IV. Discussão

IV.1. Introdução

Há mais de quinze anos diversos aspectos dos estudos de *R*elações *Q*uantitativas entre *E*strutura química e *A*tividade biológica (*QSAR*) e (*QSAR-3D)* estão sendo abordados em nosso grupo de pesquisa no Laboratório de *QSAR* e Modelagem Molecular, no Instituto de Química da Universidade de São Paulo, São Paulo. Entre eles, o estudo e a determinação de propriedades físico-químicas/estruturais (eletrônicas, lipofílicas, estéricas e àquelas relativas à polarizabilidade) de compostos com atividade biológica, em especial de anestésicos locais e de bloqueadores da transmissão do impulso nervoso, estruturalmente análogos da procaína (Tavares. L.C. 1987; Amaral,.*et al*.1991; Amaral,. *et al*. 1993; Miguel 1993; Sousa 1997; Amaral *et al.* 1997; Malvezzi, *et al.* 2001; Tavares & Amaral 1997); Siqueira 2001; Malvezzi 2003); de antibacterianos (Baroni 1987; Tavares. L.C. 1993) incluindo-se derivados nitrofurânicos 5-substituídos (Pires 1998) (Pires *et al.* 2001) (Cheng, E., *et al* 2002); de antitumorais derivados de sais de alquilamônio de N-alquil arilcetonas substituídas (Ramineli 2001) e, mais recentemente de derivados não nucleosídeos inibidores da transcriptase reversa (NNRTI) do vírús HIV-1 (Ishiki 1999; Ishiki, *et al.* 2001), de derivados pirídínicos–alfa-substituídos inibidores potencias da Ribonucleotideo Reductase da *Mycobacterium tuberculosis* (Ishiki & Amaral 2005) e de derivados nitrofurânicos antichagásicos (Rando *et al.* 2002; Malvezzi, *et al* 2002).

Estudos de Relação Quantitativa entre Estrutura química e Atividade biológica (*QSAR/QSAR-3D*) utilizando diferentes estratégias metodológicas complementares, aplicadas iterativamente, se estende a diferentes Áreas de aplicação, seja no planejamento racional de novos fármacos e defensivos agrícolas, nos estudos de seus mecanismos de ação, na previsão da toxicidade de compostos e no controle ambiental. Na literatura são encontrados trabalhos de revisão que mostram vantagens e limitações das metodologias aplicadas em *QSAR*, sendo aplicada em diversos sistemas e atividade biológica [Fujita, 1990; Kubinyi, 1993; Hansch, 1995].

Considera-se, fundamentalmente, que a atividade biológica de um composto é resultado das interações deste com as diferentes biofases, podendo ser expressa pela

contribuição de propriedades físico-químicas do composto (Análise de Hansch ou Abordagem Extratermodinâmica) ou através da contribuição das características estruturais, expressas pelas variáveis indicadoras,(**I),** (Análise de Free-Wilson). Nestas abordagens, procura-se estender os conceitos utilizados nos estudos dos mecanismos de reação em Química Orgânica, Físico-Química Orgânica para sistemas mais complexos, ou seja, no entendimento e previsão de mecanismos que ocorrem em sistemas mais complexos, como bioquímicos ou em animais.

Em todos os projetos de nosso grupo de pesquisa no Laboratório de *QSAR* e Modelagem Molecular, no IQUSP procuramos, de um modo geral, desenvolver e/ou aplicar as metodologias utilizadas em *QSAR/QSAR-3D* para descrever, de modo quantitativo, a(s) interação(ões) entre o composto químico e o sistema biológico e, assim procuramos contribuir tanto para elucidar o(s) mecanismo(s) das interações expressas por cada um dos parâmetros ou de suas características estruturais, bem como para esclarecer o mecanismo de ação e prever derivados mais potentes.

O emprego destas abordagens envolve tanto a proposição de um modelo matemático como a medida, ou o cálculo dos parâmetros responsáveis pela atividade e de suas contribuições relativas para a mesma.

A análise dos resultados, em geral, apresentados por equações e/ou por gráficos permite verificar tanto a validade, limitações e poder de previsão do modelo proposto. O grau de complexidade do modelo dependerá de quão exatamente essas interações são conhecidas e/ou podem ser expressas.

Nos diversos sistemas estudados no laboratório, os modelos de QSAR gerados foram validados estatisticamente considerando-se diferentes aspectos, entre eles: (i)o planejamento das séries de compostos e dos substituintes (ii) os erros sistemáticos intrínsecos de cada metodologia utilizada e, (iii) a seleção de variáveis.

Nos últimos anos, no entanto, os avanços recentes em várias Áreas do conhecimento como na Informática; Química Combinatória, cristalografia de raio-X fizeram com que as abordagens metodológicas até então empregadas, fossem reavaliadas.

Assim, embora modelos de sucesso tenham sido gerados, as abordagens empregadas foram, em princípio, propostas para séries restritas de compostos e de parâmetros (n/p=5).

Atualmente, no entanto, nos estudos em QSAR/QSAR-3D enfatiza-se a necessidade de se considerar, adequadamente e de modo racional, todo o imenso universo de informações disponíveis. Estes exigem que se incluam, entre outros procedimentos, (i) a seleção das informações relevantes, envolvendo o pré-tratamento de dados (ii) o emprego de métodos que permitam reduzir o volume de informações, sem acarretar, no entanto, a perda de informações relevantes. (iii) a proposição de métodos de validação dessas abordagens e, dos modelos gerados.

Com relação a pesquisa de compostos antichagásicos, nos seus diferentes aspectos, é grande o número de dados disponíveis na literatura ou, mesmo ainda não publicados (Coura & de Castro 2002).

Com o objetivo de racionalizar as informações contidas neste grande número de dados, estamos desenvolvendo e aplicando metodologias que possibilitem descodificá-las nas características responsáveis pela atividade antichagásica e, assim, utilizá-las na pesquisa de possíveis candidatos à fármacos.

No presente trabalho, destinado à dissertação de mestrado, aplicamos metodologias, na sua maioria descritas na literatura, às propriedades físico-químicas e estruturais, calculadas pelo programa DRAGON, de séries de análogos de semicarbazonas com atividade antichagásica, selecionados da literatura, visando pesquisar e, propor critérios para obtenção e validação de modelos QSAR. E, utilizando metodologias ainda novas para o grupo, pode-se estudar e estabelecer relação(ões) quantitativa(s) entre a estrutura química e atividade biológica, desta série de inibidores da cruzaína, contribuindo para o melhor conhecimento da natureza e da contribuição das características estruturais responsáveis pelas interações entre esta série de ligantes e a cruzaína.

Foram selecionadas da literatura 3 séries de compostos: série I contendo 29 derivados da 2-formilpiridina tiosemicarbazonas com atividade inibitória frente à ribonucleotídeo redutase (IRNR) de células H.Ep.-2; série II contendo 37 derivados de tiossemicarbazonas substituídas na cadeia lateral e no anel aromático, com atividade frente à cruzaína do *Trypanosoma cruzi* e, Série III contendo 45 derivados de tiossemicarbazonas, sendo 37 da série II, citada anteriormente e, 16 derivados de isatinas. Todas as séries foram divididas em 3 séries de treinamento e, nas 3 correspondentes, de teste, respectivamente séries A, B e C.

A seguir, para cada uma das séries (I, II e III), a partir das estruturas das moléculas, representadas em 3 dimensões, foram gerados 1497 descritores através do programa DRAGON. Estes descritores foram submetidos a um pré-tratamento de dados, analisando-se e, excluindo-se aqueles que não contribuiriam para as análises *PLS – Partial Least Squares*, denominação em inglê para mínimos quadrados parciais.

Através das análises PLS foram selecionados os descritores mais significativos das equações de regressão linear, sendo que este número variou nas três. A seguir, estes foram submetidos à uma análise de freqüência de presença, ou seja, foram selecionados os descritores presentes em pelo menos dois dos três modelos gerados a partir das séries de treinamento A, B e C, respectivamente. A partir dos descritores selecionados foram gerados modelos de QSAR clássico com 5 descritores que foram submetidos e, aceitos por diversos filtros de validação de modelos.

Adicionalmente, com a sistematização destas metodologias nesta dissertação de mestrado, pretende-se aplicar e estender as abordagens aqui propostas para outros sistemas estudados e, de relevância no grupo.

IV.2. Escolha das séries

Como já detalhado na metodologia, nesta dissertação de mestrado foram selecionados da literatura original (French *et al.* 1974; Du *et al.* 2002; Chiyanzu et al. 2003; Greenbaum *et al.* 2004) 90 compostos, incluindo-se 29 α-(N)-heterocíclica carboxaldeído tiossemicarbazonas substituídas, com atividade inibitória frente a ribonucleotídeo redutase (IRNR) (French *et al.* 1974) (série I; compostos I.1. a I.29); 37 tiossemicarbazonas substituídas na cadeia lateral e no anel aromático(Du *et al.* 2002) (série II; compostos II.1 a II.37) e, 61 compostos estruturalmente diferentes (série III; compostos II.1 a III.61, sendo 45 tiossemicarbazonas (Greenbaum *et al.* 2004), das quais 37 da série II (Du *et al.* 2002) e, 16 derivados de isatinas (Chiyanzu *et al.* 2003). As séries II e III apresentam atividade inibitória frente à cruzaína, uma cisteína protease do *T.cruzi*. E a série I, embora, apresente atividade inibitória frente a ribonucleotídeo redutase (IRNR) de células H.Ep.-2, é, no entanto, estruturalmente similar às séries II e III. E, ainda mais, os modelos gerados nesta dissertação para a série I foram incluídos na tese de doutorado de H.Ishiki e, lá comparados com os modelos CoMFA e CoMSIA por ele propostos.

Considerando-se os critérios de seleção de séries; descritos no item metodologia (II.1.1. *Critérios de Seleção*), os dados utilizados neste trabalho, embora retirados da literatura, podem ser considerados homogêneos e, adequados para uma análise de QSAR, efetuada respectivamente para a série I e, para as séries II e III estudados em conjunto

Desta forma, pode-se assegurar que os dados analisados foram obtidos utilizando os mesmos protocolos experimentais e, por um único grupo de pesquisa, com tradição em ensaios nas respectivas atividades biológicas, ou seja, nas atividades inibitórias frente à ribonucleotídeo redutase (IRNR) de células H.Ep.-2, (erro especificado de ± 10%) e, frente à cruzaína.

Observa-se nas tabelas, respectivamente, tabelas II.1.2.1, II.1.3.1 e II.1.4.1 que nas três séries, as faixas de variação nos valores de pIC_{50} foram de mínimo 2 unidades Estes intervalos em atividade biológica são considerados suficientes para se gerar modelos de QSAR, válidos estatisticamente. Para as séries II e III, os compostos com valores de IC_{50} maiores ou iguais a 10 μM foram considerados inativos (DU *et al*., 2002), sendo que os erros não foram especificados.

