I. Introdução

I.1. Antichagásicos

I.1.1. Alguns Aspectos da Doença de Chagas

A doença de Chagas foi descoberta em 1909 por Carlos Chagas, sendo uma protozoose causada pelo hemoflagelado *Trypanossoma cruzi*, presente no continente americano principalmente entre os pequenos mamíferos selvagens. Os vetores deste protozoário são os insetos triatomíneos, sendo principalmente os *Triatoma infestans* e o *Triatoma dimidiata*, os quais são conhecidos como "barbeiro". A doença de Chagas é uma das principais causas das doenças do coração na América Latina (Du *et al.* 2002; Libow *et al.* 1991), sendo que 16 a 18 milhões de pessoas estão infectadas com *Trypanossoma cruzi* (Du et al. 2002), resultando em mais de 14000 mortes por ano e diminuindo consideravelmente a expectativa e a qualidade de vida dos infectados (figura I.1.1.1) (WHO 2004).

O *Trypanossoma cruzi* possui um ciclo de vida complexo, adotando diferentes formas no interior do hospedeiro vertebrado e do inseto vetor, sendo elas:

- Tripomastigotas - ocorrem corrente sanguínea dos vertebrados e no tubo digestivo do vetor, possuem mobilidade;
- Epimastigotas - forma de reprodução do parasito no vetor, possuem mobilidade;
- Amastigotas – constituem os estágios de multiplicação no hospedeiro vertebrado, são destituídas de mobilidade.

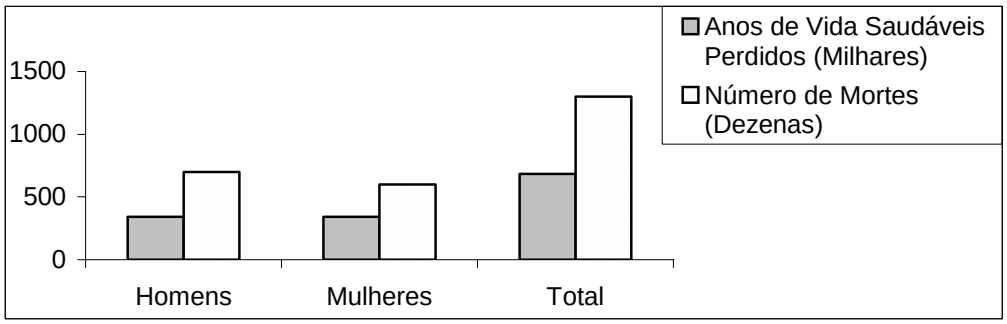
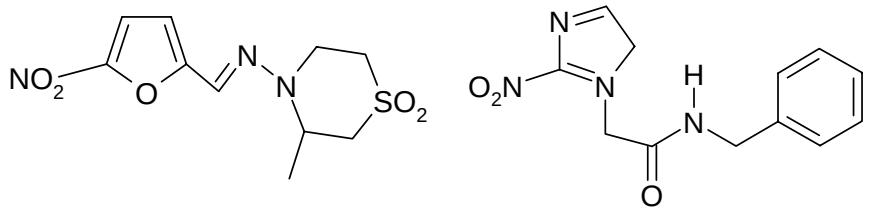


Figura I.1.1.1. Dados da OMS sobre a distribuição da doença de Chagas no mundo (WHO 2004).

I.1.2. Drogas Utilizadas

No final dos anos 60 e início dos anos 70 surgiram novas drogas para o tratamento da fase aguda da doença de Chagas. O "nifurtimox" (*Lampit*®) (Voigt *et al.* 1972) (Figura I.1.2.1.), um derivado nitrofurânico e, mais tarde o "benzonidazol" (*Rochagan*®) (Polak & Richle 1978) (Figura I.1.2.1.) um derivado imidazólico, sendo que este último apresenta IC_{50} igual a 19 μM, ou seja, a concentração necessária para eliminar 50 % da forma tripomastigota do parasita , *Trypanosoma cruz* que é em torno de 8 vezes menor que a observada para o nifurtimox (IC_{50} = 150 μM) (Melo *et al.* 2000).

