostragem de conformação e do espaço de alinhamento. Uma nova abordagem introduzida no presente trabalho e chamada de LQTA-QSAR (LQTA – Laboratório de Quimiometria Teórica e Aplicada) é

baseada na geração de um perfil de amostragem conformacional (CEP do inglês, *Conformational Ensemble Profile*) para cada composto, ao invés de somente uma conformação, seguido pelo cálculo de descritores 3D para um conjunto de compostos. Esta metodologia contempla, simultaneamente, as principais características dos métodos CoMFA e QSAR 4D. O LQTA-QSAR faz uso do pacote de acesso livre GROMACS [103] para calcular as dinâmicas moleculares (MD do inglês, *Molecular Dynamics*), simulações e estimar o CEP gerado para cada composto ou ligante. As simulações de MD podem ser desenvolvidas considerando moléculas de solvente explicitamente, o que gera uma aproximação mais real do ambiente biológico. O algoritmo de seleção de preditores ordenados (OPS) [23], introduzida na sessão 2.5.5 do capítulo 2, foi aplicado como método de seleção de variáveis na construção dos modelos PLS. O método LQTA-QSAR está disponível na internet no endereço eletrônico <http://lqta.iqm.unicamp.br>.

4.2. Metodologia

Antes das análises de QSAR 4D, as simulações de MD para as moléculas em estudo são realizadas empregando o *software* GROMACS. As coordenadas das trajetórias nos arquivos de saída do GROMACS são armazenadas em arquivos no formato “gro”. Cargas e tipos de átomos para calcular as energias de van der Waals e de Coulomb são recuperadas a partir de um arquivo de topologia gromos96 (“top” ou “ipt”) [104] criado no servidor PRODRG[105]. Estes dois arquivos são utilizados como entrada para o módulo LQTAgrid, o qual gera os descritores de energia de interação 3D.

No programa LQTAgrid, o usuário pode definir as coordenadas iniciais e o tamanho da rede virtual 3D com *grid* definido, considerando as coordenadas a partir dos arquivos “gro”. É recomendado o uso de um *grid* de tamanho suficiente para conter todos os confôrmeros do conjunto investigado. Um *grid* com espaçamento de 1 Å é

selecionado para gerar milhares de pontos nas interseções de uma rede 3D regular (Figura 4.1).

Diferentes tipos de átomos, íons ou grupos funcionais, chamados de sondas (p. ex. um grupo NH_3 positivamente carregado, que corresponde a uma porção amino-terminal de peptídeos; grupos carbonilas ou carboxilas; cátions e ânions), são usados para computar os valores de energia para as interações que a sonda selecionada experimenta em uma respectiva posição na rede 3D regular. As sondas disponíveis no LQTAgrid são definidas com base na parametrização do campo de força `ff43a1` [103] para fragmentos de átomos ou moléculas e elas são apresentadas na Tabela 4.1.

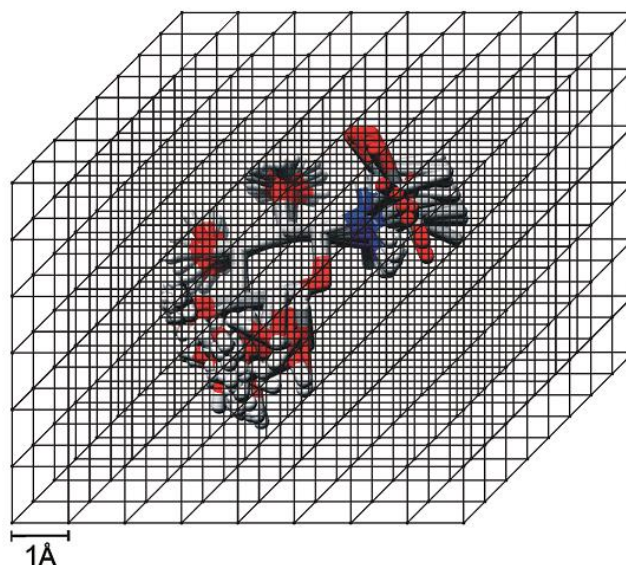


Figura 4.1. Representação da caixa 3D virtual ou grid gerada pelo módulo LQTAgrid. A distância recomendada entre as coordenadas CEP e as bordas da rede 3D são de pelo menos 5 Å. A distância do grid entre cada ponto adjacente é de 1 Å.

Tabela 4.1. Sondas disponíveis no módulo LQTAgrid.

Sondas
COO^- , C = O, NH_3^+ , SH, CH_3 , NH_2 (Arginina), C-H (Aromático), OH (H_2O), OH, Zn^{2+} , NH_2 (Amida), Cl, N-H (Aromático), Na^+

Cada sonda selecionada pelo usuário percorre o *grid*, e as propriedades eletrostáticas e estéricas 3D são computadas para cada ponto individual do *grid*, baseado nas funções potenciais de Coulomb e Lennard-Jones, de acordo com as equações 4.1 e 4.2, respectivamente.

$$E_{ele} = \frac{1}{n} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad 4.1$$

$$E_{vdW} = \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^6} \quad 4.2$$

onde

$$C_{ij}^{(12)} = \left(\frac{1}{n} C_{ii}^{(12)} \times C_{jj}^{(12)} \right)^{1/2}; C_{ij}^{(6)} = \left(\frac{1}{n} C_{ii}^{(6)} \times C_{jj}^{(6)} \right)^{1/2}$$

e q_i é a carga da i -ésima sonda; q_j é a carga do j -ésimo átomo a partir do CEP; ϵ_0 é a permissividade no vácuo; $C_{ii}^{(12)}$, $C_{ii}^{(6)}$ e $C_{jj}^{(6)}$ são parâmetros adaptados a partir do campo de força Gromos ffg43a1 [103] para sondas e átomos no CEP, respectivamente; n indica o número de *frames* alinhados no CEP; e r_{ij} representa a distância entre a i -ésima sonda e o j -ésimo átomo do CEP. Note que em ambas as equações as energias são divididas por n com o objetivo de obter uma média das energias calculadas para todas as amostras de um ligante no (CEP) em cada ponto do *grid*.

O resultado de uma análise pelo LQTAgrid é uma matriz na qual as colunas contêm os descritores que são as energias calculadas para cada ponto na rede (de acordo com as Equações 4.1 e 4.2), e as linhas representam as moléculas do conjunto investigado. Esta matriz é usada para fazer a regressão multivariada usando, por exemplo, a regressão linear múltipla (MLR), a regressão de componentes principais (PCR), ou o método PLS, com a atividade biológica como variável dependente, para construir o modelo QSAR.

4.2.1. Conjuntos de dados investigados – comparação de metodologias

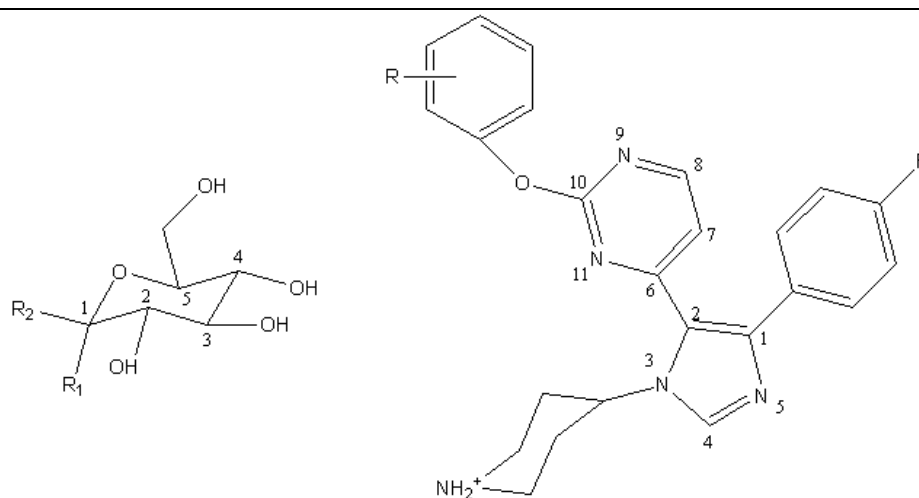
Considerando que a abordagem apresentada nesse estudo pretende incorporar as principais vantagens das metodologias de QSAR 4D e CoMFA, dois conjuntos de dados reportados na literatura, aplicando a metodologia QSAR 4D independente do receptor (RI do inglês, *Receptor Independent*) (conjunto I) [106] e o formalismo CoMFA (conjunto II) [107] foram usados para avaliar a metodologia LQTA – QSAR. O conjunto I consistiu de 47 inibidores da glicogênio fosforilase b. A enzima glicogênio fosforilase pode ajudar a controlar o balanço entre a síntese e a degradação de glicogênio favorecendo a síntese de glicogênio nos músculos e fígado e, portanto, tais inibidores podem ser agentes terapêuticos úteis para o tratamento de diabetes. Desta forma, inibidores da glicogênio fosforilase análogos à glicose podem ser de interesse clínico na regulação do metabolismo de glicogênio em organismos com diabetes. As atividades biológicas foram expressas como as energias livres de ligação aparentes entre ligante e receptor (ΔG , kcal/mol) [106] calculada a partir dos valores de constantes de ligações dos inibidores (K_i , mM) empregando a Equação 4.3, onde T é a temperatura e R é a constante dos gases.

$$\Delta G = -RT \ln K_1 \quad 4.3$$

O conjunto II é composto de 44 inibidores da quinase p38. A quinase p38 desempenha um papel vital em mecanismos de inflamações mediadas pelo fator de necrose tumoral α (TNF α) e interleucina-1 β (IL-1 β), e os inibidores de quinase p38 fornecem uma abordagem efetiva para o tratamento de doenças inflamatórias. Piridinil e pirimidinil-imidazol, que inibem a MAP quinase p38a seletivamente, são úteis no tratamento de doenças inflamatórias como artrites reumatóides. As atividades biológicas foram expressas como pIC₅₀ [107]. Sete compostos do conjunto I (3, 8, 11, 13, 20, 30 e

38) e sete compostos do conjunto II (4, 10, 13, 17, 23, 30 e 30) formaram o conjunto de validação externa e, subsequentemente, foram utilizados para testar a capacidade de predição do modelo QSAR selecionado. As estruturas e respostas biológicas dos dois conjuntos de dados estão apresentadas na Tabela 4.2.

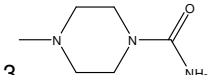
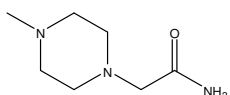
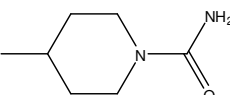
Tabela 4.2. Estruturas e atividades experimentais dos conjuntos de dados 1 [106] e 2 [107]. Os átomos numerados foram usados para o alinhamento dos CEPs de todos os ligantes.



Conjunto de dados 1

Conjunto de dados 2

	R_1	R_2	ΔG	R	pIC_{50}
1	H	NHC(=O)CH ₃	6,23	2,4 CH ₃	8,22
2	H	NHC(=O)CH ₂ CH ₃	6,11	2 HO	8,18
3	H	NHC(=O)CH ₂ Br	6,04	4 CH ₃ CH ₂	8,13
4	H	NHC(=O)CH ₂ Cl	6,03	2,5 CH ₃	7,97
5	H	NHC(=O)C ₆ H ₅	5,67	2 F	7,89
6	H	NHC(=O)CH ₂ CH ₂ CH ₃	5,58	4 HO	7,85
7	H	NHC(=O)NH ₂	5,34	2 CH ₃	7,82
8	H	C(=O)NHCH ₃	5,26	2,3 CH ₃	7,82
9	H	NHC(=O)CH ₂ NH ₂	4,76	4 CH ₃	7,72
10	C(=O)NH ₂	H	4,76	3 CH ₃ O	7,70
11	H	C(=O)NH ₂	4,65	2,4 CH ₃	7,66
12	H	C(=O)NHNH ₂	4,17	3,4 -OCH ₂ O-	7,64
13	H	SH	4,16	3 CH ₃ O	7,60
14	CH ₂ OH	H	3,92	2,6 CH ₃	7,46
15	OH	H	3,84	4 (CH ₃) ₂ CH	7,39

16	H	C(=O)NHC ₆ H ₅	3,14	2,5 CH ₃	7,39
17	H	OH	2,95	3 NH ₂ C=O	7,25
18	H	CH ₂ CN	2,84	4 C ₆ H ₅	7,22
19	OH	CH ₂ OH	2,50	3 CH ₃ NHC=O	7,12
20	H	OCH ₃	2,23	4 (CH ₃) ₃ C	7,10
21	CH ₂ NH ₂	H	2,03	4 COOH	7,09
22	C(=O)NHCH ₃	H	1,99	4 CH ₃ CH ₂ O(C=O)	7,07
23	CH ₃	H	1,77	4 NH ₂ C=O	7,05
24	C(=O)NH ₂	NHCOOCH ₃	6,65	3 F	7,02
25	H	NHCOOCH ₂ Ph	4,79	4 Cl	6,94
26	H	NHC(=O)CH ₂ NHCOCH ₃	4,17	4 CH ₃ C=O	6,92
27	H	C(=O)NHNHCH ₃	3,81	4 C ₆ H ₅ O	6,89
28	OH*	H	3,74	3 CF ₃	6,88
29	H	C(=O)NHCH ₂ CH ₂ OH	3,58	2 CH ₃ C=ONH	6,87
30	H	COOCH ₃	3,54	2 CH ₃ CH ₂ CH ₂ C=ONH	6,82
31	C(=O)NHNH ₂	H	3,50	3,4Cl	6,78
32	H	SCH ₂ C(=O)NHPh	3,39	4 CH ₃ CH ₂ CH ₂ C=O	6,76
33	H	C(=O)NH-4-OHPh	3,27	4 CN	6,75
34	H	CH ₂ CH ₂ NH ₂	3,25	3 (CH ₃) ₂ CHNHC=O	6,72
35	C(=O)NH-4-OHPh	H	3,12	3,4 F	6,68
36	OH	CH ₂ N ₃	2,95	4 CF ₃	6,67
37	OH	CH ₂ CN	2,94	4 F	6,52
38	H	C(=O)NHCH ₂ CF ₃	2,90	3 (CH ₃) ₂ NC=O	6,51
39	C(=O)NHPh	H	2,63	4 C ₆ H ₅ CH ₂ O	6,50
40	COOH	H	2,52		6,39
41	H	CH ₂ NH ₂	2,46		6,34
42	C(=O)NHCH ₂ CH ₂ OH	H	2,46		6,22
43	H	SCH ₂ C(=O)NH-2,4-F ₂ Ph	2,39	4 CH ₃ SO ₂	6,19
44	H	SCH ₂ C(=O)NH ₂	2,32	H	7,72
45	CH ₂ N ₃	H	2,29		
46	COOCH ₃	H	2,24		
47	C(=O)NHCH ₂ -2,4-F ₂ Ph	H	2,17		

Os valores de ΔG são expressos em kcal/mol.

Neste estudo, a geometria inicial utilizada para construir os modelos 3D, dados pelas geometrias otimizadas de cada ligante no espaço, foi extraída do *Brookhaven Protein Data Bank* (PDB), e os códigos de entrada foram os seguintes: 2gpb (resolução de 2,20 Å) [108] e 1bI7 (resolução de 2,50 Å) [109] para os conjuntos de dados 1 e 2, respectivamente. Embora as estruturas 3D das biomacromoléculas estivessem disponíveis, elas não foram consideradas na construção dos modelos QSAR, pois foi feito um estudo independente do receptor, conforme mencionado anteriormente. Os modelos 3D de todos os ligantes (conjunto de dados 1 e 2) passaram por minimizações de energia aplicando-se a teoria do funcional de densidade (DFT do inglês, *Density Functional Theory*) [110] com o funcional híbrido B3LYP usando o conjunto base cc-pVDZ (programa Gaussian 03) [78]. As cargas atômicas parciais eletrostáticas (CHELPG do inglês, *CHarges from ELectrostatic Potentials using a Grid based method*) [111] foram utilizadas no cálculo dos descritores das energias de interação de Coulomb pelo programa LQTAgrid. As estruturas moleculares com as energias minimizadas foram submetidas ao servidor PRODRG [106] para gerar a topologia GROMACS e as coordenadas cartesianas. Os esquemas de cargas atômicas parciais Gasteiger [112], calculadas pelo programa AutoDockTool[113] e adaptadas para o campo de força ffG43a1 foram usadas realizar as simulações de dinâmica molecular.