Do ponto de vista estrutural, para a série I. ser considerada homogênea, os compostos foram selecionados de modo que as estruturas não apresentavam ramificações na cadeia lateral (tabela II.1.2.1.). Por outro lado, como pode ser verificado nas tabelas II.1.3.1 e II.1.4.1, para as séries II e III, os compostos selecionados apresentaram considerável diversidade estrutural, com o objetivo de se avaliar o efeito da cadeia lateral sobre a atividade inibitória frente à cruzaína.

IV.3. Pré-tratamento de Dados

Para as séries I, II e III, a partir das estruturas geradas, como descrito mo II.3. *Obtenção da Geometria das Moléculas*, foram obtidos 1497 parâmetros estruturais, utilizando o programa DRAGON.

O programa DRAGON foi utilizado, considerando-se que é descrito na literatura como uma ferramenta confiável (para a maioria dos cálculos), bastante rápido (com baixo custo computacional) e, gratuito (na época). Estes critérios estão de acordo com o recomendado em literatura recente (Martin 1998), na qual sugere-se que a etapa de obtenção de dados deva apresentar precisão e ser rápida. Salienta-se ainda, a importância de

haver um compromisso entre ambas, para que discrepâncias causadas pelas aproximações nos métodos de cálculo, possam ser apontadas e, minimizadas e/ou eliminadas e, consideradas com a devida cautela na análise final.

Considerando-se os 1497 descritores gerados, aplicou-se o pré-tratamento de dados, para excluir, como recomendado na literatura (Livingstone 1995), as variáveis que não forneceram informações relevantes sobre as atividades biológicas, sugerindo-se que estas contribuiriam apenas, para aumentar a quantidade de dados e de ruídos a serem tratados.

A partir dos resultados obtidos (apresentados nos itens III.1.1, III.2.1, III.3.1 e, III.4.1) verifica-se que o número de descritores que apresentou valores constantes ou, apenas um valor diferente na série, variou nas 3 séries estudadas.

Em uma primeira etapa, aplicou-se o pré-tratamento considerando-se a exclusão de descritores com valores constantes ou, com apenas um valor diferente na série. Observando-se os dados apresentados nas tabelas II.1.2.1, II.1.3.1 e II.1.4.1, pode-se verificar que quanto maior o número de compostos presentes em cada série e, em particular, quanto maior sua diversidade estrutural (avaliada visualmente) menor é o número de descritores que apresentam valores constantes ou, com apenas um valor diferente na série. Assim, nos resultados apresentados nos itens III.2.1 e III.4.1. pode-se observar que a redução de variáveis foi maior para a série II quando comparada à série III. Este fato pode ser explicado considerando-se que o número de compostos na série III é 35% maior do que o da série II e, a redução do número de descritores na série III foi cerca de 65% menor do que a da série II. O aumento da diversidade estrutural na série III em relação à série II pode ser observado, pois esta inclui além das tiossemicarbazonas da série II os derivados de isatina.

Aplicando-se o segundo critério para o pré-tratamento de dados, que exclui os descritores que apresentaram valores superiores a 4 desvios padrão (σ) em relação à média, verifica-se que o número desses descritores variou nas 3 séries. Verifica-se ainda, que a série III (item III.4.1) apresentou o maior número (183) de descritores com tais características enquanto que a série II (item III.3.1) apresentou o menor (74). Com exceção da série I (item III.1.1) que apresentou número intermediário desses descritores (90), verifica-se que os efeitos dos aumentos do número de compostos bem como de sua diversidade estrutural causam um aumento do número de descritores com valores

superiores a 4 (σ) em relação à média. Este fato pode ser explicado, considerando-se que em uma série, o aumento tanto do número de compostos bem como de sua diversidade diminui o número de descritores que apresentam valores constantes e aqueles com apenas um valor diferente. Como conseqüência, observa-se um aumento do número de descritores que podem apresentar valores maiores do que 4 (σ) em relação à média.

Para cada uma das séries, verifica-se que valores acima de 4(σ) indicam que estes são *"outliers"* (apresentado na Introdução, item I.6) para o descritor considerado. A retirada destes descritores se faz necessária, pois as análises de regressão linear paramétrica, aplicadas em seguida, se baseiam na distribuição normal dos dados.

Aplicando-se um terceiro critério, considerando-se a intercorrelação de descritores, verifica-se que o número de descritores que apresentam valores de intercorrelação maiores que 0,95 (r>0,95) aumenta com o tamanho e com a diversidade da série, como pode ser observado nos itens III.1.1, III.2.1, III.3.1, e III.4.1. Este fato pode ser explicado pelo maior número de descritores válidos remanescentes, que não foram previamente eliminados de acordo com o primeiro critério.

Adicionalmente, para a série II comparando-se os resultados do terceiro critério, sem e, com a retirada dos descritores 4(σ), apresentados nos ítens III.2.1 e III.3.1, respectivamente, observa-se que o número de descritores intercorrelacionados diminuiu de 335 para 328 (2%). Este fato pode ser explicado considerando-se que alguns descritores com valores acima de 4(σ) foram eliminados previamente pelo segundo critério, ou seja, apresentavam valores de intercorrelação maiores que 0,95 (r>0,95) com outros descritores e/ou entre si.

IV.4. Série I

IV.4.1. Obtenção de Modelos *PLS*

É conhecido na literatura (Topliss & Costello 1972) que a análise de regressão linear múltipla (*MLR*) aplicada à séries de dados, que apresentam um número muito maior de variáveis em relação ao de amostras (p>>n) aumenta as chances de se gerar um modelo falso que, no entanto, apresenta um bom ajuste. Deste modo, por este método, não se obtém

uma correlação "verdadeira", mas sim apenas uma correlação "ao acaso" (Topliss & Costello 1972; Topliss & Edwards 1979) sendo que estas chances aumentam quanto maior o número de variáveis em relação ao número de amostras.

Por outro lado, utilizando-se o método de análise *PLS* (***Partial Least Squares***) verifica-se que as correlações obtidas "ao acaso" mostram-se não significativas. E, ainda mais, constata-se na literatura (Clark & Cramer 1993) que para o método de análise por PLS, ao contrário, observa-se que o aumento do número de variáveis pode gerar modelos estatisticamente mais significativos, observando-se, ao mesmo tempo, uma diminuição dos valores dos coeficientes de predição (Q_{cv}^2) do modelo PLS (Clark & Cramer 1993). Estas características do *PLS*, que são intercorrelacionadas, podem ser explicadas considerando-se o procedimento do algoritmo utilizado (Geladi & Kowalski 1986). Assim, tanto no *PLS* (Geladi & Kowalski 1986; Wold *et al.* 2001) como no *PCA* (Wold *et al.* 1987) a operação fundamental é repetir a extração de componentes. Estas são geradas a partir de uma combinação linear de todas as variáveis consideradas, mantendo-se cada novo componente, ortogonal a quaisquer outro componente,. previamente extraído. Estão apresentados e discutidos na Introdução (Item I.8. *Data Mining*) aspectos teóricos de alguns métodos, descritos na literatura, visando a extração de informação de um grande número de variáveis, entre eles PLS, PCA e PCR.

Para a série I, os resultados correspondentes aos modelos PLS gerados estão apresentados na tabela III.1.2.3.

Cada um dos 3 modelos *PLS* foi selecionado, considerando-se o maior valor do coeficiente de predição interno Q_{cv}^2 obtido pelo método "*full cross-validation"* (Cramer *et al.* 1988; Wakeling & Morris 1993), como apresentado e discutido anteriormente (Introdução, item I.6).

Para as séries de treinamento IA, IB e IC, os valores observados para Q_{cv}^2 foram, respectivamente iguais a 0,698, 0,802 e, 0,771. Estes, sendo maiores do que 0,5 foram neste trabalho considerados significativos, de acordo com critério proposto na literatura (Golbraikh & Tropsha 2002; Tropsha *et al.* 2003; Clark & Cramer 1993).

Como pode ser observado na tabela III.1.2.3, todos os modelos *PLS* selecionados para as 3 séries de treinamento IA, IB e IC apresentam 3 variáveis latentes, embora o número de descritores originais utilizados nos modelos PLS selecionados tenha sido

diferente. Os modelos *PLS* selecionados para as séries de treinamento IA, IB e IC continham 25, 15 e 15 descritores originais, respectivamente. Assim, constata-se que o numero total de descritores selecionados foi igual a 55, porém, alguns destes estiveram presentes em mais de um modelo.

Como apresentado e discutido no item I.9.4, valores de Q_{cv}^2 maiores do que 0,5, é um requisito necessário, porém, não suficiente para garantir o poder de predição do modelo. Comparando-se os valores de Q_{cv}^2 para as séries de treinamento (tabela III.1.2.1), com os correspondentes valores de r^2 para as séries de teste (tabela III.1.2.2) verifica-se que não há correlação significativa entre eles, como já observado e discutido na literatura por Tropsha (Golbraikh & Tropsha 2002; Tropsha *et al.* 2003). Como conseqüência, para a série I, não se observa correlação significativa entre os valores dos coeficientes de predição interna Q_{cv}^2 (tabela III.1.2.1) e externa, Q^2 (tabela III.1.2.2.). Esta correlação não significativa pode ser observada, nitidamente, nas figuras III.1.2.1 e III.1.2.2.

Adicionalmente, como podem ser observados na tabela III.1.2.1, os valores de Q_{cv}^2 aumentam a medida que são retirados do modelo, os descritores estatisticamente menos significativos, ou seja, aqueles que menos contribuíram para explicar a variância dos valores de atividade biológica nos modelos *PLS*, exceto para os modelos com número de descritores menor do que 15. Para esses, a diminuição dos valores de (Q_{cv}^2) pode ser explicada pela retirada dos descritores relevantes para o modelo *PLS* , ou seja, aqueles que explicam a variância dos valores de atividade biológica.

Como podem ser observados na figura III.1.2.1, os valores de coeficiente de predição, Q^2, para a série de teste IA, são altos (~ 0,9) e, apresentam pequenas variações em função do número de variáveis presentes. Na figura III.1.2.2 observa-se que os valores do coeficiente de predição,Q^2, da série de teste IB, apresentam um comportamento inverso dos respectivos Q_{cv}^2, ou seja, diminuem com a retirada de descritores. Na figura III.1.2.3 observa-se que os valores do coeficiente de predição, Q^2, da série de teste IC, inicialmente diminuem e, depois aumentam com a retirada dos descritores menos significativos do modelo *PLS*. Ainda mais, observa-se que ao redor de 40 descritores verificam-se tendências similares para Q^2 e para os respectivos Q_{cv}^2 da série de treinamento IC.

Para os compostos da serie I com atividade inibitória frente a ribonucleotídeo redutase (IRNR), não foi possível aplicar o critério proposto por Tropsha (Golbraikh &

Tropsha 2002), pois o programa computacional ANALYZE não armazena os dados gerados de atividade preditos pelos modelos *PLS*.