Ambas as drogas não são eficientes principalmente para a fase crônica da doença e ocasionam diversos efeitos colaterais (Coura & de Castro 2002). Para o Nirfutimox os efeitos colaterais observados envolvem náuseas, tremores, excitação, insônia, crises convulsivas e dermatite. Enquanto que para o Benzonidazol observa-se dermatite, febre, dores articulares e musculares, vômitos e diarréia (Kirchhoff 1993). Benzonidazol é o único fármaco comercializado no Brasil, pois o Nirfutimox deixou de ser fabricado pela sua alta porcentagem de efeitos colaterais tóxicos.



Nifurtimox Benzonidazol

Figura I.1.2.1. Estruturas químicas do Nifurtimox (3-metil-4-5 (nitrofurfurilidenoamina)-tetrahidro-4H-1, tiazina 1-1 dióxido)) e do Benzonidazol (N-benzil-2-nitroimidazol acetamida).

I.1.3. Alvos Biológicos

Vários processos bioquímicos foram apontados como alvos terapêuticos potenciais, entre eles, a tripanotiona redutase (Girault *et al.* 1998), a cisteína protease (Mcgrath *et al.* 1995), a diidrofolato redutase (Zuccotto *et al.* 1999) e, as enzimas do glicossomo, envolvidas no metabolismo energético, destacando a gliceraldeído-3-fosfato desidrogenase (GAPDH) (Souza *et al.* 1998).

A cruzipaína, também chamada de cruzaína, é a principal cisteína protease do *Trypanosoma cruzi* e, é expressa em todos os estágios do ciclo de vida do parasita, porém liberada em diferentes compartimentos celulares, em cada estágio. Esta enzima é essencial para a replicação intracelular do parasita. A inibição da cruzipaína tem se mostrado capaz de dificultar a invasão das células e bloquear a replicação amastigota, como também a diferenciação de tripomastigota-amastigota, portanto, inibindo o desenvolvimento intracelular (Du *et al*. 2002).

É conhecido da literatura (Lecaille *et al.* 2002) que as cisteínas proteases possuem em comum um sítio ativo constituído de resíduos de cisteína (Cys), de histidina (His) e de asparagina (Asn). A estrutura da cruzipaína apresenta ácido glutâmico (Glu) no fundo da cavidade S_2, sendo o responsável principal pela especificidade do substrato. Similarmente a catepsina B humana, que contêm Glu205 na mesma posição, a especificidade da cavidade S_2 da cruzipaína permite a ligação de ambos resíduos P_2 arginina e fenilalanina (figura I.1.3.1).

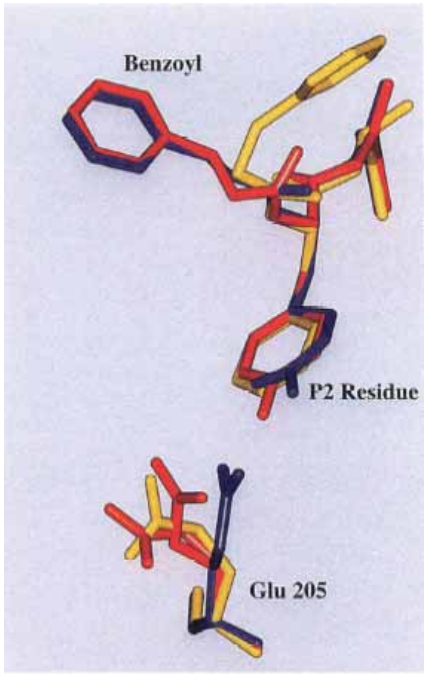


Figura I.1.3.1. Conformações da Glu205 ligada com fenilalanina (estruturas em amarelo), com arginina (estruturas em azul) ou com tirosina (estruturas em vermelho) na cavidade S_2 (Gillmor *et al.* 1997).