4.2.2. Simulações de dinâmica molecular

As simulações de dinâmica molecular de todos os ligantes isolados foram realizadas considerando um meio aquoso explícito (*extended single point charge SPC/E water models*) [114]. Quando necessário, contra-íons foram adicionados para satisfazer as condições de neutralidade elétrica. Caixas cúbicas com fronteiras periódicas foram construídas com largura suficiente, com uma distância de 10 Å entre o soluto (modelos de ligantes) e as moléculas do solvente (água). O método *Particle Mesh Ewald* (PME)

[115] foi utilizado para calcular energias de interações eletrostáticas e de van der Waals de longo alcance, com um raio de corte de 10 Å. Todas as ligações químicas foram restritas aos seus valores nominais utilizando o algoritmo de resolução de restrição linear (LINCS do inglês, *LINear Constraint Solver*) [116]. Cada componente (íon, soluto e solvente) foi separadamente acoplado no conjunto NPT (número de partículas constante, pressão e temperatura). A pressão do sistema foi controlada pelo acoplamento Parrinello-Rahman [117], e a temperatura foi gerenciada pelo termostato Berendsen [118].

As posições atômicas foram otimizadas usando os algoritmos de gradiente conjugado e declive máximo. O critério de convergência para minimização de energia foi de 50 N de força aplicada máxima sobre os átomos nos sistemas investigados onde o volume foi balanceado usando um incremento de aquecimento de sistema. O esquema de aquecimento foi o seguinte: 50, 100, 200 e 350 K para um tempo de simulação de 20 ps realizados com incrementos de 1 fs. Em seguida, o sistema foi resfriado até 300 K, e então uma simulação de dinâmica molecular de 500 ps foi realizada. O arquivo de trajetória foi registrado a cada mil passos de simulações. O CEP de todos os ligantes foram amostrados em um mesmo arquivo considerando as conformações dos ligantes registradas entre 50 a 500 ps, e estes dados foram usados para construir os modelos QSAR. Na Figura 4.2, são apresentados os CEPs para o mais e o menos ativo dos compostos de cada conjunto de treinamento. Pode ser observado que os confôrmeros do segundo conjuntos de dados não configuram uma grande faixa no espaço conformacional 3D, mesmo sem a presença da biomacromolécula. Foi verificado em testes preliminares que longas simulações não seriam necessárias para fornecer modelos PLS mais confiáveis, pois a partir do momento da estabilização do ligante na dinâmica, não se observam variações significativas nas conformações.

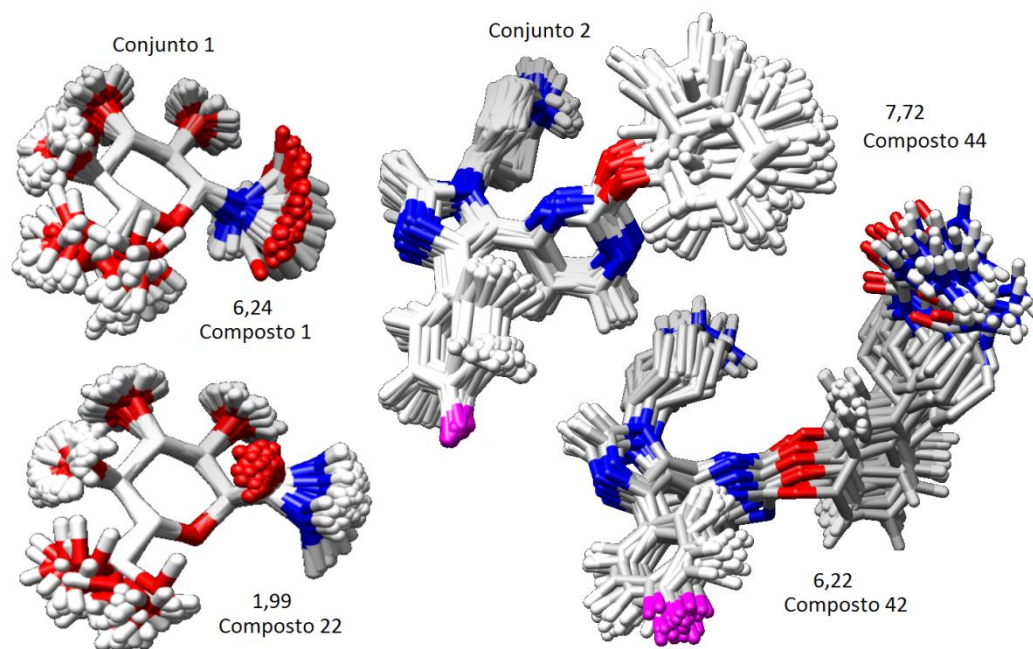


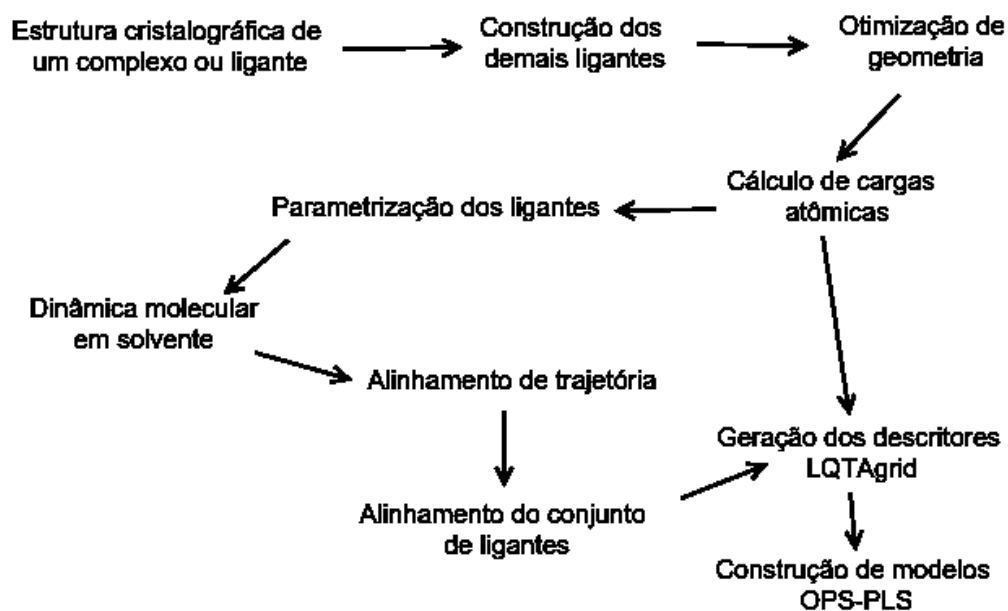
Figura 4.2. Comparação dos CEPs resultantes das simulações de DM para um dos compostos mais ativos e um dos mais inativos de cada conjunto de dados investigado. Os dados biológicos dos conjuntos 1 e 2 são expressos como ΔG (kcal/mol) e pIC_{50} , respectivamente.

4.2.3. Análises LQTAgrid

Em uma análise LQTA-QSAR são necessários dois tipos de alinhamento: alinhamento de trajetória e alinhamento do conjunto de ligantes. No primeiro, todas as amostras em um CEP de cada ligante, obtidas a partir das simulações de dinâmica molecular no GROMACS, são alinhadas entre si. No segundo, os CEPs já alinhados dos diferentes ligantes são alinhados entre si. Os átomos usados nos dois alinhamentos são os mesmos e o resultado final dos alinhamentos é colocado na caixa virtual para gerar os descritores de energia de interação molecular. Os átomos usados para os alinhamentos nessa análise para os dois conjuntos de dados utilizados são aqueles numerados na figura da Tabela 4.2 e a escolha desses átomos foi a mesma encontrada na literatura [106,107]. Como já mencionado, as sondas introduzidas no programa LQTAgrid são baseadas na

parametrização do campo de força gromos ff43a1 para simular átomos ou fragmentos moleculares como NH_3^+ , por exemplo, que corresponde a uma porção amino-terminal de peptídeos. Esse campo de força foi escolhido por estar disponível no programa GROMACS e estar parametrizado para o estudo de sistemas biomoleculares, incluindo proteínas, nucleotídeos, açúcares etc. [103,104]. As sondas exploram cada ponto de um *grid* com resolução 1Å. O tamanho do *grid* foi 24 x 22 x 20 Å para o conjunto 1, e 38 x 28 x 28 Å para o conjunto 2, e somente a sonda de NH_3^+ foi usada para calcular os descritores 3D. Testes preliminares indicaram que resultados similares poderiam ser obtidos empregando uma única sonda para gerar os descritores de energia. Uma vez obtidos os descritores de energia de interação 3D, foi feita uma seleção de variáveis com o método OPS. O Esquema 4.1 ilustra os passos envolvidos em uma análise completa pela abordagem LQTA-QSAR.

Esquema 4.1. Passos envolvidos em uma análise LQTA-QSAR completa.



4.2.4. Seleção de variáveis e validação do modelo

Matrizes de descritores geradas pelo módulo LQTAgrid (21.120 variáveis para o conjunto 1, e 59.584 para o conjunto 2) foram previamente autoescaladas para realizar os procedimentos de seleção de variáveis e de construção do modelo de regressão. Inicialmente, os valores absolutos dos coeficientes de correlação entre cada descritor e a atividade biológica foram calculados e aqueles menores do que 0,2 foram eliminados das análises. Neste ponto, permaneceram 2.449 variáveis independentes para o conjunto 1, e 19.924 para o conjunto 2. Adicionalmente, os descritores cujos gráficos em função da variável dependente mostraram distribuição não uniforme foram também eliminados [119,120]. Os conjuntos iniciais de descritores utilizados para realizar a seleção de variáveis usando o algoritmo OPS [23] foram 1.570 descritores no conjunto 1 e 8.265 descritores no conjunto 2. A ideia básica deste algoritmo é atribuir uma importância para cada descritor com base em um vetor informativo. As colunas da matriz são rearranjadas de tal modo que os mais importantes descritores são apresentados nas primeiras colunas. Então, sucessivas regressões PLS são desenvolvidas com um aumento no número de descritores visando encontrar o melhor modelo PLS. Nesta análise, o vetor de regressão foi utilizado com um vetor informativo e o coeficiente de correlação de validação-cruzada, Q^2 , como um critério para selecionar os melhores modelos.

Os modelos de regressão foram validados aplicando a validação-cruzada pela abordagem Leave- N -out (LNO) e aleatorização do vetor \mathbf{y} (\mathbf{y} -randomization) [58,59,90,68]. No procedimento LNO, N compostos ($N = 1, 2, \dots, 10$) foram deixados fora do conjunto de treinamento. Para um N particular, os dados foram aleatorizados 20 vezes, e a média e o desvio padrão dos valores de Q^2 foram utilizados. No \mathbf{y} -randomization, o vetor da variável dependente \mathbf{y} foi aleatoriamente permutado 50 vezes para os dois conjuntos investigados.

4.3. Resultados e discussão

Os modelos de regressão foram construídos após a seleção de variáveis com o algoritmo OPS, que resultou em boas avaliações estatísticas (Tabela 4.3). Para o conjunto 1, com 40 compostos no conjunto de treinamento e 12 variáveis selecionadas, o modelo com duas variáveis latentes foi indicado como o melhor modelo pela validação cruzada *leave-one-out* (LOO). Os valores de Q^2 e R^2 para este modelo são 0,72 e 0,81, respectivamente (veja Tabela 4.3). Os resíduos [atividade experimental (y_{exp}) – atividade calculada ou estimada (y_{cal})] para cada composto do conjunto 1 não excederam 1 kcal/mol nas predições de ΔG . Para o conjunto 2, com 37 compostos no conjunto de treinamento, o melhor modelo foi construído com 10 variáveis (OPS-PLS) e 5 variáveis latentes que resultou em valores de Q^2 e R^2 iguais a 0,82 e 0,90, respectivamente. Os parâmetros estatísticos dos modelos OPS-PLS resultantes, para ambos os conjuntos de dados são próximos daqueles encontrados na literatura [106,107] (Tabela 4.3).

Tabela 4.3. Parâmetros estatísticos obtidos para os modelos OPS-PLS e modelos da literatura [106,107]. Os valores entre parênteses correspondem ao número de variáveis latentes usadas nos modelos PLS.

	Q^2	R^2	RMSECV	RMSEC
Conjunto 1 (2 LVs)	0,72	0,81	0,70	0,60
Ref. [106] (MLR)	0,80	0,87	–	–
Conjunto 2 (4 LVs)	0,82	0,90	0,23	0,21
Ref. [107] (5 LVs)	0,55	0,91	0,41	0,19

Os modelos obtidos neste trabalho também foram validados aplicando os testes de aleatorização de y e de validação cruzada com a abordagem LNO, com o objetivo de avaliar sua confiabilidade e robustez. Como já foi mencionado anteriormente, bons

modelos de QSAR devem ter um valor médio de Q^2_{LNO} , $\overline{Q^2_{LNO}}$, próximo a Q^2_{LOO} e o desvio padrão para os valores de Q^2_{LOO} para cada valor de N não deve exceder 0,1. Recomenda-se que N represente uma fração significativa das amostras (por exemplo, *leave-30%-out*) em um teste LNO satisfatório.

O modelo para o conjunto de dados 1 é robusto até $N = 10$, onde o valor de $\overline{Q^2_{LNO}}$ foi 0,71 sendo próximo do modelo com $Q^2_{LOO}(0,72)$. O desvio de $\overline{Q^2_{LNO}}$ para cada valor de N oscilou entre o mínimo de 0,017 e o máximo de 0,036. Com relação ao conjunto 2 para N variando entre 1 e 9, o modelo apresentou valor de $\overline{Q^2_{LNO}}$ (0,78) próximo do valor Q^2_{LOO} (0,82) e o desvio padrão não excedeu 0,08. Contudo, para $N = 10$, o desvio de $\overline{Q^2_{LNO}}$ (0,11) excedeu o limite 0,1. Em ambos os casos, os parâmetros indicaram uma robustez satisfatória (veja Figura 4.3).

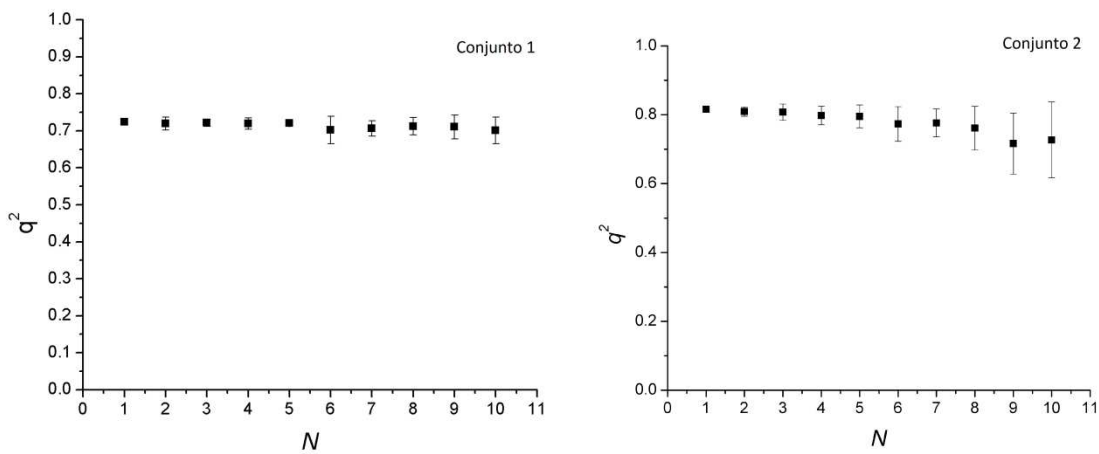


Figura 4.3. Gráficos dos resultados de LNO obtidos para os conjuntos 1 e 2, respectivamente.