IV.4.2. Análise dos Valores de Frequência

A análise da presença de um determinado descritor nos modelos PLS gerados para as três diferentes séries de treinamento, pode ser utilizada como uma indicação da relevância deste descritor para o sistema estudado, em especial para a atividade inibitória dos compostos da série I frente a RNR.

Os resultados das análises de frequência da presença de cada um dos 27 diferentes descritores para os três modelos PLS selecionados, estão apresentados, na tabela III.1.3.1.

Para a série I, considerando-se somente os descritores diferentes entre si, constata-se uma redução do número desses, ou seja, de 55 para 27 descritores (tabela III.1.3.1). Desse total, 21 deles foram selecionados para posterior análise de *QSAR*, levando-se em conta que esses estão presentes em mais de um modelo e, de acordo com critério por nós sugerido e, descrito na Metodologia (item II.5.3.3). A partir dos 21 descritores selecionados, que estão apresentados em negrito na tabela III.1.3.1, diversos modelos *QSAR* foram gerados e, a seguir validados, pelos diversos critérios discutidos nesta dissertação e, foram analisados como descrito no próximo item.

IV.4.3. Análise do Modelo de *QSAR* Clássico Gerado

As análises dos modelos QSAR gerados e, validados, permitem verificar e avaliar a natureza e a contribuição de cada um dos descritores anteriormente selecionados para a atividade inibitória, dos compostos da série I, frente a RNR.

O modelo *QSAR* foi selecionado, considerando-se o maior valor do coeficiente de predição interno (Q_{cv}^2) e, a seguir validado pelas regras, respectivamente *QUIK*, Q^2 Assintótico e, as baseadas nas funções R^P e R^N apresentadas e descritas na Introdução (itens I.9.1; I.9.2 e, I.9.3) No modelo QSAR selecionado, expresso pela equação III.1.4.1, observa-se uma variância explicada de 83,4 % com coeficiente de predição interno Q_{cv}^2 no valor igual a 0,758.

Adicionalmente, para as variáveis independentes, verifica-se na tabela III.1.4.1 um baixo valor para a entropia relativa (S_R = 0,2052) e, portanto, como apresentado na Introdução (item I.9.1), um alto índice de degeneração multivariada (D = 0,7948). Este alto valor para o índice de degeneração foi observado, apesar do modelo de QSAR selecionado ter sido, previamente, submetido e, validado pela regra QUIK. Sendo que, esta regra considera que o valor da correlação entre as variáveis independentes (Kx) é menor do que o observado para a correlação entre todas as variáveis (Kxy).

Na figura III.1.4.1, para a série I, a análise da matriz de correlação dos valores dos 5 descritores selecionados (CIC2; C-025; Mor07m-$_{REV}$; H1p-$_{VER}$ e, RDF065v-$_{REV}$) e da atividade biológica (pIC_{50}) apresenta um valor baixo, estatisticamente não significativo, para a correlação entre os descritores (r<0,45). Esse baixo valor era esperado, pois a pré-seleção de variáveis a partir da análise *PLS* aliada à aplicação da regra QUIK e da regra baseada na função R^P, foram aplicadas visando prevenir altos valores de correlações entre as variáveis selecionadas no modelo de *QSAR* gerado.

Na figura III.1.4.2, a análise visual do gráfico de regressão entre os valores de atividade $pIC_{50,}$ preditos e observados mostrou ajuste significativo à reta de regressão. Esta constatação foi confirmada observando-se o perfil do gráfico dos valores de resíduos ($pIC_{50observado} - pIC_{50predito}$), das diferenças entre os valores de pIC_{50} observados e preditos pelo modelo de *QSAR*, expresso pela equação III.1.4.1. Neste gráfico os pontos se apresentaram distribuídos de forma aleatória.

Para série I, estão apresentados e discutidos, a seguir, os descritores presentes no modelo de QSAR expresso pela equação III.1.4.1 bem como suas contribuições relativas.

$$pIC_{50} = 0{,}175\ (\pm 0{,}10)\ CIC2 + 0{,}138\ (\pm 0{,}11)\ C\text{-}025 + 0{,}300\ (\pm 0{,}11)\ Mor07m\text{-}_{REV} + 0{,}144\ (\pm 0{,}10)\ H1p\text{-}_{REV} + 0{,}207\ (\pm 0{,}12)\ RDF065v\text{-}_{REV} + 5{,}791\ (\pm 0{,}13)$$

(n = 29; r = 0,913; s = 0,248; F = 23,105; Q_{cv}^2 = 0,758; s-PRESS = 0,299) Equação III.1.4.1.

Na equação III.1.4.1., o termo Mor07m-$_{REV}$ apresenta a maior contribuição relativa (0,300) e, positiva para o modelo QSAR.

O descritor Mor07m é um dos descritores código 3D-MoRSE (*Molecule Representation of Structure based on Electron diffraction*) que representa algumas

características estruturais do composto, como por exemplo, massa e, o número de ramificações, como apresentado e descrito na Introdução (item I.4.4). O código 3D-MoRSE leva em consideração, nos cálculos, a estrutura da molécula em três dimensões.

A análise do modelo QSAR expresso pela equação III.1.4.1 indica que quanto menor o valor atribuído ao descritor Mor07m maior será o valor predito para pIC_{50}. Ainda mais, para o modelo QSAR gerado para a série I, observando-se os valores apresentados na tabela III.1.4.1 e figura III.1.4.1, verifica-se que o descritor Mor07m mostrou uma pequena contribuição (15,62%) para o valor de índice de degeneração multivariada e, um valor estatisticamente significativo (r=0,628) para a intercorrelação com pIC_{50}.

Concluindo, o descritor Mor07m mostra, portanto, uma contribuição estatisticamente significativa 0,300 (±0,11) para o modelo de *QSAR* validado para a série I.

Para a série I, o modelo QSAR expresso pela equação III.1.4.1, mostra uma menor contribuição relativa para CIC_2 (0,175).

Na equação III.1.4.1, o descritor CIC_2 é um dos descritores topológicos, como apresentado na Introdução (item I.4.7). Este pode ser definido como o índice de informação complementar para 2ª ordem de vizinhança dos vértices (Basak *et al.* 1997) sendo expresso pela equação I.4.7.2.(item I.4.7). A análise desta equação indica que quanto maior o valor de IC_k, menor será o valor de CIC_k. Ainda mais, o termo IC_k , na equação I.4.7.1, indica que quanto maior a diversidade entre os vértices de mesma ordem k, maior será o valor do índice de informação das moléculas. Desta forma, quanto maior a diversidade entre os vértices de mesma ordem k, menor será o valor de CIC_k.

Para o modelo QSAR gerado para a série I, observando-se os valores apresentados na tabela III.1.4.1 e figura III.1.4.1, verifica-se que o descritor CIC_2 mostrou uma grande contribuição (25,05%) para o valor de índice de degeneração multivariada e, um valor estatisticamente não significativo (r=0,141) para a intercorrelação com pIC_{50}.

Concluindo, o descritor CIC2 mostra, portanto, uma contribuição pequena e, no limite de sua significância estatística, (0,175 ±0,10) para o modelo de *QSAR* validado para a série I.

Para a série I, o modelo QSAR expresso pela equação III.1.4.1, mostra uma contribuição relativa para C-025 igual a 0,138.

Na equação III.1.4.1, o descritor C-025 é um dos descritores de fragmentos de átomo centrado, como apresentado na Introdução (item I.4.10). Este pode ser definido como a representação de um átomo de carbono (C) do anel aromático, ligado a outros dois átomos de carbono deste anel bem como a um outro átomo de carbono qualquer (Ghose *et al.* 1989).

Para a série I, a análise do modelo QSAR expresso pela equação III.1.4.1 indica que quanto maior o valor de C-025 maior será o valor predito para pIC_{50}. Este fato sugere, portanto, que se incluindo o fragmento C-025 verifica-se um aumento nos valores de pIC_{50}.

A análise da tabela II.1.2.1, mostra que apenas os compostos 1 a 4 e, 14 apresentam um fragmento C-025, respectivamente grupos alquila (3-CH_3; 4-CH_3; 5-CH_3 e 5-C_2H_5) e trifluorometila (5-CF_3). Observa-se ainda na tabela II.1.2.1, que os compostos 1 a 4 apresentam altos valores de pIC_{50}, respectivamente 6,59; 6,57; 6,51 e, 6,66, enquanto que o composto 14 mostra um valor mais baixo, igual a 5,62. Assim, na série I, os compostos que apresentam grupos metila ou etila nas posições 3- ou 4- ou 5- do anel piridínico estão entre os mais ativos.

Para o modelo QSAR gerado para a série I, observando-se os valores apresentados na tabela III.1.4.1 e na figura III.1.4.1, verifica-se que o descritor C-025 mostrou uma contribuição significativa e, grande (23,37%), para o valor de índice de degeneração multivariada e, um valor estatisticamente significativo (r=0,527) para a intercorrelação com pIC_{50}.

Concluindo, o descritor C-025 mostra, portanto, uma contribuição pequena e, no limite de sua significância estatística (0,138 ±0,11) para o modelo de *QSAR* validado para a série I.

Para a série I, o modelo QSAR expresso pela equação III.1.4.1, mostra uma contribuição relativa para H1p igual a 0,144.

Na equação III.1.4.1, o descritor H1p é um dos descritores *GETAWAY*, como apresentado na Introdução (item I.4.1). Este pode ser definido como sendo o valor da autocorrelação entre os átomos de distância topológica 1, ponderados pela polarizabilidade. E, é expresso pela equação I.4.1.2. (item 1.4.1.) que indica que quanto maior o valor de acessibilidade dos átomos e, considerando-se a polarizabilidade entre eles, maior será o valor de H1p.

Para a série I, a análise do modelo QSAR expresso pela equação III.1.4.1 indica que quanto maior o valor de H1p, maior será valor predito para pIC_{50} sugerindo-se, portanto, que quanto menor for a acessibilidade dos átomos e maior for a diferença de polarizabilidade dos átomos vizinhos das moléculas da série, maior será o valor de $pIC_{50.}$

Para o modelo QSAR gerado para a série I, observando-se os valores apresentados na tabela III.1.4.1 e na figura III.1.4.1, verifica-se que o descritor H1p mostrou uma grande contribuição (22,09%) para o valor de índice de degeneração multivariada e, um valor estatisticamente não significativo (|r|=0,274) para a intercorrelação com pIC_{50}.

Concluindo, o descritor H1p mostra, portanto, uma contribuição pequena e, no limite de sua significância estatística, (0,144 ± 0,10) para o modelo de *QSAR* validado para a série I.

Para a série I, o modelo QSAR expresso pela equação III.1.4.1, mostra uma contribuição relativa para RDF065v igual a 0,207.

Na equação III.1.4.1, o descritor RDF065v é um dos descritores da classe *RDF* como apresentado na Introdução (item I.4.3). Este pode ser definido, como sendo o valor associado à probabilidade de se encontrar átomos à uma determinada distância fixa. E, esta no descritor RDF065v, é igual à 6,5 Å. (Hemmer *et al.* 1999). Valores altos de RDF065v são verificados para moléculas que apresentam átomos à uma distância de 6,5 Å, considerando-se também os valores dos volumes de van der Waals dos respectivos átomos.