No final da década de 90 surgiram vários grupos estudando e propondo ligantes para a inibição da cruzaina. Entre eles, Rongshi Li e colaboradores (Li *et al.* 1996)ao estudarem séries de derivados de chalconas e de hidrazidas, identificaram um derivado de hidrazida (figura I.1.3.2.a) como o melhor inibidor da cruzaina (IC_{50} = 0,6 μM). Mais tarde, Roush estudou série de derivado de epóxi-cetonas como inibidores da cruzaína e verificou que composto mostrado na figura figura I.1.3.2.b era o inibidor mais potente (IC_{50} = 0,01 μM) da série estudada (Roush *et al.* 1998). No mesmo ano Roush (Scheidt *et al.* 1998) e colaboradores estudaram séries de derivados de aldeídos apresentando pirrolidinona na estrutura e, de derivados de vinil sulfonas e, identificaram os compostos mostrados nas figuras I.1.3.2.c (IC_{50} = 0,01 μM) e figura I.1.3.2.d (IC_{50} = 1 nM) como os mais potentes inibidores da série.

Roush e colaboradores continuaram a estudar possíveis novos inibidores contra a cruzaina (Roush *et al.* 2000), identificaram um derivado de *O*-benzil hidroxamato (figura I.1.3.2.e) como um inibidor potente e específico (IC_{50} < 0,01 μM) da cruzaína com uma alta seletividade relativa quando comparado com outras cisteínas proteazes: Cathepsin B bovina (247 maior) e da Leishmania, (51 maior) e, Papaína (305 maior). Em 2001, ao estudar um série de derivados de *N*-alcoxivinilsulfonamida, identificaram o composto mostrado na figura I.1.3.2.f (constante de taxa de inativação de segunda ordem de $6{,}48.10^{6}$ $s^{-1}.M^{-1}$) como promissor inibidor da cruzaina (Roush *et al.* 2001).

Mais adiante, Ellman e colaboradores estudaram séries de derivados de cetonas como inibidores potenciais da cruzaina e, verificaram que os compostos mostrados nas figuras I.1.3.2.g (Ki = 2,0 nM) e figura I.1.3.2.h (Ki = 4,4 nM) como os mais potentes das séries (Huang & Ellman 2002; Huang *et al.* 2003). Contemporaneamente Du e colaboradores estudaram séries de moléculas com estruturas menores do que as anteriores (derivados de tiossemicarbazonas figura I.1.3.2.i) visando melhorar a biodisponibilidade oral (Du et al. 2002).

Os valores das constantes de inibição não devem ser comparados entre os compostos mostrados na figura I.1.3.2, pois são obtidos por procedimentos diferentes.