Sobreajuste e correlação ao acaso entre a variável dependente e os descritores foram avaliados empregando a validação aleatorização de y . Modelos de regressão pobres, com valores de R^2 e Q^2_{LOO} baixos, são esperados quando os valores de y (variável

dependente) são embaralhados. Por outro lado, se bons modelos de regressão são obtidos no teste de aleatorização de y (relativamente altos R^2 e Q^2_{LOO}), isto implica que o modelo QSAR proposto não é aceitável, provavelmente devido a uma correlação ao acaso ou redundância estrutural do conjunto de treinamento [68,118].

Os valores de Q^2_{LOO} e R^2 , resultantes do teste de aleatorização de y para o conjunto 1 foram de $-0,32 \pm 0,20$ e $0,18 \pm 0,06$, respectivamente. Adicionalmente, os valores de Q^2_{LOO} e R^2 encontrados para o conjunto 2 foram $-0,94 \pm 0,64$ e $0,21 \pm 0,06$, respectivamente. O teste de aleatorização de y realizado implica que modelos QSAR aceitáveis foram obtidos para os conjuntos de dados estudados pelo método de modelagem apresentado. Os resultados destas validações internas são apresentados na Figura 4.4.

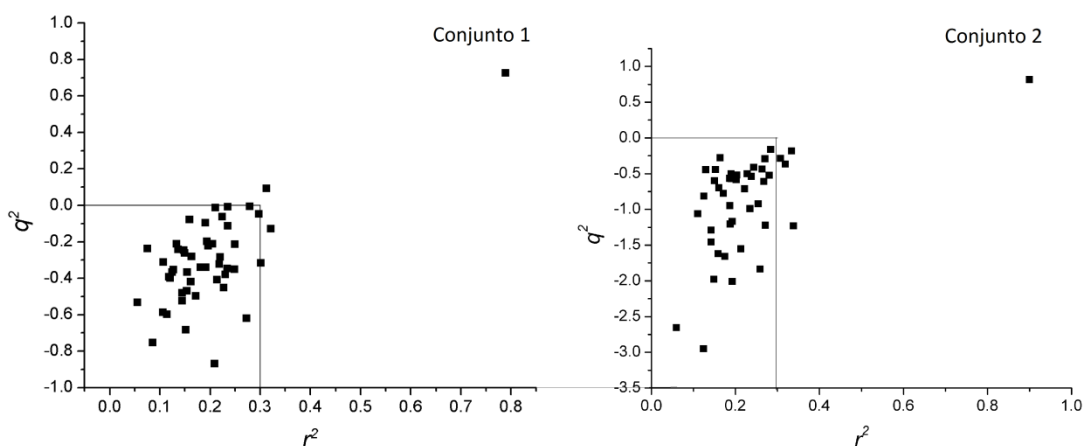


Figura 4.4. Gráficos de q^2 versus r^2 obtidos para 50 aleatorizações de y .

Infelizmente, os modelos da literatura [106,107] não foram cuidadosamente validados. Entretanto, atualmente tais procedimentos são altamente recomendados, particularmente no caso do modelo da literatura [107] para o qual a diferença entre q^2 e r^2 (0,36) é maior que 0,2, sugerindo que o modelo sofreu sobre ajuste.

Para verificar o poder de predição dos modelos OPS-PLS selecionados, dois conjuntos de testes foram usados contendo sete ligantes para cada conjunto. As estatísticas de validação externa (q^2_{ext}) encontradas para os conjuntos 1 e 2 foram de 0,60

e 0,69, respectivamente, demonstrando boa previsibilidade externa. Os valores residuais individuais [atividade experimental (y_{exp}) – atividade predita (y_{pred})] estão apresentados na Tabela 4.4, e os gráficos entre as atividades experimentais e preditas encontradas para os conjuntos de treinamento e de testes 1 e 2 estão mostrados na Figura 4.5.

Tabela 4.4. Valores de resíduos obtidos para os conjuntos teste usando os modelos OPS-PLS.

Conjunto 1	y_{exp}	y_{pred}	Resíduos (kcal/mol)	Conjunto 2	y_{exp}	y_{pred}	% Resíduos
3	6,04	4,80	1,24	4	7,97	7,50	5,9%
8	5,26	4,17	1,09	10	7,7	7,65	0,6%
11	4,65	3,93	0,72	13	7,6	7,41	2,6%
13	3,81	3,31	0,50	17	7,25	6,97	3,9%
30	3,39	3,66	-0,27	23	7,05	7,10	-0,7%
38	2,90	3,52	-0,62	30	6,82	7,17	-5,1%
20	2,32	3,31	-0,99	38	6,51	6,77	-3,9%

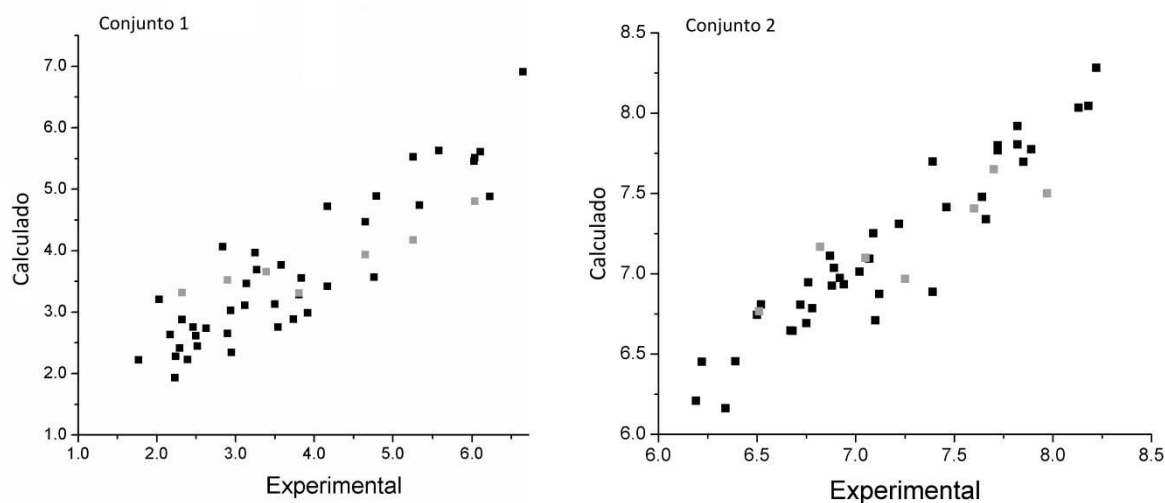


Figura 4.5. Gráfico das atividades observadas (experimentais) *versus* preditas (calculadas) para os conjuntos de treinamento (preto) e teste (cinza claro) (conjuntos 1 e 2).

4.3.1. Interpretação dos descritores

Os descritores selecionados pelo algoritmo OPS podem ser visualizados nas Figuras 4.6 e 4.7 como superfícies de acessibilidade do solvente (ViewerLite 5.0, Accelrys, Inc.; 2002).

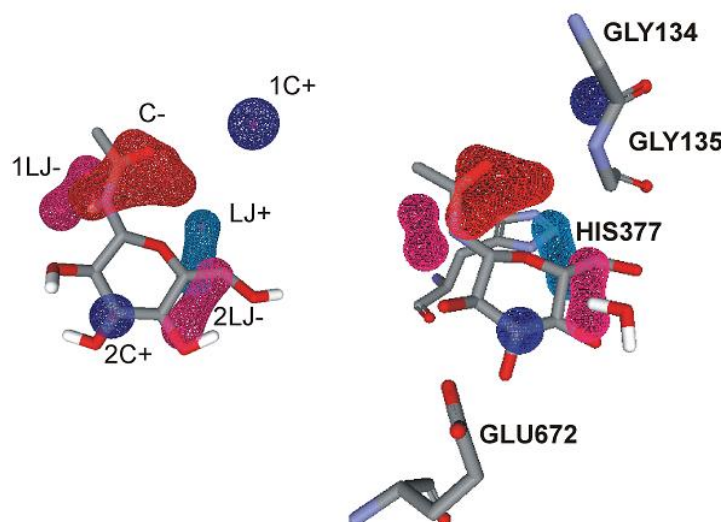


Figura 4.6. Visualização dos descritores de campo obtidos pelo método LQTAgrid e selecionados pelo algoritmo OPS para a molécula mais ativa do conjunto 1 (ViewerLite 5.0, Accelrys, Inc., 2002).

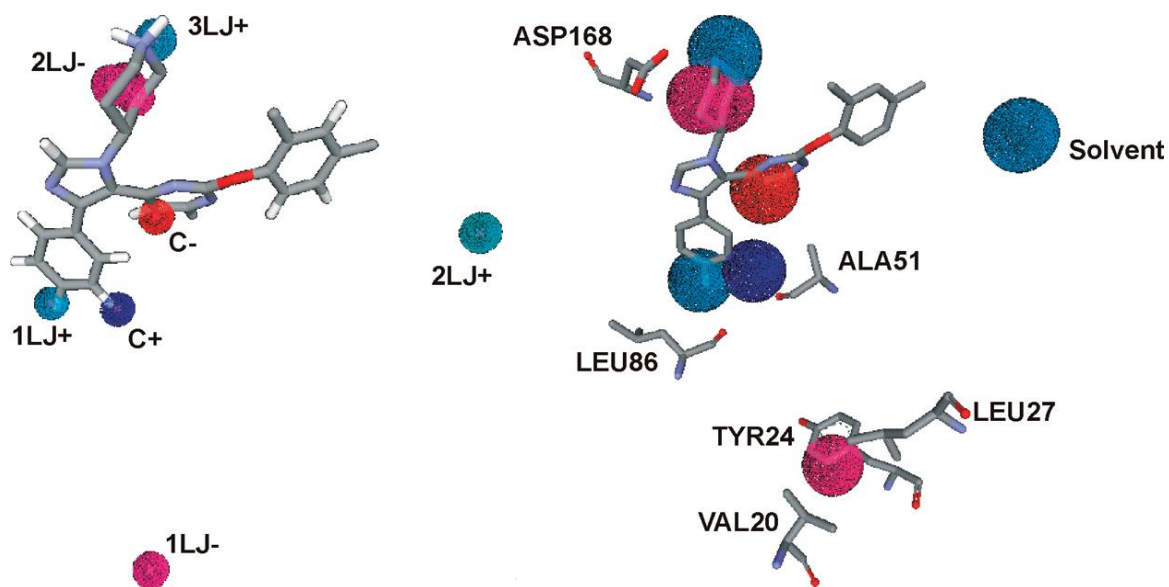


Figura 4.7. Visualização dos descritores de campo obtidos pelo método LQTAgrid e selecionados pelo algoritmo OPS para a molécula mais ativa do conjunto 2 (ViewerLite 5.0, Accelrys, Inc., 2002).

As regiões em azul claro denotam interações estéricas correspondendo aos coeficientes de regressão PLS positivos, enquanto as regiões de coloração rosa representam interações estéricas relacionadas aos coeficientes de regressão negativos. Da mesma forma, regiões de coloração azul escuro e vermelhas denotam descritores eletrostáticos com coeficientes de regressão positivo e negativo, respectivamente. Uma conformação do composto mais ativo para cada conjunto investigado e sua relação com as interações no sítio de ligação são mostradas.

Descritores apontados pela seleção de variáveis podem ser relacionados a interações encontradas no sítio de ligação, principalmente considerando as moléculas mais ativas. Para o conjunto de dados 1, a região do descritor **LJ+** pode ser associada com as interações hidrofóbicas envolvendo o resíduo aminoácido H1S377, que contribui com a estabilização do ligante no sítio de ligação pelo estabelecimento de ligações de hidrogênio. Os descritores **2LJ-** são inversamente relacionados às atividades biológicas. Eles podem ser correlacionados com a afinidade do ligante pelas moléculas de água no

sítio de ligação, sugerindo uma orientação desfavorável no sítio ativo. Outro descritor interessante é o **1LJ-**, que está relacionado ao grupo carbonílico diretamente ligado ao anel da glicose, responsável por reduzir em 10 vezes a afinidade do ligante. Considerando o descritor eletrostático **C-**, ele pode ser relacionado ao perfil do grupo carbonílico, que está longe do anel da glicose por somente um átomo. O descritor **2C+** está possivelmente relacionado aos grupos hidroxila do anel da glicose, apontando sua relevância na interação com o resíduo GLU672 no sítio de ligação. O descritor **1C+** está provavelmente relacionado à interação hidrofílica do ligante com os esqueletos dos resíduos da GLY134 e GLY135.

Os descritores selecionados para o conjunto de dados 2 podem também ser interpretados com base na interações chave que ocorrem no sítio de ligação, incluindo as interações de longo alcance. A variável **1LJ+** está posicionada em uma região hidrofóbica no sítio de ligação. Quanto mais alta a frequência dos átomos de flúor na região **1LJ+**, mais forte é a interação hidrofóbica ao redor do resíduo LEU86 (Figura 4.7). O descritor **C+** provavelmente descreve a densidade eletrônica do anel contendo o átomo de flúor como substituinte. O descritor **2LJ+** está positivamente relacionado ao valor de pIC_{50} , sugerindo que anéis mais eletropositivos interagem mais fortemente com o esqueleto do resíduo ALA51. Os descritores **2LJ-** e **3LJ+** podem ser em grande parte, relacionados à interação com resíduo ASP186 do ligante, consistente com uma região do solvente próxima à superfície da proteína. O descritor **C-** pode ser interpretado como um ponto no grid ocupado por uma região mais rica em elétrons no anel da pirimidina. Apesar do descritor **1LJ-** estar a 11 Å de distância do ligante, ele está provavelmente relacionado à região do resíduo TYR24 no sítio de ligação. Recomenda-se sempre testar outros pré-tratamentos para evitar descritores distantes do CEP, principalmente quando a estrutura do receptor não se encontra disponível. Neste trabalho, o escalamento em blocos (blockscaling) [121] foi aplicado dividindo-se a matriz de descritores em dois blocos formados pelos descritores de Coulomb e LJ, respectivamente. No entanto, os modelos obtidos não puderam ser bem validados pelas metodologias descritas

anteriormente. Desta forma, o tratamento de autoescalamento foi mantido para este conjunto de dados.

Quando o modelo da literatura para o conjunto 1 [106] é comparado ao modelo OPS-PLS, pode ser visto que os descritores são muito similares com relação às regiões C- e 1LJ-. Ambos os modelos forneceram a mesma interpretação, exceto que o modelo OPS-PLS não inclui descritores para interações de ligações de hidrogênio. Entretanto, diferenças entre os modelos aparecem em descritores encontrados na porção do anel glicosídico, os quais não foram reportados na literatura [106]. Assim, a presente abordagem apresentada neste estudo fornece descritores para uma região mais extensa do sistema sob investigação.

Os descritores selecionados para o conjunto 2 no modelo OPS-PLS final não foram bem relacionados às superfícies CoMFA reportadas por Ravindra e colaboradores [107]. Os descritores calculados pelo LQTAgrid foram bastante distintos daqueles reportados na literatura [107], dificultando qualquer tipo de comparação.