Para a série I, a análise do modelo QSAR expresso pela equação III.1.4.1 indica que quanto maior o valor de RDF065v, menor será o valor predito para pIC_{50} sugerindo-se, que a presença de fragmentos de cadeia longa ligados ao anel piridínico, diminui o valor de pIC_{50}, como por exemplo, o observado para o composto 24 (R=5-$OOCCH_2OC_2H_5$ e pIC_{50} =5,25) quando comparado com composto 3 (R=5-CH_3 e pIC_{50} =6,51).

Para o modelo QSAR gerado para a série I, observando-se os valores apresentados na tabela III.1.4.1 e na figura III.1.4.1, verifica-se que o descritor RDF065v mostrou uma pequena contribuição (13,88%) para o valor de índice de degeneração multivariada e, um valor alto e, estatisticamente significativo (|r|=0,649) para a intercorrelação com pIC_{50}.

O descritor RDF065v mostra, portanto, uma contribuição estatisticamente significativa (0,207 ±0,12) para o modelo de *QSAR* validado para a série I.

Finalizando, pode-se observar ainda na tabela III.1.4.1 e na figura III.1.4.1, que os descritores RDF065v juntamente com o Mor07m mostraram as menores contribuições para o valor de índice de degeneração multivariada, sendo iguais a 13,88% e 15,62%, respectivamente. Observando-se, ainda, que os valores da intercorrelação com pIC_{50} foram significativos e, iguais a 0,649 e 0,628, respectivamente.

Como critério adicional para avaliar a qualidade do modelo de QSAR gerado pode-se incluir a análise dos valores das diferenças entre os valores de pIC_{50} observados e preditos pelo modelo. Para série I este valor foi de no máximo |0,68| para o composto 5, como pode ser observado na tabela III.1.4.4.

IV.5. Série II – Não retirando as variáveis 4 sigma (4σ)

IV.5.1. Modelos *PLS*

Analogamente ao discutido para série I, para a série II, torna-se necessário utilizar modelos PLS para selecionar as variáveis mais significativas (que serão utilizadas nos modelos de QSAR clássico) considerando-se que o número de variáveis é muito maior que o de amostras (n>>p).

Para a série II, os resultados correspondentes aos modelos PLS obtidos estão apresentados na tabela III.2.2.3.

Cada um dos 3 modelos *PLS* foi selecionado, considerando-se o maior valor do coeficiente de predição interno Q_{cv}^2 obtido pelo método "*full cross-validation*" (Cramer *et al.* 1988; Wakeling & Morris 1993), como apresentado e discutido anteriormente (Introdução, item I.6).

Para as séries de treinamento IIA, IIB e IIC, os valores observados para Q_{cv}^2 foram, respectivamente iguais a 0,925, 0,919, e 0,891. Estes, sendo maiores do que 0,5 foram considerados significativos, de acordo com critério proposto na literatura (Golbraikh & Tropsha 2002; Tropsha *et al.* 2003; Clark & Cramer 1993).

Como observado na tabela III.2.2.3, os modelos *PLS* selecionados para as séries de treinamento IIA, IIB e IIC apresentam 3, 3 e 2 variáveis latentes, respectivamente. Os modelos *PLS* selecionados para as séries de treinamento IIA, IIB e IIC continham 30, 15 e

35 descritores originais, respectivamente. Assim, constata-se que o número total de descritores selecionados foi igual a 80, porém, alguns destes estiveram presentes em mais de um modelo.

Como apresentado e discutido no item I.9.4, valores de Q_{cv}^2 maiores do que 0,5, é um requisito necessário, porém não suficiente para garantir o poder de predição do modelo. Comparando-se os valores de Q_{cv}^2 para as séries de treinamento (tabela III.2.2.1) com os correspondentes de r^2 para as séries de teste(tabela III.2.2.2), verifica-se que não há correlação estatisticamente significativa entre eles, como já observado e discutido na literatura por Tropsha (Golbraikh & Tropsha 2002; Tropsha *et al.* 2003). Como conseqüência, para a série II, não se observa correlação significativa entre os valores dos coeficientes de predição interna Q_{cv}^2 (tabela III.2.2.1) e externa, Q^2 (tabela III.2.2.2.). Esta correlação não significativa pode ser observada, nas figuras III.2.2.1, III.2.2.2 e III.2.2.3.

Como podem ser observados na tabela III.2.2.1, os valores de Q_{cv}^2 aumentam a medida que são retirados os descritores estatisticamente menos significativos, ou seja, aqueles que menos contribuíram para explicar a variância dos valores de atividade biológica nos modelos *PLS*, exceto para os modelos com número de descritores menor do que 15. Como verificado anteriormente, esta diminuição dos valores de (Q_{cv}^2) pode ser explicada pela retirada de descritores relevantes para o modelo *PLS* , ou seja, aqueles que explicam a variância dos valores de atividade biológica.

Como podem ser observados na figura III.2.2.1, os valores do coeficiente de predição, Q^2, para a série de teste IIA, aumentam com a retirada dos descritores menos significativos do modelo *PLS*. Ainda mais, observam-se, ao redor de 100 descritores, grandes variações. Na figura III.2.2.2 observa-se que os valores do coeficiente de predição, Q^2, da série de teste IIB aumentam com a retirada dos descritores menos significativos do modelo *PLS.* Ainda mais, observa-se que, ao redor de 50 descritores, os valores apresentam grandes variações.Na figura III.1.2.3 observa-se que os valores do coeficiente de predição, Q^2, da série de teste IIC, são altos (~ 0,8) e, apresentam pequenas variações. Ainda mais, observa-se que, ao redor de 50 descritores, os valores do coeficiente de predição, Q^2, apresentam grandes variações.

Na figura III.2.2.4, observando-se os gráficos da correlação entre os valores de pIC_{50} observados e preditos pelos modelos *PLS* para as séries IIA, IIB e IIC verifica-se o ajuste dos pontos à reta de regressão, em todos os três modelos selecionados.

Esta constatação pode ser confirmada, observando-se o perfil do gráfico dos valores de resíduos das diferenças ($pIC_{50observado} - pIC_{50predito}$),entre os valores de pIC_{50} observados e preditos pelos modelos PLS, para as séries de treinamento IIA, IIB e IIC (figura III.2.2.4), pois os pontos estão distribuídos de forma aleatória.

Aplicando-se os critérios de validação de modelos propostos por Tropsha, apresentados e discutidos na introdução (item I.9.4) (Tropsha *et al.* 2003; Golbraikh & Tropsha 2002) para as 3 séries de teste IIA, IIB, e IIC, (figura III.2.2.5 e tabela III.2.2.4), verifica-se que todos os modelos PLS foram validados, exceto o modelo PLS para a série IIB. Este fato pode ser explicado considerando-se o valor observado para o coeficiente de regressão externo, r^2. Este foi igual a 0,575, sendo portanto menor que 0,6.

IV.5.2. Análise dos Valores de Frequência

A análise da presença de um determinado descritor, nos modelos PLS gerados para as três diferentes séries de treinamento, pode ser utilizada como uma indicação da relevância deste descritor para o sistema estudado, em especial para a atividade inibitória dos compostos da série II frente à cruzaína.

Para a série II, os resultados das análises de frequência da presença de cada um dos 48 diferentes descritores para os três modelos PLS selecionados, estão apresentados na tabela III.2.3.1.

Considerando-se somente os descritores diferentes entre si, para a série II constata-se uma redução do número dos mesmos, ou seja, há a redução de 80 para 48 descritores (tabela III.2.3.1). Desse total, 19 deles foram selecionados para posterior análise de *QSAR*, levando-se em conta que esses estão presentes em mais de um modelo e, de acordo com critério por nós sugerido como descrito na Metodologia (item II.5.3.3). A partir dos 19 descritores selecionados, que estão apresentados em negrito na tabela III.2.3.1, vários modelos *QSAR* foram gerados e, a seguir validados, pelos diversos critérios discutidos nesta dissertação e, analisados como descrito no próximo item.

IV.5.3. Análise do Modelo de *QSAR* Clássico Gerado

As análises dos modelos QSAR gerados e validados permitem verificar e avaliar a natureza e, a contribuição de cada um dos descritores anteriormente selecionados, para os valores de atividade inibitória dos compostos da série II frente à cruzaina.

O modelo *QSAR* foi selecionado, considerando-se o maior valor do coeficiente de predição interno (Q_{cv}^2) e, a seguir validado pelas regras, respectivamente *QUIK*, Q^2 Assintótico e, as baseadas nas funções R^P e R^N, descritas na Introdução (itens I.9.1; I.9.2 e, I.9.3). No modelo QSAR selecionado, expresso pela equação III.2.4.1, observa-se uma variância explicada de 81,3 % com coeficiente de predição interno Q_{cv}^2 igual a 0,743.

Adicionalmente, para as variáveis independentes (descritores), verifica-se na tabela III.2.4.1 um baixo valor para a entropia relativa (S_R = 0,453) e, portanto um alto índice de degeneração multivariada (D = 0,547), como apresentadas e descrito na Introdução (item I.9.1). Este alto valor para o índice de degeneração foi observado, apesar do modelo de QSAR selecionado ter sido previamente submetido e, validado pela regra QUIK. Nesta regra, considera-se que o valor da correlação entre as variáveis independentes (Kx) tem de ser menor do que o observado para a correlação entre todas as variáveis (Kxy).

Na figura III.2.4.1, para a série II, a análise da matriz de correlação dos descritores selecionados (GATS6e, PJI2, R6u, RDF125u, C-033) e, da atividade biológica (pIC_{50}) apresenta um valor baixo, estatisticamente não significativo, para a correlação entre os valores dos descritores ($r<0,30$). Esse baixo valor era esperado, pois a pré-seleção de variáveis a partir da análise *PLS* aliada à aplicação da regra QUIK e da regra baseada na função R^P, previnem altos valores de correlações entre as variáveis selecionadas para o modelo de *QSAR* gerado.

Na figura III.2.4.2, a análise visual do gráfico de regressão entre os valores de atividade $pIC_{50,}$ preditos e observados, mostrou ajuste significativo à reta de regressão. Esta observação foi confirmada pelo perfil obtido no gráfico de resíduos entre os valores de pIC_{50} (observados e preditos pelo modelo de *QSAR*, expresso pela equação III.2.4.1. Neste gráfico os pontos se apresentaram distribuídos de forma aleatória.

A seguir, para série II, estão apresentados e discutidos os descritores presentes e, suas contribuições relativas, para o modelo QSAR expresso pela equação III.2.4.1

pIC_{50} = - 0,919 (±0,44) GATS6e - 2,516 (±1,20) PJI2 + 1,095 (±0,42) R6u
- 0,335 (±0,14) RDF125u - 1,283 (±0,48) C-033 + 8,866 (±1,18)

(n = 37; r = 0,902; s = 0,316; F = 26,990; Q_{cv}^2 = 0,743; s-PRESS = 0,370) Equação III.2.4.1

Na equação III.2.4.1., o termo PJI2 apresenta a maior contribuição relativa (-2,516) para o modelo *QSAR*, que para este modelo é negativa.