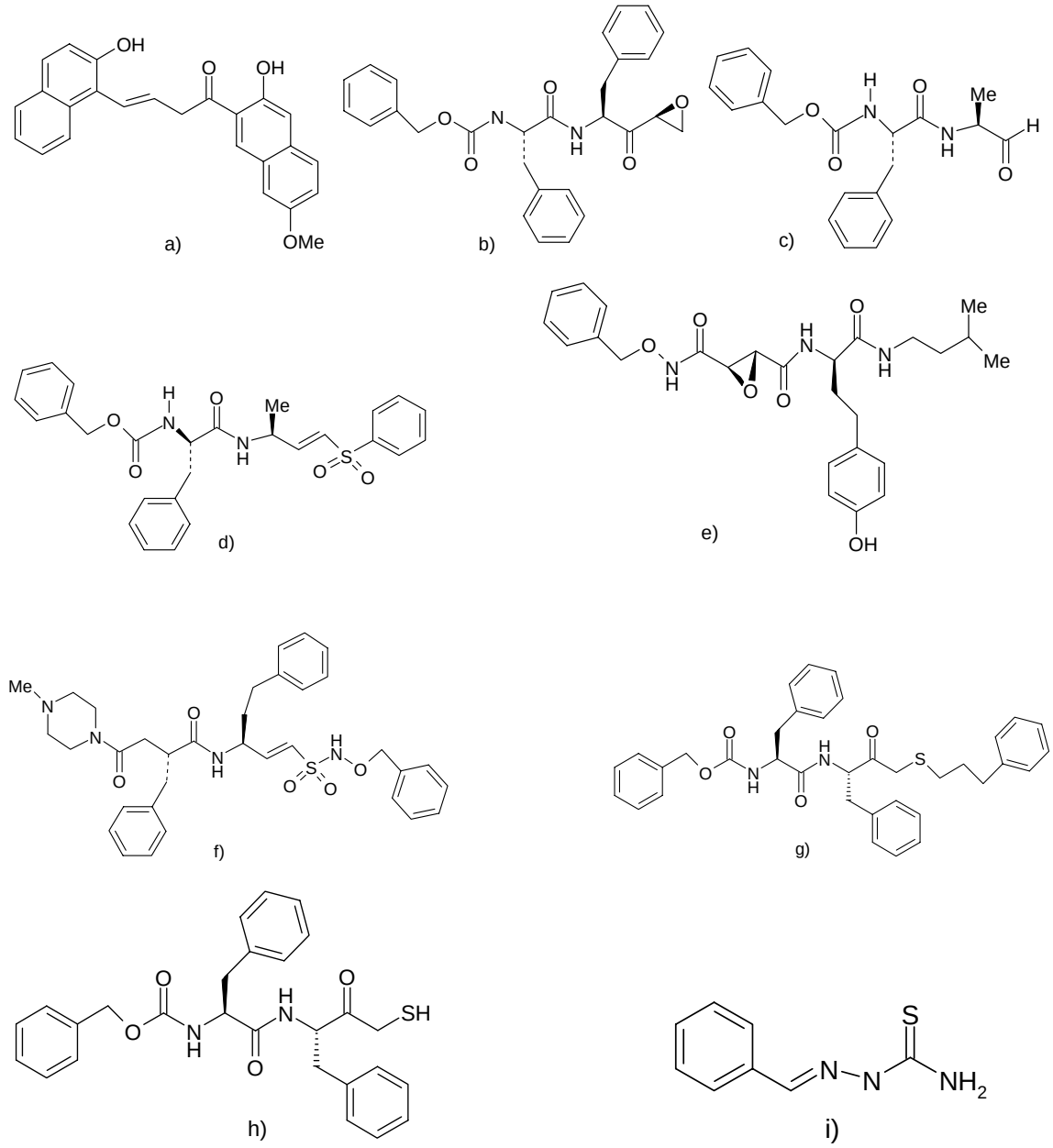


Figura I.1.3.2. Algumas estruturas selecionadas de compostos estudados como inibidores da cruzaina: a) (Li *et al.* 1996); b) (Roush *et al.* 1998); c) e d) (Roush *et al.* 1998); e) (Roush *et al.* 2000); f) (Roush *et al.* 2001); g) (Huang & Ellman 2002); h) (Huang *et al.* 2003); i) (Du *et al.* 2002).

I.2. Interação Ligante-Alvo Biológico

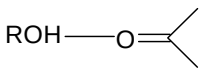
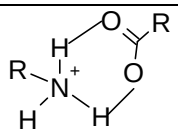
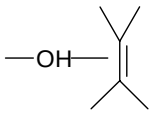
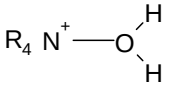
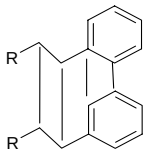
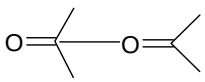
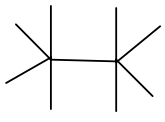
A atividade biológica de um composto é o resultado da(s) interações biológicas deste com o sistema biológico (Andrews *et al.* 1984).

A interações que ocorrem entre um composto e o sistema biológico são de diferentes intensidades e naturezas química. As interações envolvendo ligações covalentes apresentam energia de formação alta, sendo irreversíveis, não sendo importantes para a maioria dos fármacos de interesse terapêutico, exceto nos compostos que apresentam atividade anticancerígena (ex: agentes alquilantes (Pires 1998; Hansch *et al.* 2001)); compostos que inibem de forma irreversível a acetilcolinesterase (ex: inseticidas organofosfatos e, compostos intercalantes do DNA (ex: malfalam (Pires 1998), (Lattin 1995).

A intensidade da interação entre um composto, ou ligante. com o sistema biológico, na formação do complexo composto-receptor, depende das complementaridades estéricas e eletrostáticas destes (Seydel *et al.* 1979).

A variação da energia livre (ΔG) associada à formação do complexo ligante-receptor (LR) pode ser descrita como a somatória das variações de energia livre associadas às interações de naturezas, respectivamente, eletrostática, polar, não polar e hidrofóbica que ocorrem entre as moléculas do ligante e do receptor (tabela I.2.1).