4.4. Conclusões

Um novo formalismo que aproveita a vantagem dos *frames* obtidos pela dinâmica molecular, usando o pacote computacional GROMACS, para construir modelos de energia de interação foi apresentado neste estudo. O formalismo LQTA-QSAR pode ser usado para alcançar os usuários que necessitam construir modelos de QSAR-4D, utilizando um algoritmo recente para seleção de variáveis, OPS, que tem provado ser rápido e capaz de fornecer variáveis adequadas para uma análise multivariada por PLS.

Para os exemplos propostos, os parâmetros estatísticos encontrados para o procedimento de validação cruzada pela abordagem LOO e validação externa apresentaram valores similares àqueles obtidos nas referências [106] e [107]. Contudo, os modelos LQTA-QSAR foram cuidadosamente validados aplicando os métodos de

validação cruzada interna pela abordagem LNO e aleatorização de y , os quais não foram empregados nos estudos originais. Assim, os melhores modelos OPS-PLS se mostram robustos e com boa capacidade de previsão para ambos os conjuntos investigados, utilizando ligantes isolados em um meio de solvente.

Como os CEPs são calculados utilizando o programa GROMACS, os usuários são livres para criar e alinhar o perfil dos ligantes usando confôrmeros a partir de uma representação mais realística do sistema investigado (meio de solvente explícito, complexos ligante-receptor, etc.). Neste sentido, o formalismo LQTA-QSAR é uma ferramenta promissora para estratégias de planejamento de fármacos baseados na estrutura do ligante.

É notório que a metodologia LQTA-QSAR é também um método computacional bastante amigável ao usuário, com cálculos apresentando baixo tempo computacional e as opções de cálculos podendo ser adaptadas para melhor descrever cada sistema investigado. Esta metodologia pode ser usada empregando somente *softwares* de acesso livre, que garante livre acesso para a exploração das ferramentas disponíveis, e a possibilidade de monitorar todas as etapas envolvidas na construção dos modelos QSAR-4D. Como já mencionado, ele se encontra disponível para avaliação pela comunidade acadêmica no endereço eletrônico lqta.iqm.unicamp.br.

Capítulo 5

QSAR *modeling*: um pacote computacional open source para gerar e validar modelos QSAR

A construção e validação de modelos matemáticos são etapas fundamentais em qualquer estudo de QSAR. O uso de ferramentas quimiométricas é de extrema importância para que essas etapas sejam realizadas com sucesso.

A utilização dessas ferramentas ocorre através de softwares disponíveis no mercado, sendo que a maior parte deles é desenvolvida para quimiometria em geral e uma menor parte desenvolvida especificamente para QSAR.

Quando se utiliza um software de quimiometria de propósito geral, muitos analistas de QSAR encontram dificuldades em entender o funcionamento e os resultados gerados pelo software. Além disso, como não foram projetados para tal fim, muitos processos de validação não são contemplados por esses softwares. Em relação aos softwares específicos para QSAR, dificilmente se encontra um que reúna todas as características desejadas para a construção e validação de modelos.

Além disso, a grande maioria dos softwares, tanto de quimiometria quanto de QSAR, não é livre, o que pode representar um obstáculo para a construção e validação de modelos em QSAR.

Assim, um software livre para a construção e validação de modelos QSAR, chamado de QSAR *modeling*, foi desenvolvido com o objetivo de contemplar a maior parte dos métodos usados nesta tese e em outros trabalhos desenvolvidos em nosso laboratório. Assim como o software LQTAgrid, o software QSAR *modeling* tem código aberto e espera-se que receba diversas melhorias ao longo do tempo. O resultado desse estudo será apresentado a seguir.

5.1. Introdução

Conforme pode ser visto ao longo desta tese, um modelo quantitativo QSAR (ou QSPR) é representado por meio de uma equação matemática que relaciona as propriedades dos compostos investigados com suas atividades biológicas e que possui significância estatística. Essa equação deve não somente possuir um bom poder de predição, mas deve também ser validada mostrando-se robusta e não obtida ao acaso [59,67,87,90,122,123]. A equação e as validações mencionadas acima são obtidas através de métodos matemáticos e estatísticos implementados em algum programa de computador.

Existem diversos programas disponíveis na literatura que podem ser utilizados para gerar modelos QSAR. Entre eles, alguns dos mais conhecidos são: MobyDigs [124], BuildQSAR [125], VCCLAB [126,127], QSAR+ [128], BILIN [129], MOLGEN QSPR [130], CORAL [132], CODESSA PRO [132], WOLF [133]. A Tabela 5.1 mostra uma comparação das principais características presentes no programa *QSAR Modeling* com os programas supracitados. É notório que dentre os programas livres, apenas o *QSAR modeling* incorpora todos os testes sugeridos na literatura para a validação [87] e obtenção de modelos robustos, não obtidos por correlações espúrias e com a avaliação crítica dos compostos com comportamento atípico.

Tabela 5.1. Comparativo entre as principais características do programa QSAR modeling e outros programas disponíveis na literatura.

Programa	Teste de robustez ^a	Teste de correlações ao acaso ^b	Detecção de amostras anômalas	Programa livre
MobyDigs	Não	Sim	Não	Não
BuildQSAR	Não	Não	Sim	Sim
VCCLAB	Não	Não	Não	Sim

QSAR+	Não	Sim	Sim	Não
BILIN	Não	Não	Não	Sim
MOLGEN QSPR	Não	Sim	Não	Não
CORAL	Não	Não	Não	Sim
CODESSA PRO	Não	Sim	Sim	Não
WOLF	Não	Não	Sim	Não
QSAR Modeling	Sim	Sim	Sim	Sim

^a Validação cruzada feita excluindo N amostras (*leave-N-out*).

^b Aleatorização de y

Neste trabalho, é apresentado um novo programa *open source*, denominado *QSAR modeling*, cujo objetivo é construir e validar modelos de QSAR utilizando as ferramentas quimiométricas. Esse é o primeiro programa que implementa o método de seleção de variáveis recentemente desenvolvido *ordered predictors selection* (OPS) [23], incorpora os processos de validação cruzada *leave-N-out* e aleatorização de y (*y-randomization*) além de realizar a detecção de amostras anômalas. A detecção destes compostos, frequentemente negligenciada em programas de QSAR, é implementada combinando os valores de influência (*leverage*) das amostras aos seus respectivos resíduos de Student. Este é um procedimento usual em quimiometria, mas que se mostra ausente nos programas livres citados anteriormente. O programa BuildQSAR é o único que inclui uma metodologia para a detecção de amostras anômalas, analisando o desvio padrão dos resíduos.

O processo de construção de modelos usando o programa *QSAR modeling* é descrito através de um conjunto de dados formado por 37 hidrocarbonetos poliaromáticos tendo o $\log P$ (logaritmo do coeficiente de partição octanol-água) como variável dependente [134]. Além dos descritores disponibilizados no conjunto de dados, foram utilizados descritores topológicos calculados com o programa DRAGON 6 [46].

5.2. Metodologia

O programa *QSAR modeling* foi desenvolvido na linguagem Java [135] e tem uma estrutura orientada a objetos. Ele foi projetado para ser executado em qualquer sistema operacional (Windows XP, Windows Vista, Windows 7, Linux, Mac OS, Solaris, entre outros), pois a máquina virtual java (JVM) está disponível para esses sistemas. Para executar o programa *QSAR modeling* é necessário ter o ambiente de execução java (JRE) versão 6 instalado no sistema operacional.

5.3. Resultados e discussão

As entradas para a execução do programa *QSAR modeling* são dois arquivos texto contendo, respectivamente, a matriz com os valores numéricos dos descritores (geralmente chamada de matriz \mathbf{X} com I linhas e J colunas) e o vetor contendo as atividades biológicas (designado vetor \mathbf{y} com I elementos) para os I compostos sob investigação. No arquivo contendo os descritores o usuário pode, opcionalmente, adicionar o nome de cada um deles na primeira linha. A tela principal do programa, bem como a tela de entrada de dados, disponível para o usuário está mostrada na Figura 5.1.

O programa *QSAR modeling* incorpora as seguintes ferramentas:

1. Pré-processamento dos dados
2. Seleção de variáveis – Algoritmo OPS
3. Construção do modelo de regressão – Método PLS
4. Detecção de amostras com comportamento atípico – influência e resíduos de Student
5. Validações do modelo – Validação cruzada excluindo- N -amostras e teste de aleatorização de y .

The screenshot shows the main window of the QSAR Modeling software. The window title is "QSAR Modeling". Below the title bar, there are three tabs: "Arquivo", "Pré-processamento", and "Executar". The "Pré-processamento" tab is active, displaying a data table with 25 rows (labeled "Amostra1" to "Amostra25") and 8 columns. The columns are labeled: "SpPosA_RG", "SM11_AEA...", "SpMin3_Bh...", "RDF040u", "E1m", "H8u", and "y". The "y" column is highlighted in blue. The status bar at the bottom indicates "37 Linhas X 8 Colunas".

	SpPosA_RG	SM11_AEA...	SpMin3_Bh...	RDF040u	E1m	H8u	y
Amostra1	0,423	1,618	1,295	5,008	0,266	0	3,35
Amostra2	0,424	1,618	1,53	6,324	0,275	0	3,87
Amostra3	0,423	1,618	1,425	5,457	0,267	0	4
Amostra4	0,425	1,618	1,622	9,568	0,269	0	4,39
Amostra5	0,425	1,618	1,557	6,325	0,31	0	4,38
Amostra6	0,424	1,948	1,533	5,814	0,242	0	4,31
Amostra7	0,424	1,952	1,57	6,657	0,267	0	4,44
Amostra8	0,423	1,618	1,539	7,197	0,247	0	4,31
Amostra9	0,424	1,737	1,569	7,059	0,269	0	4,31
Amostra10	0,42	2,044	1,572	7,857	0,276	0,31	4,73
Amostra11	0,419	1,802	1,417	8,02	0,33	0	4,5
Amostra12	0,421	2,362	1,672	7,875	0,308	0	5,69
Amostra13	0,423	1,806	1,526	9,015	0,301	0	4,52
Amostra14	0,422	1,927	1,544	9,753	0,303	0	5,08
Amostra15	0,424	2,028	1,598	9,053	0,279	0	5,24
Amostra16	0,424	2,08	1,623	9,594	0,306	0,033	5,15
Amostra17	0,417	2,342	1,681	11,128	0,371	0,091	5,76
Amostra18	0,42	2,26	1,681	11,999	0,353	0,081	5,91
Amostra19	0,421	2,158	1,65	11,782	0,341	0,053	5,86
Amostra20	0,42	2,517	1,876	9,035	0,253	0	5,49
Amostra21	0,419	2,247	1,683	8,537	0,257	0	5
Amostra22	0,418	2,556	1,827	15,161	0,395	0,137	6,81
Amostra23	0,421	2,473	1,804	15,211	0,383	0,24	5,8
Amostra24	0,419	2,353	1,701	12,351	0,312	0,023	5,97
Amostra25	0,418	2,496	1,745	12,891	0,325	0	6,25
							6,5

Figura 5.1. Tela principal do programa *QSAR modeling*.

5.3.1. Pré-processamento dos dados

O pré-tratamento dos dados é um procedimento de rotina na construção dos modelos de QSAR. Quando as variáveis têm diferentes unidades ou quando a faixa de variação dos dados é grande, o que ocorre com frequência nos estudos de QSAR, recomenda-se o autoescalamamento das variáveis. Com este procedimento, a influência de uma variável dominante é minimizada em cálculos posteriores. O auto-escalamamento implica em subtrair de cada elemento de uma coluna da matriz de dados o valor médio da respectiva coluna e dividir o resultado pelo desvio padrão da mesma, conforme mostrado no capítulo 2. Este tratamento é aplicado à matriz dos descritores e ao vetor contendo as atividades biológicas.

Em alguns casos, a centragem dos dados na média é utilizada ao invés do auto-escalamamento e neste caso, de cada elemento de uma coluna da matriz de dados é subtraído do valor médio da respectiva coluna. O programa *QSAR modeling* oferece

estes dois tipos de pré-processamento dos dados, se bem que no geral, os dados são autoescalados.

5.3.2. Construção de modelos de regressão com o método PLS

Os modelos matemáticos usados em QSAR são frequentemente obtidos através de uma regressão linear [10,20,122,136] entre a matriz de descritores e a atividade biológica. Geralmente essa regressão pode ser feita de três maneiras diferentes: *i*) regressão linear múltipla (MLR); *ii*) regressão por componentes principais (PCR); e *iii*) regressão por quadrados mínimos parciais (PLS).

Historicamente, a regressão multivariada era feita usando-se o método MLR, que sempre funcionou bem porque o número de descritores era menor do que o número de amostras. Atualmente, quando se utiliza esta metodologia em estudos de QSAR, é comum fixar um mínimo de 5 ou 6 compostos para cada descritor e considerar que eles não possuem alta correlação entre si ($r > 0,7$). Entretanto, programas modernos de modelagem usados em estudos de QSAR geram milhares de descritores que frequentemente são altamente correlacionados entre si, especialmente em análises de QSAR 3D e 4D [48,51,52,63]. Assim, o método MLR não pode ser usado nesses casos, ao menos que se faça uma seleção de variáveis criteriosa. Para evitar esses problemas, uma boa alternativa é o uso dos métodos de projeção, também conhecidos como métodos bi-lineares, como a regressão por componentes principais (PCR) ou a regressão por quadrados mínimos parciais (PLS) [10,12,20,27,28]. Quando esses métodos são aplicados, o número de descritores e as correlações entre eles deixam de ser um problema. Entre os métodos PLS e PCR, o primeiro é mais popular em estudos de QSAR e foi o método de regressão escolhido para ser implementado no programa *QSAR modeling*. Embora os métodos PLS e PCR apresentem resultados similares, PLS geralmente produz modelos mais parcimoniosos, com um número menor fatores e mantendo um bom ajuste.

O número ótimo de variáveis latentes (LV) no modelo é comumente determinado pela validação interna cruzada. Esta metodologia é aplicada, pois os métodos de projeção produzem modelos tendenciosos e é necessário evitar o sobreajuste (*overfitting*). Na validação cruzada, o conjunto de dados é dividido em certa quantidade de grupos (de tamanho N) e vários modelos são gerados sempre deixando um desses grupos de fora do modelo. Em seguida, o modelo de regressão obtido é usado para prever a variável dependente (atividade biológica ou propriedade físico química) das amostras deixadas de fora da análise. Esse processo é repetido até que todas as amostras tenham sido excluídas da análise uma vez. Essa estratégia, chamada de validação cruzada *leave-N-out*, é muito importante para ter uma ideia inicial a respeito da capacidade preditiva e da robustez do modelo. O uso mais comum dessa estratégia é com o valor de N igual a um, ao se fazer a validação cruzada *leave-one-out*.

O programa *QSAR modeling* oferece, como resultado da validação interna cruzada, tabelas contendo os valores dos parâmetros estatísticos 1 a 9 listados na Tabela 2.2, os coeficientes de regressão do modelo PLS ($b(j)$ para $j = 1, 2, \dots, J$), os valores previstos para a variável dependente na validação cruzada ($\hat{y}_{cv}(i)$ para $i = 1, 2, \dots, I$) e os valores previstos para a variável dependente no modelo ($\hat{y}_{cal}(i)$ para $i = 1, 2, \dots, I$)

O procedimento de validação cruzada disponível no programa *QSAR modeling* permite que o usuário escolha o número máximo de variáveis latentes (LV) e o número de amostras a serem removidas durante o processo de validação cruzada (Figura 5.2). A Figura 5.3 mostra os resultados obtidos para o conjunto de dados usado neste estudo depois de feita a seleção de variáveis.