Na equação III.2.4.1, o descritor PJI_2 (*2D Petitjean shape index*) é um dos descritores topológicos, como apresentado na Introdução (item I.4.7). Este pode ser definido como uma medida do quanto a estrutura da molécula é cíclica ou acíclica, sendo expresso pela equação I.4.7.4 (item I.4.7). A análise desta equação indica que quanto maior o valor de PIJ_2, maior será o caráter acíclico da "forma" da molécula (Petitjean 1992).

Para o modelo QSAR gerado para a série II, observando-se os valores apresentados na tabela III.2.4.1 e na figura III.2.4.1, verifica-se que o descritor PIJ_2 mostrou uma grande contribuição (30,39%) para o valor de índice de degeneração multivariada e, um valor estatisticamente significativo (|r|=0,470) para a intercorrelação com pIC_{50}.

Concluindo, o descritor PIJ_2 mostra, portanto, uma contribuição significativa (-2,516 ±1,20) para o modelo de *QSAR* validado para a série II.

Para a série II, o modelo QSAR expresso pela equação III.2.4.1, o termo GATS6e apresenta uma menor contribuição relativa (- 0,919) e, negativa.

O descritor GATS6e é um dos descritores de autocorrelação como apresentado na Introdução (item I.4.1.6) (Geary 1954). Este pode ser definido como o valor da diferença entre as eletronegatividades dos átomos, que estão separados por uma distância topológica 6, sendo expresso pela equação I.4.5.6. (item 1.4.5). A análise desta equação indica que quanto maior o valor de GATS6e, maior é a diferença de eletronegatividade entre os átomos separados, por uma distância topológica 6.

Para a série II, a análise do modelo QSAR expresso pela equação III.2.4.1 indica que quanto maior o valor de GATS6e, menor será o valor predito para pIC_{50}, sugerindo-se portanto, que quanto menor a diferença entre os valores de eletronegatividade, dos átomos

que estão separados por uma distância topológica de valor 6, maior será o valor observado para $pIC_{50.}$

Para o modelo QSAR gerado para a série II, observando-se os valores apresentados na tabela III.2.4.1 e na figura III.2.4.1, verifica-se que o descritor GATS6e mostrou uma contribuição nula (0%) para o valor de índice de degeneração multivariada e, um valor estatisticamente não significativo (|r|=0,229) para a intercorrelação com pIC_{50}.

Concluindo, o descritor GATS6e mostra, portanto, uma contribuição significativa (-0,919 ±0,44) para o modelo de *QSAR* validado para a série II.

Para a série II, o modelo QSAR expresso pela equação III.2.4.1, o termo R6u apresenta a maior contribuição relativa (1,095) e, positiva.

Na equação III.2.4.1, o descritor R6u é um dos descritores *GETAWAY*, como apresentado na Introdução (item I.4.1). Este pode ser definido como sendo o valor da autocorrelação entre os átomos de distância topológica 6, considerando a distância geométrica entre estes átomos e, sua contribuição para a forma da molécula, sendo expresso pela equação I.4.1.5. (item 1.4.1.). A análise desta equação indica que quanto maior for o valor da contribuição dos átomos para o "formato" da molécula (maiores *leverages*) bem como, quanto menor for a distância geométrica entre estes átomos, maior será o valor observado para R6u.

Para a série II, a análise do modelo QSAR expresso pela equação III.2.4.1 indica que quanto maior o valor de R6u, maior será o valor predito para pIC_{50}, sugerindo-se, portanto, que com o aumento da "influência" (átomos posicionados em regiões mais periféricos com relação ao "centro" da molécula) dos átomos sobre a "forma" da molécula e diminuindo a distância geométrica dos átomos separados por uma distância topológica 6, maior será o valor observado para $pIC_{50.}$

Para o modelo QSAR gerado para a série II, observando-se os valores apresentados na tabela III.2.4.1 e figura III.2.4.1, verifica-se que o descritor R6u mostrou uma pequena contribuição (1,31 %) para o valor de índice de degeneração multivariada e, um valor estatisticamente significativo (r=0,542) para a intercorrelação com pIC_{50}.

Concluindo, o descritor R6u mostra, portanto, uma contribuição significativa (1,095 ± 0,42) para o modelo de *QSAR* validado para a série II.

Para a série II, o modelo QSAR expresso pela equação III.2.4.1, o termo RDF125u apresenta uma contribuição relativa igual a - 0,335.

Na equação III.2.4.1, o descritor RDF125u, é um dos descritores da classe *RDF* como apresentado na Introdução (item I.4.3). Este pode ser definido como sendo o valor associado à probabilidade de se encontrar átomos à uma determinada distância fixa, sendo esta no descritor RDF125u igual à 12,5 Å (Hemmer *et al.* 1999) Valores altos de RDF125u são verificados para moléculas que apresentem átomos à uma distância de 12,5 Å.

Para a série II, a análise do modelo QSAR expresso pela equação III.2.4.1 indica que quanto maior o valor de RDF125u, menor será o valor predito para pIC_{50}. Podendo-se, então, sugerir, que para compostos que apresentam cadeias longas na suas estruturas menores valores de pIC_{50}, serão observados. Como por exemplo, o verificado na tabela II.1.3.1. para os compostos da 3, 34 e 37, apresentando baixos valores de pIC_{50} iguais a 5,15; 5,80e 5,80, respectivamente

Para o modelo QSAR gerado para a série II, observando-se os valores apresentados na tabela III.2.4.1 e na figura III.2.4.1, verifica-se que o descritor RDF125u mostrou uma grande contribuição (28,55%) para o valor de índice de degeneração multivariada e, um valor estatisticamente não significativo ($|r|=0,460$) para a intercorrelação com pIC_{50}.

Concluindo, o descritor RDF125u mostra, no entanto, uma contribuição significativa (-0,335 ±0,14) e, negativa, para o modelo de *QSAR* validado para a série II.

Para a série II, o modelo QSAR expresso pela equação III.2.4.1, mostra uma contribuição relativa para C-033 igual a -1,283.

Na equação III.2.4.1, o descritor C-033 é um dos descritores de fragmentos de átomo centrado, como apresentado na Introdução (item I.4.10). Este pode ser definido como a representação de um átomo de carbono (C) do pirrol, ligado a dois outros átomos deste anel, respectivamente carbono e nitrogênio, bem como, a um átomo de hidrogênio qualquer (Ghose *et al.* 1989).

Para a série II, a análise do modelo QSAR expresso pela equação III.2.4.1 indica que quanto maior o valor de C-033, menor será o valor predito para pIC_{50}, sugerindo-se, portanto, que na série II, incluindo-se o fragmento C-033 observa-se uma diminuição no valor de pIC_{50}.

A análise da tabela II.1.3.1, mostra que na série II, apenas os compostos 6 e 8 apresentam um fragmento C-033. Observa-se ainda na tabela, que os valores de pIC_{50} para os compostos 6 e 8 são baixos e iguais a 5,55. (O valor do erro experimental da medida não foi especificado, na literatura original)

Para o modelo QSAR gerado para a série II, observando-se os valores apresentados na tabela III.2.4.1 e na figura III.2.4.1, verifica-se que o descritor, C-033, mostrou uma grande contribuição significativa (39,75%) para o valor de índice de degeneração multivariada e, um valor estatisticamente não significativo (|r| =0,339), para a intercorrelação com pIC_{50}.

Concluindo, o descritor C-033 mostra, no entanto, uma contribuição estatisticamente significativa (- 1,283 ± 0,48) para o modelo de *QSAR* validado para a série II.

Como critério adicional para avaliar a qualidade do modelo de QSAR gerado, pode-se incluir a análise dos valores das diferenças entre os valores de pIC_{50} observados e preditos pelo modelo. Para série II este valor foi de no máximo |0,79| para o composto 14, como pode ser observado na tabela III.2.4.4.

IV.6. Série II – Com a retirada das variáveis 4 sigma (4σ)

Sabe-se da literatura (Livingstone 1995) que a presença de valores de descritores acima de 4 desvios padrão, além da média (4σ) indica que estes são *"outliers"*, para o descritor considerado e, portanto retirados, como apresentado e discutido na Introdução e Metologia (itens I.8 e II.5.2) bem como no início deste capítulo de Discussão (pré-tratamento de dados).

IV.6.1. Modelos *PLS*, com a retirada das variáveis 4σ

Analogamente ao discutido para as séries I e, série II sem a retirada das variáveis 4**σ,** o número de variáveis é muito maior que o de amostras (n>>p).

Para a série II, com a retirada das variáveis 4**σ,** os resultados correspondentes aos modelos PLS obtidos estão apresentados na tabela III.3.2.3.

Cada um dos 3 modelos *PLS* foi selecionado, considerando-se o maior valor do coeficiente de predição interno Q_{cv}^2 obtido pelo método "*full cross-validation"* (Cramer *et al.* 1988; Wakeling & Morris 1993), como apresentado e discutido anteriormente (Introdução, item I.6).

Com a retirada das variáveis 4**σ,** para as séries de treinamento IIA, IIB e IIC, os valores observados para Q_{cv}^2 foram, respectivamente iguais a 0,871, 0,832, e 0,901. Estes, sendo maiores do que 0,5 foram considerados significativos, de acordo com critério proposto na literatura (Golbraikh & Tropsha 2002; Tropsha *et al.* 2003; Clark & Cramer 1993).

Com a retirada das variáveis 4σ, como observado na tabela III.3.2.3, os modelos *PLS* selecionados para as séries de treinamento IIA, IIB e IIC apresentam respectivamente 3, 2 e 4 variáveis latentes. Os modelos *PLS* selecionados para as séries de treinamento IIA, IIB e IIC continham 10, 30 e 30 descritores originais, respectivamente. Assim, constata-se que o número total de descritores selecionados foi igual a 70, porém, alguns destes estiveram presentes em mais de um modelo.

Com a retirada das variáveis 4σ, como apresentado e discutido no item I.9.4, valores de Q_{cv}^2 maiores do que 0,5, é um requisito necessário, porém não suficiente para garantir o poder de predição do modelo. Comparando-se os valores de Q_{cv}^2 para as séries de treinamento (tabela III.3.2.1) com os correspondentes de r^2, para as séries de teste(tabela III.3.2.2), verifica-se que não há correlação significativa entre eles, como já observado e discutido na literatura por Tropsha (Golbraikh & Tropsha 2002; Tropsha *et al.* 2003). Como conseqüência, para a série II, não se observa correlação significativa entre os valores dos coeficientes de predição interna Q_{cv}^2 (tabela III.2.2.1) e externa, Q^2 (tabela III.2.2.2.). Estas correlações não significativas podem ser observada, nas figuras III.3.2.1, III.3.2.2 e III.3.2.3.