Tabela I.2.1 – Principais tipos de interações entre um composto e o sistema biológico, um exemplo e as respectivas faixas de valores de energias envolvidas. (Pires 1998)

INTERAÇÃO			INTERAÇÃO		
Tipo	Energia (KJ/mol)	Exemplo	Tipo	Energia (KJ/mol)	Exemplo
Covalente	170-600	CH_3-OH	Ligação de Hidrogênio	4-17	
Iônica	40		Transferência de Carga	4-17	
Íon-dipolo	4-17		Hidrofóbica	4	
Dipolo-Dipolo	4-17		Van der Waals	2-4	

Na figura I.2.1 estão apresentados esquematicamente os fatores energéticos (entrópicos e entálpicos) envolvidos na formação do complexo ligante/receptor.

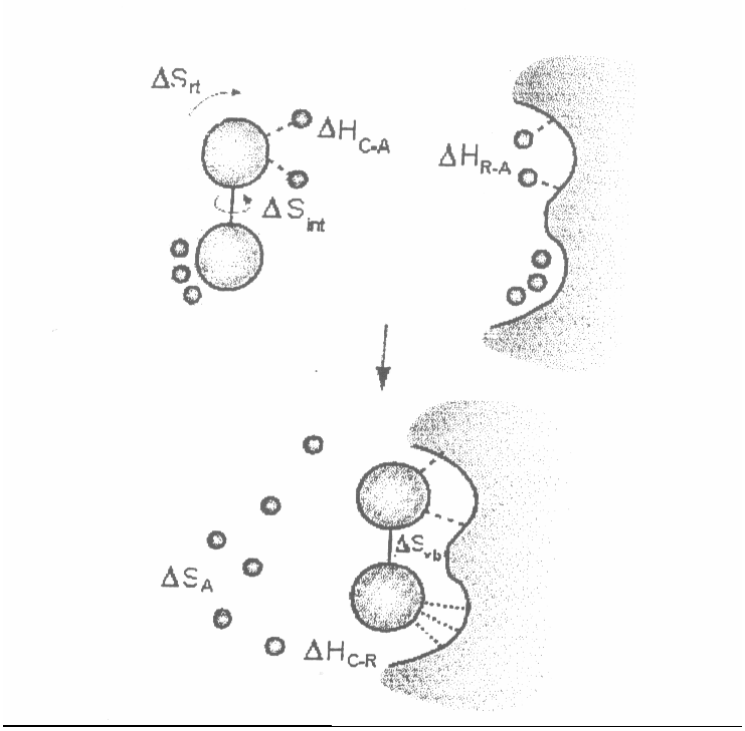


Figura I..2.1. Balanço Energético da(s) interação(ões) composto-receptor: fatores entálpicos e entrópicos envolvidos. (Andrews *et al.* 1984; Pires 1998).

Na figura I.2.1, os termos $\Delta H_{C\text{-}A}$ e $\Delta H_{R\text{-}A}$ são os valores das entalpias de solvatação, respectivamente, do ligante (C) e do receptor (R), energia que necessita ser fornecida para a dessolvatação. O termo ΔS_{rt} corresponde ao valor de energia que precisa ser fornecido devido à diminuição da entropia do sistema devido à conversão dos de graus de liberdade de rotação e de translação do composto livre para a de vibração do complexo composto-receptor. O termo ΔS_{int} corresponde ao valor de energia devido à diminuição de flexibilidade conformacional do ligante. O termo $\Delta H_{C\text{-}R}$ é o valor de energia liberada devido as complementaridades estéricas e eletrostáticas, respectivamente, do ligante e do receptor. ΔS_A corresponde ao valor de energia liberada pelo aumento de entropia das moléculas de água que deixam de estar organizadas ao redor das superfícies de contato do ligante e do receptor. ΔS_{vib} é a entropia residual de vibração do complexo composto-receptor. As interações eletrostáticas e polares entre o ligante e o solvente estão contidas no termo $\Delta H_{C\text{-}A}$, que é uma quantidade de energia que precisa ser fornecida para se separar as

moléculas de água do ligante. Da mesma forma, as interações eletrostáticas e, polares entre a água e o receptor, compõem o termo ΔH_{R-A}. O termo ΔH_{C-R} contem os valores das energias relativas às interações eletrostáticas, polares e não polares entre o ligante e o receptor enquanto que o termo ΔS_A se refere às interações hidrofóbicas, aumentando a entropia da água por ocasião da formação do complexo composto-receptor.