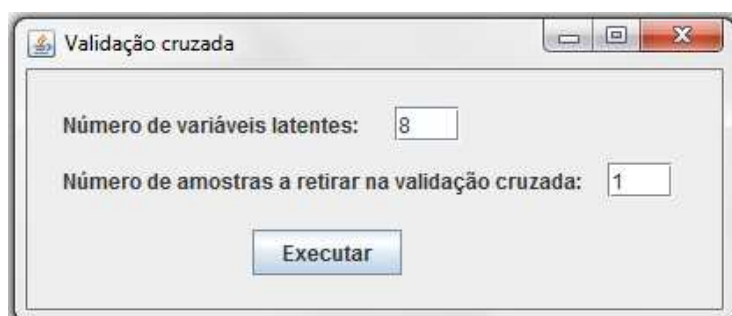


Figura 5.2. Janela do programa *QSAR modeling* na qual o usuário escolhe o número máximo de variáveis latentes e o número de amostras a serem removidas durante a validação cruzada.

5.3.3. Seleção de variáveis com o algoritmo OPS

O algoritmo de seleção de preditores ordenados (OPS) é um algoritmo desenvolvido recentemente para efetuar a seleção das variáveis [23] e já foi usado com sucesso em estudos de QSAR/QSPR [38,63,65,66,137,138]. A ideia básica desse algoritmo é atribuir importância a cada descritor com base em um vetor informativo. As colunas da matriz são rearranjadas de modo que os descritores mais importantes apareçam nas primeiras colunas. Em seguida, sucessivas regressões PLS são realizadas aumentando-se o número de descritores no modelo com o objetivo de otimizar modelo PLS. O melhor modelo de regressão pode ser escolhido de acordo com alguns dos parâmetros mostrados na Tabela 2.2.

O algoritmo OPS está implementado no programa *QSAR modeling* com os seguintes vetores informativos: *i*) vetor de correlação; *ii*) vetor de regressão PLS; e *iii*) vetor obtido pelo produto elemento a elemento destes dois vetores. A Figura 5.4 mostra a janela do programa na qual o usuário escolhe as opções apropriadas para executar o algoritmo OPS.

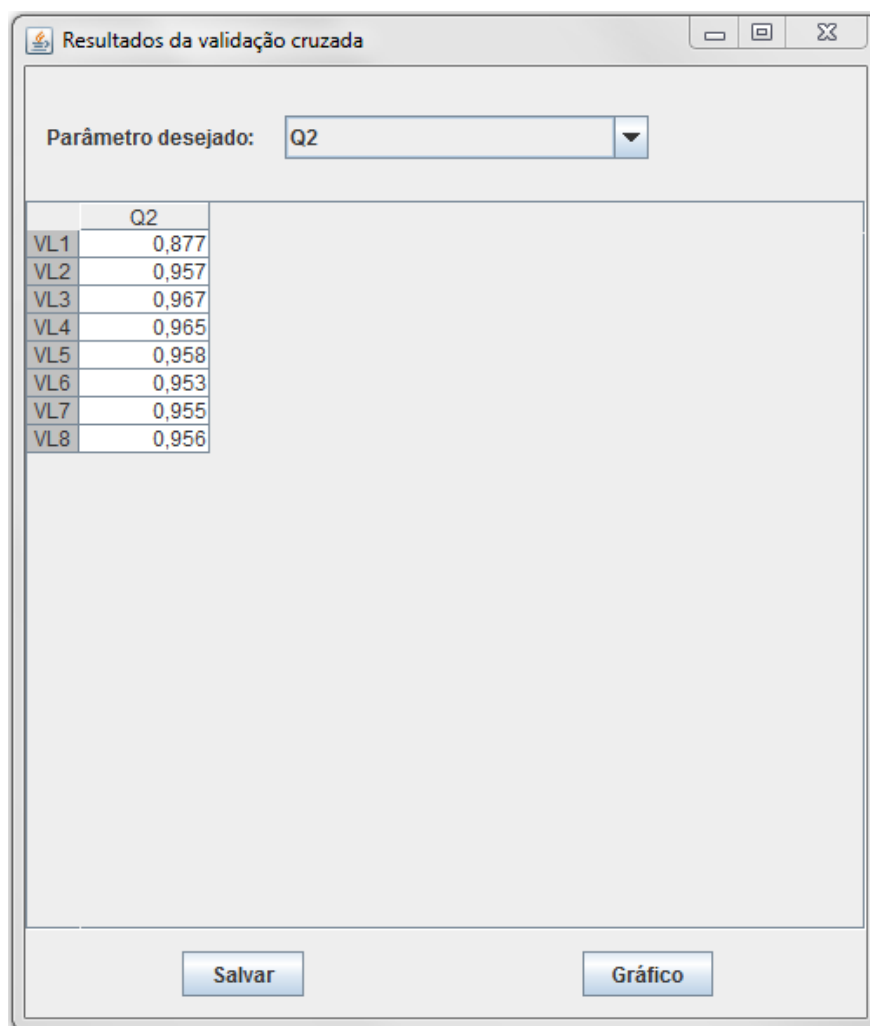


Figura 5.3. Janela do programa *QSAR modeling* na qual os resultados da validação cruzada são mostrados. Os parâmetros 1 a 9 da Tabela 2.2, os coeficientes de regressão, os valores previstos para a variável dependente na validação cruzada e os valores previstos para a variável dependente no modelo de regressão podem ser vistos nessa janela.

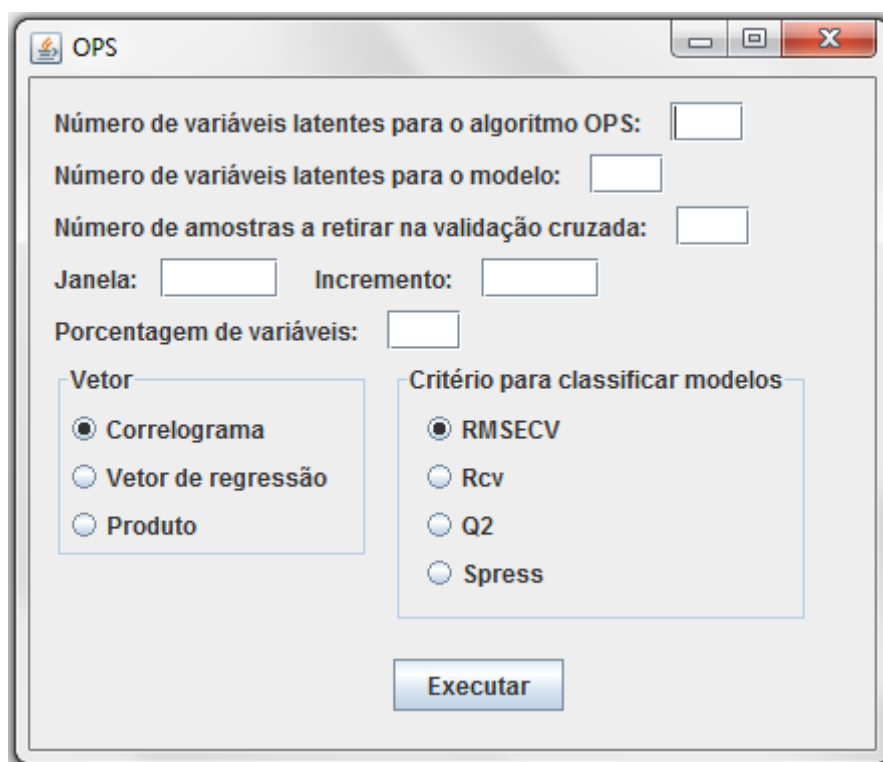


Figura 5.4. Janela do programa *QSAR modeling* na qual o usuário escolhe as opções de execução do algoritmo OPS.

O usuário tem as seguintes opções para executar o algoritmo OPS (A Figura 5.4 mostra a tela com essas opções):

- **Número de variáveis latentes para o algoritmo OPS** significa o número de variáveis latentes do modelo cujo vetor de regressão será usado para ordenar as variáveis. Diferentes números de variáveis latentes produzem vetores de regressão distintos, o que pode afetar a ordenação das variáveis.
- **Número de variáveis latentes do modelo** significa o número máximo de variáveis latentes dos modelos construídos durante a execução do algoritmo OPS (veja a referência 23 para maiores detalhes).
- **Número de amostras a serem removidas durante a validação cruzada** significa o valor de N no procedimento *leave-N-out*.

- **Janela** significa o número inicial de descritores na matriz analisada pelo algoritmo OPS.
- **Incremento** significa o número de descritores que serão adicionados à matriz analisada pelo algoritmo OPS em cada passo.
- **Porcentagem de variáveis** significa a fração de descritores que serão analisados pelo algoritmo OPS.
- **Vetor** significa o vetor informativo que será usado para ordenar os descritores.
- **Critério para classificação do modelo** significa o parâmetro que será usado para avaliar a qualidade do modelo.

Como resultado da seleção de variáveis, o programa *QSAR modeling* mostra uma tabela que lista os melhores modelos obtidos. O programa permite selecionar um dos modelos listados e executar todos os testes de validação disponíveis no programa sobre esse modelo. Além disso, o programa permite salvar a matriz de descritores selecionados para uma análise futura. A Figura 5.5 mostra a janela do programa *QSAR modeling* que apresenta os resultados obtidos com a aplicação do algoritmo OPS.

The screenshot shows a window titled "Resultados OPS" with a table of 10 models. The table has three columns: "Q2", "N° de variáveis", and "N° de variáveis latentes". Below the table, there is a text input field labeled "Escolha um modelo:" and a "Salvar" button.

	Q2	N° de variáveis	N° de variáveis latentes
Modelo 1	0,967	8	3
Modelo 2	0,957	8	2
Modelo 3	0,941	7	3
Modelo 4	0,926	7	2
Modelo 5	0,92	5	1
Modelo 6	0,907	6	2
Modelo 7	0,896	6	3
Modelo 8	0,895	4	1
Modelo 9	0,894	5	3
Modelo 10	0,894	5	2

Figura 5.5. O valor do parâmetro escolhido para avaliar o modelo, o número de variáveis selecionadas e o número de variáveis latentes dos dez modelos selecionados são mostrados como resultados do algoritmo OPS.

Para ilustrar o uso do algoritmo OPS no programa *QSAR modeling*, foi usado um conjunto de dados com 37 compostos e 407 descritores. Esse conjunto de dados, disponível em [134], é sugerido pela *International Academy of Mathematical Chemistry* como *benchmark* para avaliar a capacidade preditiva de modelos QSAR/QSPR. Os compostos são hidrocarbonetos poliaromáticos (PAH) e a propriedade de interesse utilizada foi o log P (logaritmo do coeficiente de partição octanol-água). Dentre os

descritores usados nesta análise encontram-se descritores eletrônicos, estéricos, topológicos e eletrotológicos. Como foram utilizados descritores de diferentes naturezas, o pré-processamento utilizado foi o auto-escalamento dos dados. Um corte na correlação, também disponível no programa *QSAR modeling*, foi feito antes da primeira execução do algoritmo OPS. Descritores com o valor do coeficiente de correlação de Pearson com o vetor de atividades biológicas abaixo de 0,3 foram eliminados, restando 305 descritores. A matriz de dados resultante foi submetida ao algoritmo OPS e o melhor modelo foi obtido com 15 descritores selecionados, 3 variáveis latentes e um valor de Q^2 igual a 0,959. Uma nova tentativa de se obter um melhor modelo foi feita aplicando-se o algoritmo OPS a esta matriz com 15 descritores, o que resultou no modelo final mostrado na Tabela 5.2 (8 descritores, 3 variáveis latentes e $Q^2 = 0,967$).

Tabela 5.2. Resultados da validação cruzada obtidos para um modelo com 3 LV após a seleção de variáveis feita com o programa *QSAR modeling*.

Parâmetro	$PRESS_{cv}$	$PRESS_{cal}$	r_{cv}	r_{cal}	Q^2	R^2	$RMSECV$	$RMSEC$
Valor	1,23	0,74	0,98	0,99	0,97	0,98	0,18	0,14

5.3.4. Detecção de amostras anômalas (*outliers*)

Ao verificar a qualidade do conjunto de treinamento a ser usado para a construção do modelo de regressão, deve-se assegurar de que as amostras formam um conjunto homogêneo. Compostos estruturalmente diferentes dos demais ou com valores experimentais atípicos para a atividade biológica podem ter uma influência inadequada no modelo e devem ser removidos do conjunto de treinamento antes da construção do modelo. Um procedimento comum em Quimiometria para se detectar a presença de

amostras anômalas é usar os valores de influência e dos resíduos de Student apresentados no capítulo 1 [10,61,122]. A influência indica exatamente o que o nome diz: a sua capacidade de influenciar na estimativa dos coeficientes de regressão, enquanto que o resíduo de Student é um resíduo (diferença entre o valor experimental da atividade biológica e o valor calculado pelo modelo de regressão) padronizado, obtido dividindo o resíduo por uma estimativa de seu próprio desvio padrão. A vantagem de se adotar esta definição para o resíduo é que ele apresenta média igual a zero e desvio padrão igual a um.

A detecção de amostras anômalas feita pelo programa *QSAR modeling* permite que o usuário escolha o número de variáveis latentes que será usado pelo modelo PLS e fornece como resultado uma tabela com os valores de influência e do resíduo de Student para cada um dos compostos no conjunto de treinamento (Figura 5.6). Amostras com influência maior que $3A/I$, onde A é o número de variáveis latentes e I é o número de amostras, podem ser consideradas suspeitas e devem ser analisadas cuidadosamente, caso a caso [61,122]. Em relação aos resíduos de Student, as amostras devem estar aleatoriamente espalhadas ao redor da origem indicando que eles seguem uma distribuição normal. Assumindo uma distribuição normal no nível de probabilidade 95% ($\alpha = 0,05$) o valor crítico para um teste bilateral é igual a 1,96, quando os resíduos estão limitados pelo intervalo $\pm 1,96$ (em geral se usa o intervalo 2,0).

Amostras que apresentam simultaneamente valores de influência e resíduo de Student acima dos limites indicados acima são atípicas e devem ser excluídas do conjunto de dados.

O programa *QSAR modeling* foi usado para verificar a presença de amostras anômalas no modelo final obtido depois da seleção de variáveis com o método OPS. A Figura 5.7 mostra o gráfico de influência *versus* resíduos de Student. A partir dessa Figura pode ser observado que não existem amostras que apresentem simultaneamente influência e resíduo de Student acima dos limites aceitos pela literatura. No entanto, a partir da Figura 5.7 pode-se observar que o composto 10 apresenta um alto valor de

influência quando comparado ao dos outros compostos o que o caracteriza como sendo atípico. Outra observação é que os compostos 2 e 23 apresentam um valor de resíduo de Student ligeiramente abaixo do limite inferior. Estes dois últimos devem ser temporariamente excluídos, o modelo é refeito e a melhora causada deve ser avaliada, Caso ela seja significativa, eles devem ser eliminados e caso contrário, permanecem no modelo. A amostra 2 tem influência baixa e, portanto não deve causar alterações no vetor de regressão, o que não ocorre para o composto 23 tem uma influência mais significativa. Quando os três compostos são removidos, o valor de Q^2 muda de 0,97 para 0,98, melhorando o modelo estatisticamente, embora não seja uma melhora significativa. Os resíduos altos observados para os compostos 2 e 23 podem ser um indicativo de incerteza nas medidas experimentais. No entanto, a remoção das amostras deve ser realizada cuidadosamente, pois uma explicação química ou biológica deve ser dada para cada amostra classificada como atípica.

5.3.5. Validação cruzada excluindo-N-amostras

Se o processo de validação cruzada excluindo- N -amostras for repetido inúmeras vezes para diferentes valores de N , serão obtidos diferentes valores do coeficiente de correlação de validação cruzada (Q^2) para cada execução. Além disso, mesmo que valores iguais de N sejam usados (desde que esse valor não seja igual a um), diferentes execuções do procedimento *leave-N-out* também levam a valores distintos de Q^2 , pois a ordem das amostras na matriz é aleatorizada antes da retirada dos grupos durante o procedimento de validação cruzada.