Como podem ser observados na tabela III.3.2.1, os valores de Q_{cv}^2 aumentam a medida que são retirados os descritores estatisticamente menos significativos, ou seja, aqueles que menos contribuíram para explicar a variância dos valores de atividade biológica nos modelos *PLS,* exceto para os modelos com número de descritores menor do que 15. Como verificado e discutido anteriormente, esta diminuição dos valores de (Q_{cv}^2)

pode ser explicada pela retirada de descritores relevantes para o modelo *PLS* , ou seja, aqueles que explicam a variância dos valores de atividade biológica.

Com a retirada das variáveis 4σ, como podem ser observados na figura III.3.2.1, os valores de coeficiente de predição, Q^2, para a série de teste IIA, são ~ 0,6 e, apresentam pequenas variações, em função do número de descritores. Ainda mais, observa-se que ao redor de 25 descritores, os valores de Q^2 diminuem sensivelmente. Na figura III.3.2.2 observa-se que os valores do coeficiente de predição,Q^2, da série de teste IB, são ~ 0,55 e, apresentam pequenas variações. Ainda mais, observa-se que ao redor de 15 descritores, os valores de Q2 diminuem sensivelmente.. Na figura III.3.2.3 observa-se que os valores do coeficiente de predição,Q^2, da série de teste IIC, apresentam um comportamento inverso dos respectivos Q_{cv}^2, diminuindo com a retirada de descritores.

Na figura III.3.2.4, observando-se os gráficos da correlação entre os valores de pIC_{50} observados e preditos pelos modelos *PLS* para as séries IIA, IIB e IIC verifica-se o ajuste dos pontos à reta de regressão, em todos os três modelos selecionados, com a retirada das variáveis 4σ..

Esta constatação pode ser confirmada, observando-se o perfil do gráfico dos valores de resíduos das diferenças ($pIC_{50observado} - pIC_{50predito}$), entre os valores de pIC_{50} observados e preditos pelos modelos PLS, para as séries de treinamento IIA, IIB e IIC (figura III.3.2.4), pois os pontos estão distribuídos de forma aleatória.

Aplicando-se os critérios de validação de modelos propostos por Tropsha, apresentados e discutidos na introdução (item I.9.4) (Tropsha *et al.* 2003; Golbraikh & Tropsha 2002) para as 3 séries de teste IIA, IIB, e IIC, (figura III.3.2.5 e tabela III.3.2.4), e, com a retirada das variáveis 4σ, verifica-se que todos os modelos PLS foram validados, exceto o modelo PLS para a série IIA. Este fato pode ser explicado considerando-se o valor observado para o coeficiente de regressão externo, r^2. Este foi igual a 0,254, sendo portanto menor que 0,6.

IV.6.2. Análise dos Valores de Freqüência, com a retirada das variáveis 4σ

Com a retirada das variáveis 4σ, a análise da presença de um determinado descritor nos modelos PLS gerados para as três diferentes séries de treinamento, pode ser utilizada

como uma indicação da relevância deste descritor para o sistema estudado, em especial para a atividade inibitória dos compostos da série II frente à cruzaína.

Para a série II, com a retirada das variáveis 4σ, os resultados das análises de frequência da presença de cada um dos 45 diferentes descritores para os três modelos PLS selecionados, estão apresentados na tabela III.3.3.1.

Para a série II, com a retirada das variáveis 4σ, considerando-se somente os descritores diferentes entre si, constata-se uma redução do número dos mesmos, ou seja, de 70 para 45 descritores (tabela III.3.3.1). Desse total, 23 deles foram selecionados, para posterior análise de *QSAR*, considerando-se que esses estão presentes em mais de um modelo. E, ainda mais, por estar de acordo com critério por nós sugerido e, descrito na Metodologia (item II.5.3.3). A partir dos 23 descritores selecionados, que estão apresentados em negrito na tabela III.3.3.1, vários modelos *QSAR* foram gerados e, a seguir validados, pelos diversos critérios discutidos nesta dissertação e, analisados como apresentado a seguir.

IV.6.3. Análise do Modelo de QSAR Clássico Gerado, com a retirada das variáveis 4σ

As análises dos modelos QSAR aqui gerados e validados, com a retirada das variáveis 4σ, permitem verificar e avaliar a natureza e a contribuição de cada um dos descritores anteriormente selecionados para a atividade inibitória dos compostos da série II frente à cruzaina.

O modelo *QSAR* foi selecionado, considerando-se o maior valor do coeficiente de predição interno (Q_{cv}^2) e, a seguir, com a retirada das variáveis 4σ, validado pelas regras, respectivamente *QUIK*, Q^2 Assintótico e as baseadas nas funções R^P e R^N apresentadas descritas na Introdução (itens I.9.1; I.9.2 e, I.9.3) No modelo QSAR selecionado, expresso pela equação III.3.4.1, observa-se uma variância explicada de 82,6 % e, com coeficiente de predição interno Q_{cv}^2 no valor de 0,762.

Adicionalmente, para os descritores (variáveis independentes), verifica-se na tabela III.3.4.1 um alto valor para a entropia relativa (S_R = 0,583) e, portanto um baixo índice de degeneração multivariada (D = 0,417), como apresentadas e descrito na Introdução (item I.9.1). Este alto valor para o índice de degeneração foi observado, apesar do modelo de

QSAR selecionado ter sido, previamente submetido e, validado pela regra QUIK. Esta regra considera que o valor da correlação entre as variáveis independentes (Kx) é menor do que o observado para a correlação entre todas as variáveis (Kxy).

Na figura III.3.4.1, para a série II, com a retirada das variáveis 4σ, a análise da matriz de correlação dos descritores selecionados (nS, E1m, RDF020p, R7e, nROR) e da atividade biológica (pIC_{50}) apresenta um valor baixo, estatisticamente não significativo, para a correlação entre os valores dos descritores ($r<0,21$). Esse baixo valor era esperado, pois a pré-seleção de variáveis a partir da análise *PLS*, aliada à aplicação da regra QUIK e da regra baseada na função R^P, previnem altos valores de correlações entre as variáveis selecionadas no modelo de *QSAR* gerado.

Na figura III.3.4.2, a análise visual do gráfico de regressão entre os valores de atividade $pIC_{50,}$ preditos e observados, mostrou ajuste significativo à reta de regressão, com a retirada das variáveis 4σ. Esta observação foi confirmada pelo perfil obtido no gráfico de resíduos entre os valores de pIC_{50} (observados e preditos pelo modelo de *QSAR*, expresso pela equação III.3.4.1. Neste gráfico os pontos se apresentaram distribuídos de forma aleatória.

A seguir, para série II e, com ,a retirada das variáveis 4σ, estão apresentados e discutidos os descritores presentes e suas contribuições relativas para o modelo QSAR, expresso pela equação III.3.4.1

$$pIC_{50} = -0,994\ (\pm0,34)\ nS + 0,674\ (\pm0,31)\ E1m - 0,736\ (\pm0,26)\ RDF020p + 1,469\ (\pm0,43)\ R7e + 0,433\ (\pm0,35)\ nROR + 7,366\ (\pm0,76) \qquad \text{Equação III.3.4.1}$$

$(n = 37;\ r = 0,909;\ s = 0,304;\ F = 29,558;\ Q_{cv}^2 = 0,762;\ s\text{-PRESS} = 0,357)$

Na equação III.3.4.1., o termo R7e apresenta a maior contribuição relativa (1,469) e, positiva para o modelo *QSAR*.

Na equação III.3.4.1, o descritor R7e é um dos descritores *GETAWAY*, como apresentado na Introdução (item I.4.1). Este pode ser definido como sendo o valor de autocorrelação entre os átomos de distância topológica 7. Este descritor, expresso pela equação I.4.1.5. (item 1.4.1.), considera, adicionalmente, os seguintes termos: distância geométrica entre estes átomos, contribuição para a forma da molécula e, respectivos valores de eletronegatividade. A análise da equação I.4.1.5 indica que quanto maior for o valor da

contribuição dos átomos para a "forma" da molécula (maiores *leverages*) bem como para o valor da eletronegatividade e, quanto menor for a distância geométrica entre estes átomos, maior será o valor de R7e.

Para a série II, com a retirada das variáveis 4σ, a análise do modelo QSAR expresso pela equação III.3.4.1 indica que quanto maior o valor de R7e, maior será o valor predito para pIC_{50}, sugerindo-se, portanto, que com o aumento da "influência" (átomos posicionados em regiões mais periféricos com relação ao "centro" da molécula) dos átomos sobre a "forma" da molécula e diminuindo a distância geométrica dos átomos separados por uma distância topológica 7 e, aumentando o valor de eletronegatividade, maior será o valor observado para $pIC_{50.}$

Para o modelo QSAR gerado para a série II, com a retirada das variáveis 4σ, observando-se os valores apresentados na tabela III.2.4.1 e na figura III.2.4.1, verifica-se que o descritor R7e mostrou uma pequena contribuição (0,57%) para o valor de índice de degeneração multivariada e, um valor estatisticamente significativo (r=0,545) para a intercorrelação com pIC_{50}.

Concluindo, o descritor R7e mostra, portanto, uma contribuição significativa (1,469 ±0,43) para o modelo de *QSAR* validado para a série II. e, com a retirada das variáveis 4σ.

Na equação III.3.4.1., o termo nS apresenta uma contribuição relativa no valor de -0,994 a para o modelo *QSAR.*

Na equação III.3.4.1, o descritor nS é um dos descritores constitucionais, como apresentado na Introdução (item I.4.11). Este pode ser definido como sendo o número de átomos de enxofre presentes na molécula.

Para a série II e, com a retirada das variáveis 4σ, a análise do modelo *QSAR* expresso pela equação III.3.4.1 indica que quanto maior o valor de nS, menor será o valor predito para pIC_{50}, sugerindo-se, portanto, que para os compostos da série II que possuem mais de um átomo de enxofre, uma diminuição no valor de pIC_{50} é verificada. Desta forma, na tabela II.1.3.1, verifica-se que todos os compostos da série II apresentam 1 átomo de enxofre, com exceção dos compostos 3, 6, 24 e 25 que apresentam 2 átomos de enxofre. Esses apresentam baixos valores de $pIC_{50,}$ respectivamente, 5,15, 5,55, 5,42, e 5,72.

Para o modelo QSAR gerado para a série II e, com a retirada das variáveis 4σ, observando-se os valores apresentados na tabela III.3.4.1 e na figura III.3.4.1, é importante

salientar que o descritor nS mostrou a maior contribuição significativa (49,71%) para o valor de índice de degeneração multivariada e, um valor estatisticamente significativo (|r| =0,540).

Concluindo, o descritor nS mostra, no entanto, uma contribuição estatisticamente significativa (-0,994 ± 0,34) para o modelo de *QSAR* validado para a série II.