A variação de energia livre que ocorre na formação do complexo composto-receptor é a soma de todos estes apontados acrescida de mais dois termos: ΔS_{rt} que é devido à diminuição da entropia do ligante, por perda das liberdades de rotação e de translação e ΔS_{int} que é a perda entrópica relativa ao ligante, devido à perda da liberdade conformacional por ocasião da formação do complexo. Estes dois últimos termos entrópicos se convertem apenas em uma entropia residual vibracional no complexo composto-receptor ΔS_{vib} (Pires 1998; Andrews *et al.* 1984).

O balanço líquido de energia livre (ΔG_{C-R}) resultante das interações eletrostáticas e, hidrofóbicas menos os custos energéticos associados às perdas de entropias rotacionais, translacionais e conformacionais deve ser diretamente proporcional ao valor do logaritmo da constante de dissociação do complexo ligante/receptor ($\log K_{C-R}$) (equação I.2.1) (Andrews *et al.* 1984) (Malvezzi 2003).

$$\Delta G_{C-R} = -2{,}3RT \log K_{C-R} \qquad \textbf{Equação I.2.1.}$$

Onde: ΔG_{C-R} representa a variação de energia livre da interação entre o ligante e o receptor;
R é o valor da constante universal dos gases;
T é o valor da temperatura (em graus Kelvin);
logK é o logaritmo da constante de equilíbrio da interação entre o ligante e o receptor.

I.3. Descritores baseados na reatividade química

A obtenção do valor de energia livre (ΔG_{C-R}) a partir das estruturas químicas do ligante e do receptor (situação desejada que permitiria a determinação do valor da constante de dissociação do complexo ligante/receptor – logK) é ainda uma das etapas de objeto de estudos exaustivos por vários grupos de pesquisa (Andrews *et al.* 1984; Malvezzi 2003).

Uma abordagem simplificada envolve o estabelecimento de Relações Lineares de Energia Livre, conhecida como a Análise de Hansch ou Abordagem Extratermodinâmica.

Esta foi proposta quando as variações de energia livre ($\Delta\Delta G_{C-R}$) causadas por mudanças estruturais do ligante, podem ser correlacionadas com os valores das afinidades relativas destes ligantes com o mesmo receptor (Hansch *et al.* 1962; Hansch *et al.* 1963; Hansch & Fujita 1964). Nestas correlações são utilizados parâmetros físico-químicos que são relacionados com as forças intermoleculares envolvidas na interação entre o ligante e o receptor (tabela I.3.1). Portanto, Abordagem Extratermodinâmica, refere-se aos parâmetros físico-químicos que são correlacionados com a energia livre, porém não utiliza a estrutura formal da termodinâmica mostrada na equação equação I.2.1.

Tabela I.3.1. Alguns exemplos de parâmetros físico-químicos/estruturais utilizados como descritores; as correspondentes propriedades moleculares e naturezas das interações ligante-receptor (alvo biológico) envolvidas.

Interações Ligante-Alvo Biológico		
Parâmetro Descritor [a]	**Propriedade Molecular**	**Natureza da Interação**
σ_m; σ_p; $\mathfrak{I}$; $\mathfrak{R}$;	Densidade eletrônica	Eletrostática
MR; V;	Polarizabilidade	Dispersão
E_S; r_V;	Topologia	Estérica
logP; π;	Lipofilicidade	Hidrofóbica

[a] σ_m e σ_p são constantes eletrônicas de grupos substituintes de Hammet; $\mathfrak{I}$ e $\mathfrak{R}$ são as constantes eletrônicas de Swain e Lupton; MR é a refratividade molar; V é o volume; E_S é a constante estérica de Taft; r_V é o raio de van der Waals; logP é o coeficiente de partição; π é a constante hidrofóbica de substituinte de Hansch.