	Leverage	R. Student
Amostra1	0,131	0,228
Amostra2	0,025	-2,219
Amostra3	0,032	-0,123
Amostra4	0,017	0,087
Amostra5	0,026	0,533
Amostra6	0,038	1,444
Amostra7	0,04	-0,978
Amostra8	0,051	0,747
Amostra9	0,037	-0,56
Amostra10	0,482	0,926
Amostra11	0,198	0,285
Amostra12	0,063	1,333
Amostra13	0,025	-0,197
Amostra14	0,046	-0,763
Amostra15	0,025	1,491
Amostra16	0,019	-0,357
Amostra17	0,208	0,011
Amostra18	0,043	0,607
Amostra19	0,022	1,257
Amostra20	0,074	-1,669
Amostra21	0,003	-1,122
Amostra22	0,191	0,788
Amostra23	0,204	-2,3
Amostra24	0,03	-0,557
Amostra25	0,078	0,458

Figura 5.6. Resultado da detecção de amostras anômalas mostrando os valores de influência e dos resíduos de Student para os compostos do conjunto de treinamento.

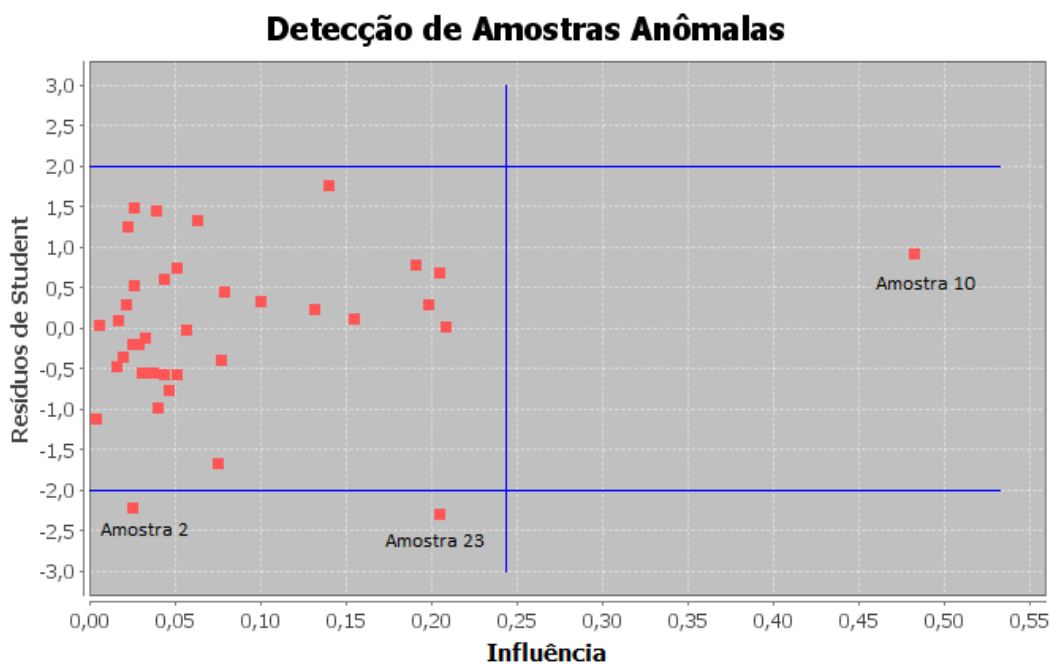


Figura 5.7. Gráfico de Influência *versus* Resíduos de Student para a detecção de amostras anômalas (*outliers*). As linhas azuis indicam os limites aceitos pela literatura.

No entanto, estes valores de Q^2 não deveriam ser muito diferentes entre si. Como o modelo é construído com o objetivo de prever as atividades de novas amostras, ele não deveria ser sensível às amostras removidas durante a validação cruzada. Assim, para avaliar a robustez do modelo, é altamente recomendável executar repetidos testes da validação cruzada *leave-N-out* para diferentes valores de N (variando de dois até 20% a 30 % do número de compostos) [59].

A robustez do modelo é avaliada pelo procedimento *leave-N-out* com o programa *QSAR modeling*. Neste processo é possível escolher o número máximo de amostras a serem removidas durante a validação cruzada, o número de variáveis latentes, que é mantido fixo durante a validação do modelo, assim como o número de repetições em cada validação para cada número de amostras removidas (Figura 5.8). O programa mostra como resultado uma tabela contendo os valores de *RMSECV* ou de Q^2 dependendo da escolha do usuário. Na Figura 5.9 estão os resultados do teste para um modelo com 3 LV, 3 repetições e um número máximo de 10 amostras removidas para um dos parâmetros estatísticos calculados por este teste.

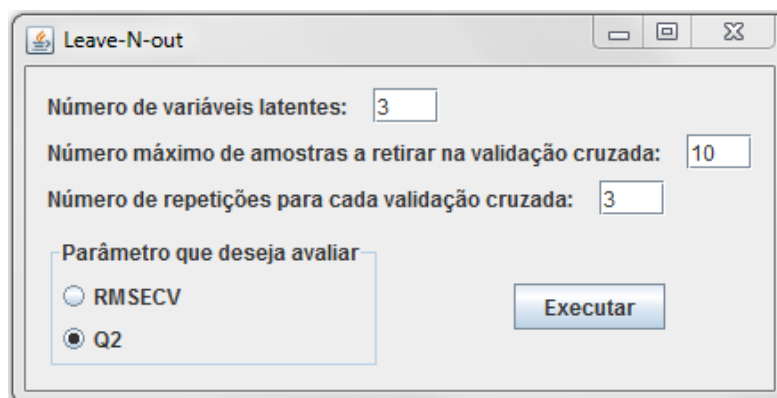


Figura 5.8. Procedimento de validação *leave-N-out* para garantir a robustez de um modelo usando o programa *QSAR modeling*.

O modelo de regressão obtido após a seleção de variáveis com o algoritmo OPS foi submetido ao procedimento de validação *leave-N-out* e os resultados são

apresentados na Figura 5.10. Como pode ser visto, o modelo pode ser considerado robusto, já que pequenas flutuações no valor de Q^2 são observadas com até 10 amostras removidas. Para cada valor de N o procedimento foi repetido três vezes (triplicata).

	Repetição 1	Repetição 2	Repetição 3
Leave-1-out	0,967	0,967	0,967
Leave-2-out	0,967	0,966	0,967
Leave-3-out	0,963	0,964	0,966
Leave-4-out	0,968	0,944	0,959
Leave-5-out	0,964	0,966	0,963
Leave-6-out	0,964	0,964	0,961
Leave-7-out	0,965	0,965	0,963
Leave-8-out	0,941	0,96	0,961
Leave-9-out	0,969	0,968	0,968
Leave-10-out	0,956	0,957	0,96

Figura 5.9. Resultados obtidos com o procedimento de validação *leave-N-out*.

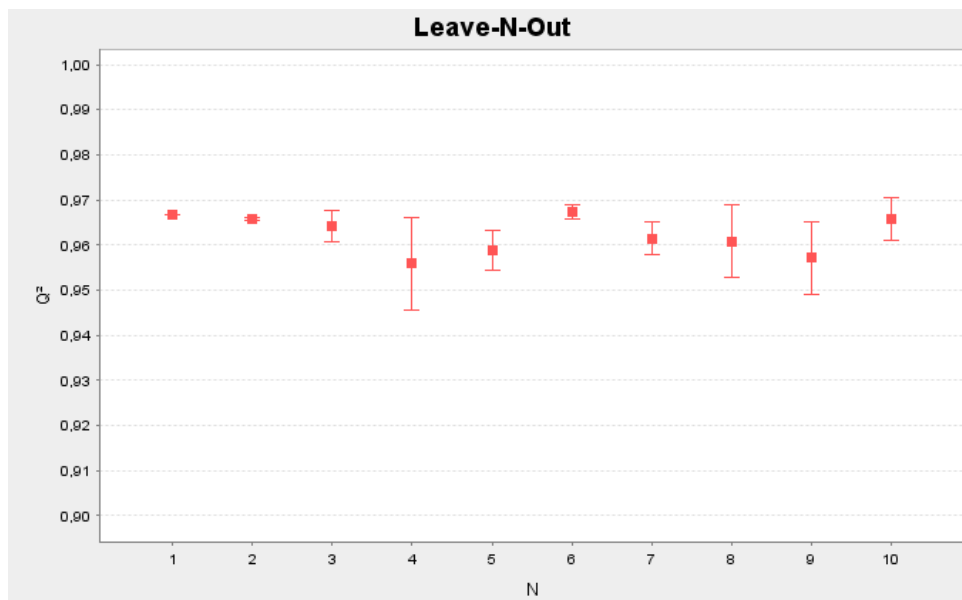


Figura 5.10. Validação *leave-N-out* aplicada ao modelo final obtido depois da seleção de variáveis com o algoritmo OPS. Os pontos representam a média e as barras indicam o desvio padrão de uma triplicata para cada valor de N . O modelo mostrou-se robusto até um valor de N igual a 11 (30% das amostras).

5.3.6. Teste de aleatorização de y (y-randomization)

A proposta do teste de aleatorização de y é detectar e quantificar correlações ao acaso entre a variável dependente e os descritores [59,67,90,122]. Para obter uma estimativa da significância de um valor de Q^2 obtido para um dado modelo, deve-se construir modelos paralelos com os valores de atividade biológica (vetor \mathbf{y}) permutados enquanto que os descritores originais (matriz \mathbf{X}) são mantidos fixos. Espera-se que os modelos paralelos construídos nestas condições sejam de péssima qualidade e com valores de Q^2 bem menores do que o valor obtido para o modelo real, garantindo assim que o modelo real não foi obtido ao acaso.

No processo de aleatorização de y usando o programa *QSAR modeling* é possível escolher o número de aleatorizações que serão executadas neste passo da validação (Figura 5.11). O programa fornece como resultado uma tabela contendo os valores R^2 e Q^2 calculados para os modelos obtidos com as atividades biológicas trocadas e o coeficiente de correlação de Pearson ($r(\mathbf{y}_{al}, \mathbf{y})$) entre o vetor \mathbf{y} com as atividades biológicas corretas e os vetores gerados com as atividades aleatorizadas, \mathbf{y}_{al} (Figura 5.12). A última linha desta tabela contém os valores de R^2 e Q^2 referentes ao modelo real para que sejam comparados com aqueles obtidos para os modelos paralelos.

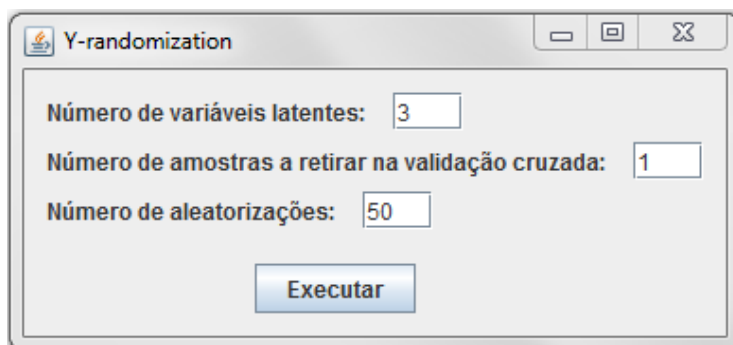


Figura 5.11. Procedimento de validação de aleatorização de y para verificar a correlação ao acaso de um modelo usando o programa *QSAR modeling*.

R2	Q2	R(yal,y)
0,173	-0,255	0,246
0,061	-0,7	0,13
0,143	-0,332	0,216
0,203	-0,35	0,011
0,201	-0,261	0,08
0,131	-0,381	0,259
0,16	-0,392	0,107
0,23	-0,345	0,04
0,228	-0,095	0,075
0,245	-0,346	0,047
0,118	-0,287	0,011
0,108	-0,581	0,082
0,203	-0,301	0,082
0,107	-0,916	0,016
0,233	-0,225	0,072
0,188	-0,853	0,245
0,067	-0,52	0,012
0,196	-0,402	0,278
0,092	-0,467	0,075
0,148	-0,719	0,087
0,164	-0,373	0,271
0,211	-0,262	0,145
0,257	-0,298	0,141
0,226	-0,386	0,062
0,122	-0,345	0,2

Figura 5.12. Resultados do teste de aleatorização de y fornecidos pelo programa *QSAR modeling*.

O modelo obtido após a seleção de variáveis com o algoritmo OPS foi submetido ao teste de aleatorização de y realizando-se 50 aleatorizações e retirando-se uma amostra (*leave-one-out*) em cada uma delas. Os resultados são apresentados na Figura 5.13. Como pode ser visto, todos os valores de R^2 e Q^2 dos modelos obtidos com o vetor aleatorizado, y_{al} , são menores que 0,4 e 0,0, respectivamente [122], confirmando que o modelo real não foi obtido ao acaso.

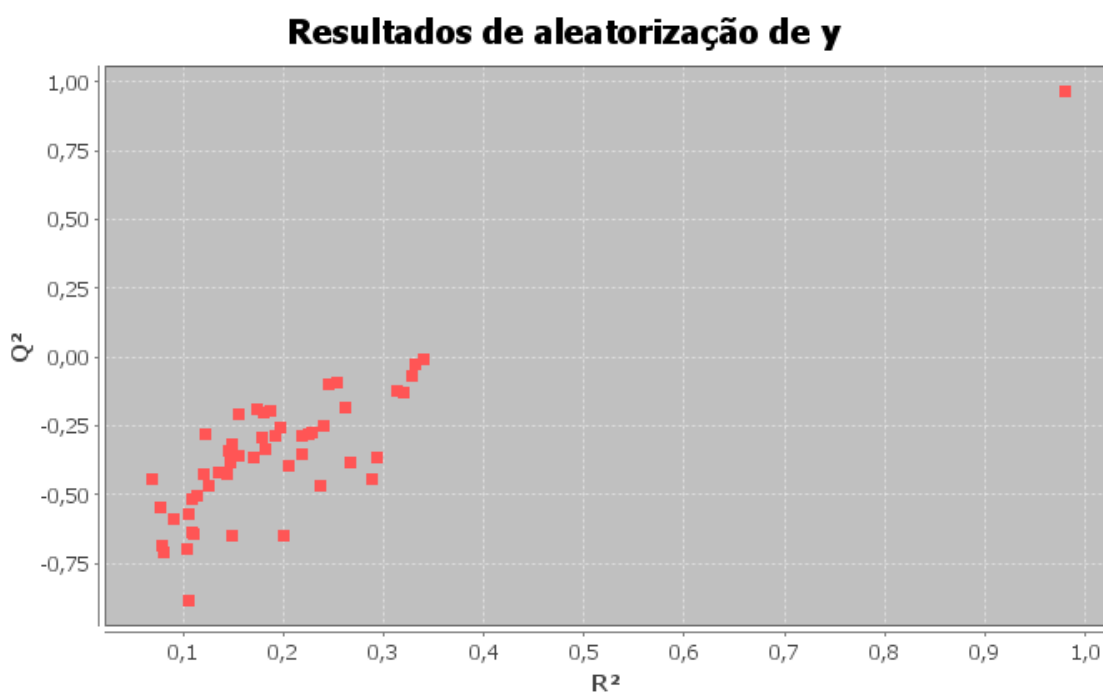


Figura 5.13. Valores de R^2 e Q^2 obtidos com o teste de aleatorização de y . O ponto distante representa os valores de R^2 e Q^2 para o modelo real.

5.3.7. Comparação com alguns dos softwares citados

Com o objetivo de atestar a qualidade do modelo obtido pelo programa *QSAR modeling*, o mesmo conjunto de dados foi utilizado para a construção de modelos com alguns dos softwares citados na Tabela 5.1 e que também são usados para a construção de modelos QSAR.