Para a série I e, com a retirada das variáveis 4σ, no modelo QSAR expresso pela equação III.3.4.1, o termo E1m apresenta uma menor contribuição relativa (0,674) e, positiva

Na equação III.3.4.1, o descritor E1m é um dos descritores direcionais da classe WHIM, como apresentado na Introdução (item I.4.2). Este pode ser definido como a distribuição dos átomos considerando a sua massa atômica e o primeiro eixo principal da matrix T, sendo expresso pela equação I.4.2.6 (item 1.4.2.). A análise desta equação indica que quanto maior o valor de E1m, mais próximos os átomos estão, no primeiro eixo principal.

Para a série II e, com a retirada das variáveis 4σ, a análise do modelo QSAR expresso pela equação III.3.4.1 indica que quanto maior o valor de E1m, maior será o valor predito para pIC_{50} sugerindo-se, que os compostos com as cadeias maiores apresentam valores de pIC_{50} mais altos.

Para o modelo QSAR gerado para a série II e, com a retirada das variáveis 4σ, observando-se os valores apresentados na tabela III.3.4.1 e na figura III.3.4.1, verifica-se que o descritor E1m mostrou uma contribuição nula (0 %) para o valor de índice de degeneração multivariada e, um valor estatisticamente não significativo (r=0,262) para a intercorrelação com pIC_{50}.

Concluindo, o descritor E1m mostra, no entanto, uma contribuição significativa (0,674 ±0,31) para o modelo de *QSAR* validado para a série II.

Para a série II e, com a retirada das variáveis 4σ, no modelo QSAR expresso pela equação III.3.4.1, o termo RDF020p apresenta uma contribuição relativa no valor de -0,736.

Na equação III.2.4.1, o descritor RDF020p, é um dos descritores da classe *RDF* como apresentado na Introdução (item I.4.3). Este pode ser definido como sendo o valor associado à probabilidade de se encontrar átomos à uma determinada distância fixa, sendo esta no descritor RDF020p igual à 2,0 Å (Hemmer *et al.* 1999). Valores altos de RDF0020p

são verificados para moléculas que apresentem átomos à uma distância de 2,0 Å bem como altos, os respectivos valores de polarizabilidade.

Para a série II e, com a retirada das variáveis 4σ, a análise do modelo QSAR expresso pela equação III.3.4.1 indica que quanto maior o valor de RDF020p, menor será o valor predito para pIC_{50}, sugerindo-se, que os compostos que apresentem átomos com altos valores de polarizabilidade e, à uma distância de 2,0 Å devam apresentar menores valores de $pIC_{50.}$

De acordo com modelo de *QSAR* selecionado (equação III.3.4.1), quanto maior o valor de RDF020p, menor é o valor de pIC_{50}. O valor do coeficiente de intercorrelação de RDF020p com a variável dependente é um dos mais altos (|r|=0,417- figura III.3.4.1) dentre os descritores selecionados. A sua contribuição para o valor de índice de degeneração multivariada é nula (0% - tabela III.3.4.1.), portanto não houve nenhum valor repetido presente neste descritor para a série II, com a retirada das variáveis 4σ.

Concluindo, o descritor RDF020p mostra, portanto, uma contribuição significativa (-0,736 ± 0,26) para o modelo de *QSAR* validado para a série II.

Na equação III.3.4.1, o termo nROR apresenta uma contribuição relativa igual a 0,433, para o modelo *QSAR,* com a retirada das variáveis 4σ.

Na equação III.3.4.1, o descritor nROR é um dos descritores da classe de grupos funcionais, como apresentado na Introdução (item I.4.9). Este pode ser definido como sendo o número de grupamentos presentes na molécula, que contem os átomos COC da função éter.

Para a série II, com a retirada das variáveis 4σ, a análise do modelo *QSAR* expresso pela equação III.3.4.1 indica que quanto maior o valor de nROR, maior será o valor predito para pIC_{50}, sugerindo-se, portanto, que na série II a presença do fragmento nROR, aumenta o valor de pIC_{50}.

A análise da tabela II.1.3.1, mostra que os compostos 1, 2, 4, 5 e, 9 apresentam 1 fragmento nROR e, altos valores de $pIC_{50,}$ respectivamente, 6,52, 6,55, 6,85, e 6,55 e, 6,25.

Para o modelo QSAR gerado para a série II e, com a retirada das variáveis 4σ, observando-se os valores apresentados na tabela III.3.4.1 e na figura III.3.4.1, verifica-se que o descritor nROR mostrou uma grande contribuição significativa (49,71%) para o valor

de índice de degeneração multivariada e, um valor estatisticamente não significativo (r =0,063) para a intercorrelação com pIC_{50}.

Concluindo, o descritor nROR mostra, portanto, uma contribuição pequena e, no limite de sua significância estatistica, (0,433 ±0,35) para o modelo de *QSAR* validado para a série II, com a retirada das variáveis 4σ.

Como critério adicional para avaliar a qualidade do modelo de QSAR gerado, pode-se incluir a análise dos valores das diferenças entre os valores de pIC_{50} observados e preditos pelo modelo. Para série II e, com a retirada das variáveis 4σ, este valor foi de no máximo |0,84| para o composto 36, como pode ser observado na tabela III.3.4.4.

IV.7. Série III

IV.7.1. Modelos *PLS*

Analogamente ao discutido para as séries I e II, para a série III o número de variáveis é muito maior que o de amostras (n>>p).

Para a série III, os resultados correspondentes aos modelos PLS obtidos estão apresentados na tabela III.4.2.3.

Cada um dos 3 modelos *PLS* foi selecionado, considerando-se o maior valor do coeficiente de predição interno, Q_{cv}^2, obtido pelo método "*full cross-validation"* (Cramer *et al.* 1988; Wakeling & Morris 1993), como apresentado e discutido anteriormente (Introdução, item I.6).

Para as séries de treinamento IIIA, IIIB e IIIC, os valores observados para Q_{cv}^2 foram, respectivamente iguais a 0,918, 0,903, e 0,904. Estes, sendo maiores do que 0,5 foram considerados significativos, de acordo com critério proposto na literatura (Golbraikh & Tropsha 2002; Tropsha *et al.* 2003; Clark & Cramer 1993).

Como observado na tabela III.4.2.3, os modelos *PLS* selecionados para as séries de treinamento IIIA, IIIB e IIIC apresentam 4 variáveis latentes, embora o número de descritores originais utilizados nos modelos PLS selecionados tenha sido diferente. Os modelos *PLS* selecionados para as séries de treinamento IIIA, IIIB e IIIC continham 30, 25 e 45 descritores originais, respectivamente. Assim, constata-se que o número total de

descritores selecionados foi igual a 100, porém, alguns desses estiveram presentes em mais de um modelo.

Como apresentado e discutido no item I.9.4, valores de Q_{cv}^2 maiores do que 0,5, é um requisito necessário, porém não suficiente para garantir o poder de predição do modelo. Comparando-se os valores de Q_{cv}^2 , para as séries de treinamento (tabela III.4.2.1), com os correspondentes de r^2, para as séries de teste(tabela III.4.2.2), verifica-se que há correlação não-significativa entre eles, como já observado e discutido na literatura por Tropsha (Golbraikh & Tropsha 2002; Tropsha *et al.* 2003). Como conseqüência, para a série III, não se observa correlação significativa entre os valores dos coeficientes de predição interna (Q_{cv}^2) (tabela III.2.2.1) e, externa (Q^2) (tabela III.2.2.2.). Esta correlação estatisticamente não significativa pode ser observada, nas figuras III.4.2.1, III.4.2.2 e III.4.2.3.

Como podem ser observados na tabela III.4.2.1, os valores de Q_{cv}^2 aumentam a medida que são retirados os descritores menos significativos, ou seja, aqueles que menos contribuíram para explicar a variância dos valores de atividade biológica nos modelos *PLS*, exceto para os modelos com número de descritores menor do que 15. Esta diminuição dos valores de Q_{cv}^2 pode ser explicada, considerando-se a retirada de descritores relevantes para o modelo *PLS*, ou seja, daqueles que explicam a variância dos valores de atividade biológica.

Por outro lado, como podem ser observados na figura III.4.2.1, os valores de coeficiente de predição, Q^2, para a série de teste IIIA, são ~ 0,65 e, apresentam pequenas variações, em função do número de descritores. Ainda mais, observa-se que ao redor de 25 descritores, inicialmente, os valores de Q^2 aumentam e, verifica-se uma diminuição a medida que se retiram os descritores menos significativos. Na figura III.4.2.2 observa-se que os valores do coeficiente de predição, Q^2, da série de teste IIIB, são altos (~ 0,8) e, apresentam pequenas variações, em função do número de descritores. Na figura III.1.2.3 observa-se que os valores do coeficiente de predição, Q^2, da série de teste IIIC, são baixos (~ 0,45) e, apresentam pequenas variações, em função do número de descritores. Ainda mais, ao redor de 30 descritores, os valores de Q^2 diminuem com a retirada dos descritores menos significativos.

Na figura III.4.2.4, observando-se os gráficos da correlação entre os valores de pIC_{50} observados e preditos pelos modelos *PLS* para as séries IIIA, IIIB e IIIC verifica-se o

ajuste dos pontos à reta de regressão, em todos os três modelos selecionados. Esta constatação pode ser confirmada, observando-se o perfil do gráfico dos valores de resíduos das diferenças ($pIC_{50observado} - pIC_{50predito}$),entre os valores de pIC_{50} observados e preditos pelos modelos PLS, para as séries de treinamento IIIA, IIIB e IIIC (figura III.4.2.4), pois os pontos estão distribuídos de forma aleatória.

Aplicando-se os critérios de validação de modelos propostos por Tropsha, apresentados e discutidos na introdução (item I.9.4) (Tropsha *et al.* 2003; Golbraikh & Tropsha 2002) para as 3 séries de teste IIIA, IIIB, e IIIC, (figura III.4.2.5 e tabela III.4.2.4), verifica-se que todos os modelos PLS foram validados.

IV.7.2. Análise dos Valores de Frequência

Analogamente ao discutido para as séries I e II, a análise da presença de um determinado descritor nos modelos PLS gerados para as três diferentes séries de treinamento, pode ser utilizada como uma indicação da relevância deste descritor para o sistema estudado, em especial para a atividade inibitória dos compostos da série III frente a cruzaína.

Para a série III, os resultados das análises de frequência da presença de cada um dos 59 diferentes descritores para os três modelos PLS selecionados, estão apresentados na tabela III.4.3.1.

Para a série III considerando-se somente os descritores diferentes entre si, constata-se uma redução do número dos mesmos, ou seja, de 100 para 59 descritores (tabela III.4.3.1). Desse total, 25 deles foram selecionados, para posterior análise de *QSAR*, considerando-se que esses estão presentes em mais de um modelo. E, ainda mais, por estar de acordo com critério por nós sugerido e, descrito na Metodologia (item II.5.3.3). A partir dos 25 descritores selecionados, que estão apresentados em negrito na tabela III.4.3.1, vários modelos *QSAR* foram gerados e, a seguir validados, pelos diversos critérios discutidos nesta dissertação e, analisados como descrito a seguir.