O programa VCCLAB [127], disponível gratuitamente na web, foi utilizado para a construção de um modelo PLS. O programa utiliza PLS como método de regressão e realiza a seleção de variáveis em duas etapas: i) elimina descritores que são praticamente constantes; ii) seleciona os demais descritores através de algoritmo genético com base nos valores de Q^2 dos modelos gerados. A seleção de variáveis levou a um modelo com 190 descritores e 2 variáveis latentes, com um valor de Q^2 igual a 0,963. Apesar de o modelo apresentar um valor de Q^2 , próximo ao obtido com o programa *QSAR modeling*

(0,967), a interpretação física do modelo é impossível devido ao número excessivo de descritores. Além disso, como o método PLS é tendencioso, a projeção de 190 descritores em apenas 2 variáveis latentes pode levar à perda de informação importante devido à ocorrência de subajuste. Infelizmente o número de variáveis latentes é selecionado automaticamente pelo programa, impedindo assim a análise de modelos com outros números de variáveis latentes.

O programa BuildQSAR [125], disponível para download gratuitamente, foi usado para a construção de um modelo MLR. Apesar de o programa disponibilizar também o método de regressão PCR, a seleção de variáveis, que pode ser feita através de busca sistemática ou com o algoritmo genético, só pode ser feita utilizando a regressão MLR. No modelo obtido depois da seleção de variáveis feita com o algoritmo genético, foram obtidos 7 descritores e nenhum *outlier* detectado e o valor de Q^2 foi de 0,936, também inferior ao valor obtido com o programa *QSAR modeling*. A matriz com estes descritores selecionados foi usada no programa *QSAR modeling* e observou-se que o modelo não passa nos testes de aleatorização de y e *leave-N-out*.

O programa Wolf [133] também foi usado para a construção de um modelo MLR. No modelo obtido depois da seleção de variáveis feita com algoritmo genético, foram obtidos 5 descritores e, depois da remoção de uma amostra detectada como anômala (amostra 23), o valor de Q^2 foi de 0,961, também inferior ao valor obtido com o *QSAR modeling*. A matriz com estes descritores selecionados foi usada no programa *QSAR modeling* e observou-se que o modelo não passa no teste de aleatorização de y .

5.4. Conclusões

O programa *QSAR modeling* permite a construção de modelos QSAR ou QSPR de uma maneira simples e rápida. Além disso, ele reúne em um único programa um algoritmo de seleção de variáveis, recentemente desenvolvido para construir modelos

PLS, um procedimento para detecção de amostras com comportamento anômalo e os principais procedimentos de validação exigidos atualmente pela comunidade científica.

Um conjunto de dados formado por 37 hidrocarbonetos poliaromáticos com log P determinado experimentalmente foi usado para ilustrar o uso de todas as ferramentas fornecidas pelo programa QSAR *modeling* e os resultados obtidos foram superiores aos obtidos por outros programas utilizados a título de comparação. Além disso, pôde-se observar que muitas funcionalidades disponíveis no programa QSAR *modeling* não estão disponíveis em outros programas.

Por ser um programa de código aberto, QSAR *modeling* é uma nova ferramenta para estudos de QSAR disponível para qualquer pessoa que desejar usá-la e, assim, pode ser melhorada para necessidades específicas em diversos campos de pesquisa.

O programa se encontra disponível para *download* no site lqta.iqm.unicamp.br.

Conclusão geral e perspectivas futuras

Em determinadas áreas da química, a interface com outras ciências como física, biologia, matemática, estatística e computação é enorme. Nessa tese temos um exemplo claro de como isso acontece.

Essa multidisciplinaridade, aliada à diversidade de ferramentas existentes no mercado, torna complexo o estudo, a construção e a interpretação de modelos de QSAR.

Portanto, o trabalho desenvolvido nesta tese foi uma pequena tentativa de minimizar a complexidade enfrentada em estudos de QSAR através da elaboração de algoritmos e softwares que reúnem de maneira simples e amigável as principais ferramentas necessárias para a geração de descritores, construção e validação de modelos QSAR. Os resultados obtidos nesse trabalho mostram que os algoritmos e programas aqui desenvolvidos podem ser aplicados com sucesso em uma variedade de estudos QSAR.

Outra dificuldade encontrada em estudos de QSAR é o fato de a grande maioria dos programas utilizados na área serem proprietários e de alto custo. Em um país como o Brasil, onde grande parte das universidades sofre com orçamentos reduzidos, esse pode ser um fator que impeça a realização de um estudo QSAR de qualidade.

Assim, acredita-se que nesse trabalho foi dado um avanço no sentido de tornar a área de estudo QSAR acessível a todos que tiverem interesse e disposição. Por serem programas livres e de código aberto, LQTAgrid e *QSAR modeling* são apenas embriões, prontos para receber as mais diversas contribuições, crescerem cada vez mais e se tornarem referência em estudos de QSAR.

Como a área de QSAR é extremamente dinâmica e evolui rapidamente com o passar do tempo, os programas aqui desenvolvidos foram projetados de modo que essa evolução possa ser acompanhada de maneira eficiente com a facilidade de implementação de quaisquer novas funcionalidades.

Portanto, com a utilização, melhoria e evolução dos programas desenvolvidos, além do desenvolvimento de novos algoritmos e programas de acordo com novas necessidades, esse pode ser o início de uma série de trabalhos que levem a estudos de QSAR de sucesso e independentes.

Referências Bibliográficas

1. Golub, G. H.; Van Loan, C. F. **Lanczos Methods**. In: *Matrix computation*. 3^a ed. Baltimore: John Hopkins University Press, 1996. p. 470-507.
2. Golub, G. H.; Kahan, W. **Calculating the singular values and pseudo-inverse of a matrix**. *SIAM J. Num. Anal. Ser. B*. 1965, 2, 205-224.
3. Wold, S.; Johansson, E.; Cocchi, M. **PLS Partial Least Squares Projections to Latent Structures**. In: Kubinyi, H. *3D QSAR: Theory, Methods and Applications*. V. 1. Leiden: Kluwer Academic Publishers, 1993. p. 523-550.
4. Givehchi, A.; Dietrich, A.; Wrede, P.; Schneider, G. **ChemSpaceShuttle: A tool for data mining in drug discovery by classification, projection, and 3D visualization**. *QSAR Comb. Sci.* 2003, 47, 549-559.
5. Bao, J. S.; Cai, Y. Z.; Corke, H. **Prediction of rice starch quality parameters by near-infrared reflectance spectroscopy**. *J. Food Sci.* 2001, 66, 936-939.
6. Alam, T. M.; Alam, M. K. **Chemometric analysis of NMR spectroscopy data: A review**. *Annual reports on NMR spectroscopy* 2005, 54, 41-80.
7. Ribeiro, J. S.; Augusto, F.; Salva, T. J. G.; Thomaziello, R. A.; Ferreira, M. M. C. **Prediction of sensory properties of Brazilian Arabica roasted coffees by headspace solid phase microextraction-gas chromatography and partial least squares**. *Anal. Chim. Acta* 2009, 634, 172-179.
8. Gil, D. B.; de la Pena, A. M.; Arancibia, J. A.; Escandar, G. M.; Olivieri, A. C. **Second-order advantage achieved by unfolded-partial least-squares/residual bilinearization modeling of excitation-emission fluorescence data presenting inner filter effects**. *Anal. Chem.* 2006, 78, 8051-8058.
9. Cruz, S. C.; Aarnoutse, P. J.; Rothenberg, G.; Westerhuis, J. A.; Smilde, A. K.; Blik A. **Kinetic and mechanistic studies on the Heck reaction using real-time near infrared spectroscopy**. *Phys. Chem. Chem. Phys.* 2003, 5, 4455-4460.
10. Martens, H.; Naes, T. *Multivariate Calibration*. New York: Wiley, 1989. cap. 3.
11. Geladi, P.; Kowalski, B. R. **Partial least-squares regression - A tutorial**. *Anal. Chim. Acta* 1986, 185, 1-17.

12. Hoskuldsson, A. **PLS regression methods.***J. Chemom.***1988**,*2*, 211-228.
13. Helland, I. S. **Some theoretical aspects of partial least squares regression.***Chemom.Intell.Lab. Syst.* **2001**,*58*, 97-107.
14. Wold, S.; Sjoström, M.; Eriksson, L. **PLS-regression: a basic tool of chemometrics.***Chemom.Intell.Lab. Syst.* **2001**,*58*, 109-130.
15. Riu, J.; Bro R. **Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models.***Chemom.Intell. Lab. Syst.***2003**, *65*, 35-49.
16. Rajah, M. N.; McIntosh, A. R. **Overlap in the functional neural systems involved in semantic and episodic memory retrieval.***J. Cogn. Neurosci.***2005**,*17*, 470-482.
17. Li, X. M.; Zhou, J. X.; Yuan, S. H.; Zhou, X. P.; Fu, Q. **The effects of tolerance for ambiguity and uncertainty on the appropriateness of accounting performance measures.***Biomed. Environ. Sci.* **2008**,*41*, 45-52.
18. Pedro, A. M. K.; Ferreira, M. M. C. **Simultaneously calibrating solids, sugars and acidity of tomato products using PLS2 and NIR spectroscopy.***Anal.Chim.Acta.***2007**, *595*, 221-227.
19. Henrique, C. M.; Teófilo, R. F.; Sabino, L.; Ferreira, M. M. C.; Cereda M. P. **Classification of cassava starch films by physicochemical properties and water vapor permeability quantification by FTIR and PLS.***J. Food Sci.* **2007**,*72*, E184-E189.
20. Ferreira, M. M. C. **Multivariate QSAR.***J. Braz. Chem. Soc.* **2002**,*13*, 742-753.
21. Kiralj, R.; Ferreira, M. M. C. **Comparative chemometric and QSAR/SAR study of structurally unrelated substrates of a MATE efflux pump VmrA from *V. parahaemolyticus*: prediction of multidrug resistance.***QSAR Comb. Sci.* **2008**,*27*, 314-329.
22. Wu, W.; Manne, R. **Fast regression methods in a Lanczos (or PLS-1) basis. Theory and applications.***Chemom.Intell.Lab.Syst.***2000**,*51*, 145-161.
23. Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C. **Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression.***J. Chemom.***2009**,*23*, 32-48.

24. Haaland, D. M.; Thomas, E. V. **Partial least-squares methods for spectral analyses.1.relation to other quantitative calibration methods and the extraction of qualitative information.***Anal.Chem.* **1988,60**, 1193-1202.
25. Dayal, B. S.; MacGregor, J. F. **Improved PLS algorithms.***J. Chemom.***1997,11**, 73-85.
26. Lindgren, F.; Geladi, P.; Wold, S. **The kernel algorithm for PLS.***J. Chemom.***1993,7**, 45-59.
27. De Jong, S. **SIMPLS - An alternative approach to partial least-squares regression.***Chemom.Intell.Lab. Syst.* **1993,18**, 251-263.
28. Manne, R. **Analysis of 2 partial-least-squares algorithms for multivariate calibration.***Chemometrics Intell. Lab. Syst.* **1987,2**, 187-197.
29. Helland, I. S. **On the structure of partial least squares regression.***Commun.Stat.- Simul. Comput.***1988,17**, 581-607.
30. Lorber, A.; Kowalski, B. R. **A note on the use of the partial least-squares method for multivariate calibration.***Appl.Spectrosc.***1988,42**, 1572-1574.
31. Phatak, A.; de Hoog, F. **Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS.***J. Chemom.***2002,16**, 361-367.
32. Elden, L. **Partial least-squares vs. Lanczos bidiagonalization - I: analysis of a projection method for multiple regression.***Comput.Stat. Data Anal.***2004,46**, 11-31.
33. Pell, R. J.; Ramos, L. S.; Manne, R. **The model space in partial least square regression.***J.Chemom.***2007,21**, 165-172.
34. Teófilo, R. F.; Ferreira, M. M. C. **Chemometrics II: Spreadsheets for experimental design calculations, a tutorial.***Quim. Nova* **2006, 29**, 338-350.
35. Montgomery, D. C.; Runger G. C. *Applied Statistics and Probability for Engineers.* New York: Wiley, **2003**.
36. Dyrby, M.; Engelsen, S. B.; Norgaard. L.; Bruhn, M.; Lundsberg-Nielsen, L. **Chemometric quantitation of the active substance (containing C=N) in a**

pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra.*Appl. Spectrosc.*2002**,*56*, 579-585.**

37. Bro, R.; Rinnan, A.; Faber, N. M. **Standard error of prediction for multilinear PLS - 2. Practical implementation in fluorescence spectroscopy.***Chemom. Intell. Lab. Syst.* **2005**,*75*, 69-76.

38. Teófilo, R. F.; Kiralj, R.; Ceragioli, H. J.; Peterlevitz, A. C.; Baranauskas, V.; Kubota, L. T.; Ferreira, M. M. C. **QSPR Study of Passivation by Phenolic Compounds at Platinum and Boron-Doped Diamond Electrodes.***J. Electrochem. Soc.* **2008**, *155*, D640-D650.

39. Carbó-Dorca, R.; Robert, D.; Amat L.; Gironés, X.; Besalú E. *Molecular Quantum Similarity in QSAR and Drug Design*. Berlin: Springer, **2000**.

40. Hammett, L. P. **The Effect of Structure upon the Reactions of Organic Compounds Benzene Derivatives.***J. Am. Chem. Soc.* **1937**, *59*, 96-103.

41. Free, S. M.; Wilson, J. W. **A Mathematical Contribution to Structure-Activity Studies.***J. Med. Chem.***1964**, *7*, 1616-1626.

42. Hansch, C.; Fujita, T. **p- σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure.***J. Am. Chem. Soc.* **1964**, *86*, 1616-1626.

43. Wiener, H. **Structural determination of paraffin boiling points.***J. Chem. Phys.* **1947**, *69*, 17-20.

44. Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic M.; **Molecular connectivity. I: Relationship to nonspecific local anaesthesia.***J. Pharm. Sci.***1947**, *69*, 17-20.

45. Randic M. **On characterization of molecular branching.***J. Am. Chem. Soc.* **1975**, *97*, 6609-6615.

46. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Dragon 3.0 Web Version, 2003.

47. <https://www.chemaxon.com/download/marvin/for-end-users/> .Acessado em 10 de Março de 2013.

48. Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. **Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins.***J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.

49. Martinez-Merino, V.; Cerecetto H. **Comfa-simca model for antichagasic nitrofurazone derivatives.***Bioorg. Med. Chem.***2001**, 9, 1025-1030.
50. Zhao, W. N.; Yu, Q. S.; Zou, J. W.; Ma, M.; Zheng, K. W. **Three-dimensional quantitative structure-activity relationship study for analogues of TQXs using CoMFA e CoMSIA.***J. Mol. Struct. (Theochem)* **2005**, 723, 69-78.
51. Nilsson, J.; de Jong, S.; Smilde, A. K. **Multiway Calibration in 3D QSAR.***J. Chemom.***1997**, 11, 511-524.
52. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. **Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism.***J. Am. Chem. Soc.* **1997**, 119, 10509-10524.
53. Fujita, T. **Recent Success Stories Leading to Commercializable Bioactive Compounds with the Aid of Traditional QSAR Procedures.***Quant. Struct.-Act. Relat.***1997**, 16, 107-112.
54. Krohn, A.; Redshaw, S.; Ritchie, J. C .; Graves, B. J.; Hatada M. H. **Novel binding mode of highly potent HIV-proteinase inhibitors incorporating the (R)-hydroxyethylamine isostere.***J. Med. Chem.* **1991**, 34, 3340-3342.
55. Kawakami, Y.; Inoue, A.; Kawai, T.; Wakita, M.; Sugimoto, H.; Hopfinger, A. J. **The rationale for E2020 as a potent acetylcholinesterase inhibitor.***Bioorg. Med. Chem.***1996**, 4, 1429-1446.
56. Cardozo, M. G.; Iimura, Y.; Sugimoto, H.; Yamanishi, Y.; Hopfinger, A. J. **QSAR Analyses of the Substituted Indanone and Benzylpiperidine Rings of a Series of Indanone-Benzylpiperidine Inhibitors of Acetylcholinesterase.***J. Med. Chem.* **1992**, 35, 584-589.
57. Phatak, A.; de Jong, S. **The geometry of partial least squares.***J. Chemom.***1997**, 11, 311-338.
58. Golbraikh, A.; Tropsha, A. **Beware of q²!***J. Mol. Grap. Modell.***2002**, 20, 269-276.
59. Kiralj, R.; Ferreira, M. M. C. **Basic validation procedures for regression models in QSAR and QSPR studies: theory and applications.***J. Braz. Chem. Soc.***2009**, 20, 770-787.

60. Naes, T.; Isaksson, T.; Fearn, T.; Davies, T. **A User-Friendly Guide to Multivariate Calibration and Classification**. Chichester: NIR publications, 2002; p177-190.
61. Ferreira, M. M. C.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L. O. **Quimiometria I: Calibração multivariada, um tutorial**. *Quim.Nova*. **1999**, 22, 724-731.
62. Pasqualoto, K. F. M.; Ferreira, E. I.; Santos-Filho, O. A.; Hopfinger, A. J. **Rational design of new antituberculosis agents: receptor-independent four-dimensional quantitative structure–activity relationship analysis of a set of isoniazid derivatives**. *J. Med. Chem.* 2004, 47, 3755-3764.
63. Martins, J. P. A.; Barbosa, E. G.; Pasqualoto, K. F. M.; Ferreira, M. M. C. **LQTA-QSAR: A new 4D-QSAR methodology**. *J. Chem. Inf. Model.* **2009**, 49, 1428-1436.
64. Rogers, D.; Hopfinger, A. J. **Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships**. *J. Chem. Inf. Comput. Sci.*, 1994, 34, 854–866.
65. de Melo, E. B.; Ferreira, M. M. C. **Multivariate QSAR study of 4,5-dihydropyrimidine carboxamides as HIV-1 integrase inhibitors**. *Eur. J. Med. Chem.* 2009, 44, 3577-3583.
66. Hernández, N.; Kiralj, R.; Ferreira, M. M. C.; Talavera, I. **Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors**. *Chemom.Intell. Lab. Syst.* 2009, 98, 65-77.
67. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. **Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs**. *Environ. Health.Perspect.* 2003, 111, 1361-1375.b
68. Bratchell, N. **Chemometric methods in molecular design**. *J. Chemometr.* **1997**, 11, 93-94.
69. Purohit, V.; Basu, A. K. **Mutagenicity of nitroaromatic compounds**. *Chem. Res. Toxicol.* **2000**, 13, 673-692.
70. Edenharder, R.; Tang, X. **Inhibition of the mutagenicity of 2-nitrofluorene, 3-nitrofluoranthene and 1-nitropyrene by flavonoids, coumarins, quinones and other phenolic compounds**. *Food Chem. Toxicol.* **1997**, 35, 357-372.

71. Maron, M.D.; Ames, B.N. **Revised methods for the Salmonella mutagenicity test.** *Mutation Res.* **1983**, *113*, 173-215.
72. Silva, C.R.M.; Naves, M.M.V. **Suplementação de vitaminas na prevenção de câncer.** *Rev. Nutr.* **2000**, *14*, 135-143.
73. Oliveira, V.M.; Aldrighi, J.M.; Rinaldi, J.F. **Quimioprevenção do câncer de mama.** *Rev. Assoc. Med. Bras.* **2006**, *52*, 453-459.
74. Tsao, A.S.; Kim, E.S.; Hong, W.K. **Chemoprevention of Cancer.** *CA Cancer J. Clin.* **2004**, *54*, 150-180.
75. Lameira, J.; Medeiros, I.G.; Reis, M.; Santos, A. S.; Alves, C.N. **Structure–activity relationship study of flavone compounds with anti-HIV-1 integrase activity: A density functional theory study.** *Bioorg. Med. Chem.* **2006**, *14*, 7105-7112.
76. Cambridge Crystallographic Data Centre Inc.; The Cambridge Structural Database. 2009.
77. Hypercube Inc. HyperChem 7.1 for Windows. 2002.
78. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Laham, A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A., GAUSSIAN03, revision C.02, Department of Chemistry, Carnegie Mellon University: Pittsburgh, PA, 2003.
79. Molfetta, F.A.; Bruni, A.T.; Rosseli, F.P.; Silva, A.B.F.; **A partial least squares and principal component regression study of quinone compounds with trypanocidal activity.** *Struct. Chem.* **2007**, *18*, 49-57.

80. Parameter Client interface. By Virtual Computational Chemistry Laboratory, **2005**.Disponível <http://www.vcclab.org/lab/pclient>.Acessado em março de 2013.
81. ALOGPS interface version 2.1. By Virtual Computational Chemistry Laboratory, **2005**.Disponível <http://www.vcclab.org/lab/alogps>.Acessado em março de 2013.
82. Informetrix Inc. Pirouette 4 for Windows. 2007.
83. Martins, J. P. A.; Teófilo, R. F.; Ferreira, M. M. C. **Computational performance and cross-validation error precision of five PLS algorithms using designed and real data set.***J. Chemom.***2010**, *24*, 320-332.
84. Rücker,C.; Rücker,G.; Meringer,M.; **y-Randomization and Its Variants in QSPR/QSAR.***J. Chem. Inf. Model.***2007**, *47*, 2345-2357.
85. Wold,S.; Eriksson,L.**Statistical Validation of QSAR Results.**In: H. van de Waterbeemd (Ed.), *Chemometric Methods in Molecular Design*. Weinheim: Wiley-VCH, **1995**. p. 309-318.
86. Gaudio,A.C.; Zandonade,E.**Proposição, Validação e Análise dos Modelos que Correlacionam Estrutura Química e Atividade Biológica.***Quim.Nova***2001**, *24*, 658-671.
87. OECD-Organization for Economic Co-Operation and Development. Guidance Document onthe Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. Paris:OECD, **2007**. Disponível em <http://www.oecd.org/ehs>.
88. Roy,P.P.; Leonard,J.T.; Roy,K.**Exploring the impact of size of training sets for the development of predictive QSAR models.***Chemom. Intell. Lab. Syst.***2008**, *90*, 31-42.
89. Roy,P.P.; Roy,K.; **On Some Aspects of Variable Selection for Partial Least Squares Regression Models.***QSAR Comb. Sci.***2008**, *27*, 302-313.
90. Tropsha,A.; Gramatica,P.; Gombar,V.K.**The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models.***QSAR Comb. Sci.***2003**, *22*, 69-77.
91. Melagraki,G.; Afantitis,A.; Sarimveis,H.;Koutentis,P.A.; Markopolus,J.; Igglessi-Markopoulou,O. **Optimization of biaryl piperidine and 4-amino-2-biarylurea MCH1**

receptor antagonists using QSAR modeling, classification techniques and virtual screening.*J. Comput.-Aided Mol. Des.*2007**, *21*, 251-267.**

92. Aptula, A.O.; Jeliaskova, N.G.; Schultz, T.W.; Cronin, M.T.D. **The Better Predictive Model: High q^2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set?***QSAR Comb. Sci.***2005**, *24*, 385–396.

93. Farkas, O.; Jakus, J.; Héberger, K. **Quantitative Structure – Antioxidant Activity Relationships of Flavonoid Compounds.***Molecules***2004**, *9*, 1079-1088.

94. Hatch, F.T.; Lightstone, F.C.; Colvin, M.E. **QSAR of Flavonoids for Inhibition of Heterocyclic Amine Mutagenicity.***Environ. Mol. Mutagen.***2000**, *35*, 279-299.

95. Heo, M.Y.; Sohn, S.J.; Au, W.W.; **Anti-genotoxicity of galangin as a cancer chemopreventive agent candidate.** *Mutat. Res., Fundam. Mol. Mech. Mutagen.***2001**, *488*, 135-150.

96. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Weinheim: Wiley-VCH, 2000. 667 pp.

97. Morales, A.H.; Duchowicz, P.R.; Pérez, M.Á.C.; Castro, E.A.; Cordeiro, M.N.D.S.; González, M.P.; **Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential.***Chemom. Intell. Lab. Syst.***2006**, *81*, 180–187.

98. Put, R.; Xu, Q.S.; Massart, D.L.; Heyden, Y.V. **Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies.***J. Chromatogr., A***2004**, *1055*, 11–19.

99. Rasulev, B.F.; Abdullaev, N.D.; Syrov, V.N.; Leszczynski, J.A **Quantitative Structure-Activity Relationship (QSAR) Study of the Antioxidant Activity of Flavonoids.***QSAR Comb. Sci.***2005**, *24*, 1056-1065.

100. Consonni, V.; Todeschini, R.; Pavan, M. **Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors.***J. Chem. Inf. Comp. Sci.***2002**, *42*, 682-692.

101. González, M.P.; Terán, C.; Teijeira, M.; González-Moa, M.J. **GETAWAY descriptors to predicting A_{2A} adenosine receptors agonists.** *Eur. J. Med. Chem.***2005**, *40*, 1080-1086.

102. Wen-Na, Z.; Qing-Sen, Y.; Jian-Wei, Z.; MA, M.; Ke-Wen, Z., **Three-dimensional quantitative structure-activity relationship study for analogues of TQXs using CoMFA and CoMSIA.***J. Mol. Struct. (Theochem)* **2005**,*723*, 69-78.
103. Lindahl, E.; Hess, B.; van der Spoel, D.**GROMACS 3.0: a package for molecular simulation and trajectory analysis.***J. Mol. Model.* **2001**,*7*, 306-317.
104. Spoel, D. v. d.; Lindahl, E.; Hess, B.; Buuren, A. R. v.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Feenstra, K. A.; Drunen, R. v.; Berendsen, H. J. C. *Gromacs User Manual version 3.3*, **2005**. GROMACS: Fast, Free and Flexible MD-Paper Manuals. Disponível <http://www.gromacs.org/>. Acessado em março de 2013.
105. Schttelkopf, A. W.; van Aalten, D. M. F.**PRODRG: a tool for high-throughput crystallography of protein ligand complexes.***Acta Cryst. D* **2004**,*60*, 1355-1363.
106. Venkatarangan, P.; Hopfinger, A. J. **Prediction of ligand-receptor binding free energy by 4D-QSAR analysis: Application to a set of glucose analogue inhibitors of glycogen phosphorylase.***J. Chem. Inf. Comput. Sci.* **1999**,*39*, 1141-1150.
107. Ravindra, G. K.; Achaiah, G.; Sastry, G. N. **Molecular modeling studies of phenoxyrimidinyl imidazoles as p38 kinase inhibitors using QSAR and docking.***Eur. J. Med. Chem.* **2008**,*43*, 830-838.
108. Watson, K. A.; Chrysina, E. D.; Tsitsanou, K. E.; Zographos, S. E.; Archontis, G.; Fleet, G. W. J.; Oikonomakos, N. G. **Kinetic and crystallographic studies of glucopyranose spirohydantoin and glucopyranosylamine analogs inhibitors of glycogen phosphorylase.***Proteins: Struct. Funct. Bioinf.***2005**,*61*, 966-983.
109. Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisà, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. **Structural basis of inhibitor selectivity in MAP kinases.***Structure* **1998**,*6*, 1117-1128.
110. Lee, C.; Yang, W.; Parr, R. G. **Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density.***Phys. Rev. B* **1988**,*37*, 785.
111. Breneman, C. M.; Wiberg, K. B.**Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis.***J. Comput. Chem.* **1990**,*11*, 361-373.

112. Gasteiger, J.; Marsili, M. **Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges.** *Tetrahedron* **1980**, *36*, 3219-3228.
113. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. **Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.** *J. Comput. Chem.* **1998**, *19*, 1639-1662.
114. Kusalik, P. G.; Svishchev, I. M. **The spatial structure in liquid water.** *Science* **1994**, *265*, 1219-1221.
115. Darden, T.; York, D.; Pedersen, L. **Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems.** *J. Chem. Phys.* **1993**, *98*, 10089-10092.
116. Berk, H.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. **LINCS: A linear constraint solver for molecular simulations.** *J. Comput. Chem.* **1997**, *18*, 1463-1472.
117. Parrinello, M.; Rahman, A. **Crystal structure and pair potentials: A molecular dynamics study.** *Phys. Rev. Lett.* **1980**, *45*, 1196.
118. Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; DiNola, A.; Haak, J. R. **Molecular dynamics with coupling to an external bath.** *J. Chem. Phys.* **1984**, *81*, 3684-3690.
119. Barbosa, E. G.; Ferreira, M. M. C. **Digital Filters for Molecular Interaction Field Descriptors.** *Mol. Inform.* **2012**, *31*, 75-84.
120. Kiralj, R.; Ferreira M. M. C. **Is your QSAR/QSPR descriptor real or trash? J. Chemometr.** **2010**, *24*, 681-693.
121. Ortiz, A.R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R.C. **Reliability of Comparative Molecular Field Analysis Models: Effects of Data Scaling and Variable Selection Using a Set of Human Synovial Fluid Phospholipase A₂ Inhibitors.** *J. Med. Chem.* **1997**, *40*, 1136-1148.
122. Ferreira, M.M.C.; Kiralj, R. In: Montanari, C. *Química Medicinal, Métodos e Fundamentos em Planejamento de Fármacos.* São Paulo: EDUSP, **2011**. cap. 12.
123. Gramatica, P. **Principles of QSAR models validation: internal and external.** *QSAR & Comb. Sci.* **2007**, *26*, 694-701.

124. Talete srl.MobyDigs, Version 1. 2004.
- 125.de Oliveira, D. B.; Gaudio, A. C.**BuildQSAR: A New Computer Program for QSAR Analysis.***Quant. Struct.-Act. Relat.***2000**, *19*, 599-601.
126. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V.**Virtual Computational Chemistry Laboratory – Design and Description.***J. Comput. Aid. Mol. Des.***2005**, *19*, 453-463.
127. VCCLAB, Virtual Computational Chemistry Laboratory, 2005. <http://www.vcclab.org>. Acessado em 10 de Março de 2013.
128. Cerius QSAR+, 2000. <http://www.esi.umontreal.ca/accelrys/pdf/qsarC45.pdf>. Acessado em 10 de Março de 2013.
129. Bilinear Model, BILIN, 1976. <http://www.kubinyi.de/bilin-program.html>. Acessado em 10 de Março de 2013.
130. Molecular Structure Generation MOLGEN QSPR, 2003.<http://www.molgen.de/?src=documents/molgenqspr.html>. Acessado em 10 de Março de 2013.
131. Correlation and Logic, CORAL, 2010. <http://www.insilico.eu/coral/> Acessado em 10 de Março de 2013.
132. Comprehensive Descriptors for Structural and Statistical Analysis, CODESSA PRO, 2001. <http://www.codessa-pro.com/index.htm>. Acessado em 10 de Março de 2013.
133. D. Rogers, WOLF Reference Manual Version 5.5, The Chem21 Group Inc., Chicago, IL 1994.
134. <http://www.moleculardescriptors.eu/dataset/dataset.htm>. Acessado em 10 de Março de 2013.
135. *Java*, version 6 update 10; java development kit; Sun microsystems, Inc: Santa Clara, CA 95054 USA, 2008.

136. Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*. New York: Wiley, 1989.

137. de Melo, E. B.; Ferreira, M. M. C. **Four-Dimensional Structure–Activity Relationship Model to Predict HIV-1 Integrase Strand Transfer Inhibition using LQTA-QSAR Methodology.** *J. Chem. Inf. Model.* **2012**, *52*, 1722-1732.

138. Barbosa, E. G.; Pasqualoto, K. F. M.; Ferreira, M. M. C. **The receptor-dependent LQTA-QSAR: application to a set of trypanothione reductase inhibitors.** *J. Comput.-Aided Mol. Des.* **2012**, *26*, 1055-1065.