IV.7.3. Análise do Modelo de QSAR Clássico Gerado

As análises dos modelos QSAR gerados e, validados, permitem verificar e, avaliar a natureza e a contribuição de cada um dos descritores, anteriormente selecionados, para a atividade inibitória dos compostos da série III frente à cruzaina.

O modelo *QSAR* foi selecionado, considerando-se o maior valor do coeficiente de predição interno (Q_{cv}^2) e, a seguir validado pelas regras, respectivamente *QUIK*, Q^2 Assintótico e as baseadas nas funções R^P e R^N apresentadas e, descritas na Introdução (itens I.9.1; I.9.2 e, I.9.3). No modelo QSAR selecionado, expresso pela equação III.4.4.1, observa-se uma variância explicada de 82,8 % e, com coeficiente de predição interno Q_{cv}^2 igual a 0,792.

Adicionalmente, para os descritores (variáveis independentes), verifica-se na tabela III.4.4.1 um alto valor para a entropia relativa (S_R = 0,502) e, portanto um baixo índice de degeneração multivariada (D = 0,498), como apresentadas e descrito na Introdução (item I.9.1). Este alto valor para o índice de degeneração foi observado, apesar do modelo de QSAR selecionado ter sido, previamente submetido e, validado, pela regra QUIK. Esta regra considera que o valor da correlação entre as variáveis independentes (Kx) deve ser menor do que o observado para a correlação entre todas as variáveis (Kxy).

Na figura III.4.4.1, para a série III, a análise da matriz de correlação dos descritores selecionados (C-035, BEHm3, MATS8m, MATS8v, nROR) e da atividade biológica (pIC_{50}) apresenta um valor baixo, estatisticamente não significativo, para a correlação entre os valores dos descritores ($r<0,35$). Esse baixo valor era esperado, pois a pré-seleção de variáveis a partir da análise *PLS*, aliada à aplicação da regra QUIK e, da regra baseada na função R^P, previnem altos valores de correlações entre as variáveis selecionadas no modelo de *QSAR* gerado.

Na figura III.4.4.2, a análise visual do gráfico de regressão entre os valores de atividade $pIC_{50,}$ preditos e observados, mostrou ajuste significativo à reta de regressão, com a retirada das variáveis 4σ. Esta observação foi confirmada pelo perfil obtido no gráfico de resíduos entre os valores de pIC_{50} (observados e preditos pelo modelo de *QSAR*, expresso pela equação III.4.4.1. Neste gráfico os pontos se apresentaram distribuídos de forma aleatória.

A seguir, para série III, estão apresentados e discutidos os descritores presentes e, suas contribuições relativas para o modelo QSAR, expresso pela equação III.4.4.1

pIC50 = - 1,287 (±0,28) C-035 - 2,335 (±0,75) BEHm3 - 18,66 (±6,67) MATS8m
+ 2,693 (±0,63) MATS8v + 1,193 (±0,48) nROR + 32,98 (±7,85)

(n = 61; r = 0,910; s = 0,433; F = 53,030; Q_{cv}^2 = 0,792; s-PRESS = 0,477) Equação III.4.4.1

Na equação III.4.4.1., o termo MATS8m apresenta a maior contribuição relativa para o modelo *QSAR*, sendo esta no valor de -18,66.

O descritor Mats8m é um dos descritores de autocorrelação, como apresentado na Introdução (item I.4.1.6) (Moran 1950). Este pode ser definido como sendo o valor das diferenças entre os valores de massa atômica, relativa aos átomos que estão separados a uma distância topológica 8, sendo expresso pela equação I.4.5.5. (item 1.4.5). A análise desta equação indica que quanto mais alto o valor de MATS8m, mais positivamente estão autocorrelacionados os valores de massa atômica, relativa aos átomos com distância topológica 8.

Para a série III, a análise do modelo QSAR expresso pela equação III.4.4.1 indica que quanto mais alto o valor de Mats8m, mais baixo será o valor predito para $pIC_{50.}$ Sugere-se, portanto, que quanto maior a diferença entre os valores de massa atômica relativa aos átomos com distância topológica 8, maior será o valor observado para pIC_{50}

Para o modelo QSAR gerado para a série III, observando-se os valores apresentados na tabela III.4.4.1 e na figura III.4.4.1, verifica-se que o descritor Mats8m mostrou uma pequena contribuição (6,36%) para o valor de índice de degeneração multivariada e, um valor estatisticamente não significativo (|r|=0,285) para a intercorrelação com pIC_{50}.

Concluindo, o descritor Mats8m mostra, no entanto, uma contribuição significativa (-18,66 ±6,67) para o modelo de *QSAR* validado para a série III.

Para a série III, no modelo QSAR expresso pela equação III.4.4.1, C-035 mostra uma contribuição relativa igual a -1,287.

Na equação III.4.4.1, o descritor C-035 é um dos descritores de fragmentos de átomo centrado, como apresentado na Introdução (item I.4.10). Este pode ser definido como um fragmento constituído por um átomo de carbono ligado simultaneamente a um

nitrogênio, a um oxigênio através de com uma dupla ligação e, a um outro átomo de carbono qualquer (Ghose *et al.* 1989)

Para a série III, a análise do modelo QSAR expresso pela equação III.4.4.1 indica que quanto maior for o valor de C-035, menor será o valor predito para pIC_{50}. Sugere-se, portanto, que na série III, a presença do fragmento C-035, diminui o valor de pIC_{50}.

A análise da tabela II.1.4.1, mostra que apenas os 16 derivados de isatinas (compostos 38 a 53) apresentam 1 fragmento C-035. Observa-se ainda na tabela II.1.4.1, que os compostos 38 a 53 apresentam baixos valores de pIC_{50}, sendo estes menores que 6,0.

Para o modelo QSAR gerado para a série III, observando-se os valores apresentados na tabela III.4.4.1 e na figura III.4.4.1, verifica-se que o descritor C-035 mostrou uma grande contribuição significativa (43,51%) para o valor de índice de degeneração multivariada e, um valor estatisticamente significativo (|r|=0,685) para a intercorrelação com pIC_{50}.

Concluindo, o descritor C-035 mostra, portanto, uma contribuição estatisticamente significativa (- 1,287 ±0,28), para o modelo de *QSAR* validado para a série III.

Para a série III, no modelo QSAR expresso pela equação III.4.4.1, o termo BEHm3 apresenta uma contribuição relativa no valor de -2,335.

O descritor BEHm3 é um dos descritores BCUT, como apresentado na Introdução (item I.4.1.4). Este pode ser definido como o terceiro autovalor obtido da matriz de conectividade, na qual os elementos diagonais contêm os valores de massa atômica.

Para a série III, a análise do modelo QSAR expresso pela equação III.4.4.1 indica que quanto mais alto o valor de BEHm3 mais baixo será o valor predito para $pIC_{50.}$

Para o modelo QSAR gerado para a série III, observando-se os valores apresentados na tabela III.4.4.1 e na figura III.4.4.1, verifica-se que o descritor BEHm3 mostrou uma pequena contribuição (1,40%) para o valor de índice de degeneração multivariada e, um valor estatisticamente não significativo (|r|=0,146) para a intercorrelação com pIC_{50}.

Concluindo, o descritor BEHm3 mostra, no entanto, uma contribuição significativa (-2,335 ±0,75) para o modelo de *QSAR* validado para a série III.

Para a série III, o modelo QSAR expresso pela equação III.4.4.1, mostra uma contribuição relativa para Mats8v no valor de 2,693.

O descritor Mats8v é um dos descritores de autocorrelação como apresentado na Introdução (item I.4.1.6) (Moran 1950). Este pode ser definido como o valor das diferenças entre os valores de volume de van der Waals, relativo aos átomos que estão separados a distância topológica 8, sendo expresso pela equação I.4.5.5. (item 1.4.5). A análise desta equação indica que quanto mais alto o valor de MATS8m, mais positivamente estão autocorrelacionados os valores do volume de van der Waals relativo aos átomos com distância topológica 8.

Para a série III, a análise do modelo QSAR expresso pela equação III.4.4.1 indica que quanto mais alto o valor de Mats8v mais alto será o valor predito para $pIC_{50.}$ Sugere-se, portanto, que quanto menor a diferença ente os valores de volume de van der Waals relativo aos átomos com distância topológica 8, maior será o valor observado para $pIC_{50.}$

Para o modelo QSAR gerado para a série III, observando-se os valores apresentados na tabela III.4.4.1 e na figura III.4.4.1, verifica-se que o descritor Mats8v mostrou uma pequena contribuição (1,12%) para o valor de índice de degeneração multivariada e, um valor estatisticamente significativo (r=0,411) para a intercorrelação com pIC_{50}.

Concluindo, o descritor Mats8v mostra, portanto, uma contribuição significativa (2,693 ±0,63) para o modelo de *QSAR* validado para a série III.

Na equação III.4.4.1, o termo nROR apresenta uma contribuição positiva relativa menor (1,193) para o modelo *QSAR*.

Na equação III.4.4.1, o descritor nROR é um dos descritores da classe de grupos funcionais, como apresentado na Introdução (item I.4.9).Este pode ser definido como sendo o número de grupamentos presentes na molécula, que contem os átomos COC da função éter. Para a série III, a análise do modelo *QSAR* expresso pela equação III.4.4.1 indica que quanto maior o valor de nROR, maior será o valor predito para pIC_{50}, sugerindo-se, portanto, que na série III a presença do fragmento nROR, aumenta o valor de pIC_{50}.

A análise das tabelas II.1.3.1 e II.1.4.1, mostra que os 8 compostos da série III, respectivamente 1, 2, 4, 5, 9, 52, 56 e, 57 apresentam 1 fragmento nROR. E, dentre estes os compostos 52, 56 e, 57 apresentam valores de atividade menores que 6,0 ($pIC_{50} < 6,0$) enquanto que os compostos 1, 2, 4, 5 e, 9 apresentam maiores valores de $pIC_{50,}$ respectivamente, 6,52, 6,55, 6,85, e 6,55 e, 6,25.

Para o modelo QSAR gerado para a série III, observando-se os valores apresentados na tabela III.4.4.1 e na figura III.4.4.1, verifica-se que o descritor nROR mostrou uma grande contribuição significativa (47,61%) para o valor de índice de degeneração multivariada e, um valor estatisticamente não significativo (r =0,202) para a intercorrelação com pIC_{50}.

Concluindo, o descritor nROR mostra, no entanto, uma contribuição estatisticamente significativa, (1,193 ±0,48), para o modelo de *QSAR* validado para a série III.

Como critério adicional para avaliar a qualidade do modelo de QSAR gerado, pode-se incluir a análise dos valores das diferenças entre os valores de pIC_{50} observados e preditos pelo modelo. Para série III, este valor foi de no máximo |1,04| para o composto 8, como pode ser observado na tabela III.4.4.4.