I.4. Descritores Moleculares

As propriedades físico-químicas como também a atividade biológica de compostos orgânicos dependem de suas estruturas moleculares. Com a finalidade de se obter relações entre as estruturas químicas e a atividades biológicas utilizando abordagens computacionais, é necessário encontrar representações apropriadas da estrutura molecular dos compostos (Hansch *et al.* 1990).

Um descritor molecular pode ser considerado como sendo o resultado obtido de procedimento lógico e matemático, aplicado às informações químicas codificadas através

de uma representação de uma molécula (Consonni *et al.* 2002a). Este procedimento transforma estas informações em um valor numérico associado a uma determinada propriedade molecular importante para posterior análise, correlacionado com uma propriedade molecular, como por exemplo, ponto de fusão, ou a uma atividade biológica. Porém, estas correlações são raramente obtidas, pois os sistemas estudados são freqüentemente complexos e uma relação entre uma propriedade molecular com os descritores moleculares não é, em geral, claramente entendido e, conseqüentemente definido ambiguamente. O mais importante para ser considerado e, definitivamente limitante é o fato dos sistemas em muitos casos não são completamente conhecidos (Kubinyi 1993b).

Os métodos que podem ser aplicados para se obter relações entre as estruturas moleculares dos ligantes e as afinidades relativas destes com o receptor dependem se a estrutura do receptor é conhecida. Se a estrutura do receptor não for conhecida, as variações da atividade biológica podem ser relacionadas com as relativas diferenças dos descritores moleculares, em uma determinada série de moléculas. Alguns destes descritores necessitam de um alinhamento estrutural (superposição) das moléculas. Este alinhamento irá determinar como um descritor diferencia uma molécula de outra (Klebe *et al.* 1994).

Por outro lado, sabendo-se que ligações não covalentes são as principais responsáveis pelas interações entre o ligante e o receptor, e, estas podem ser descritas em termos estéricos e eletrostáticos (Good *et al.* 1993). O estudo das propriedades estéricas envolvidas nas interações entre os ligantes e os receptores biológicos é freqüentemente decisivo no entendimento das características estruturais dos ligantes para a atividade biológica. Os efeitos estéricos ocorrem de diversas maneiras. Sugere-se na literatura (Hansch *et al.* 1990) que este pode aparecer como resultado da repulsão entre os átomos não ligados. Tais repulsões podem determinar não apenas a influência intramolecular estérica dos substituintes nas propriedades moleculares, mas também a influência intermolecular específica da afinidade do ligante pelo o receptor. E, em particular, nos métodos de *QSAR* clássico, consideram-se ainda insatisfatórios (Hansch *et al.* 1990), os métodos disponíveis para quantificar as características topológicas de um composto para poder compará-las com os outros descritores de propriedades físico-químicas. Apenas propriedades estéricas de substituintes ou, de certas subestruturas, podem ser

adequadamente descritas, enquanto mais informação é necessária para análises precisas dos efeitos estéricos das interações dos ligantes com o sítio ativo dos receptores (Hansch *et al.* 1990).

Neste contexto, encontram-se na literatura (Carbo *et al.* 1980; Hodgkin & Richards 1987; Reynolds *et al.* 1992; Good 1992; Serilevy *et al.* 1994) vários trabalhos envolvendo cálculos de similaridade com o objetivo de serem utilizados como um método de gerar parâmetros para as análises de *QSAR.* Em geral, os cálculos de similaridade comparam os compostos da série estudada considerando algumas propriedades, como por exemplo, densidade eletrostática, potencial eletrostático, e, formato. (Serilevy *et al.* 1994; Good *et al.* 1993). Considerando-se as relações observadas entre similaridades moleculares e as correspondentes variações nos valores de atividade biológica, diferentes expressões de similaridade química tem sido investigadas (Kubinyi *et al*. 1998).

Adicionalmente, decorrente do enorme desenvolvimento dos sistemas de modelagem molecular, encontram-se na literatura (Sadowski & Gasteiger 1993; Sadowski *et al.* 1994) muitos bancos de dados, baseados em cristalografias de raio-X, e estes estão disponíveis para fornecer dados de diferentes tipos de estruturas em 3 dimensões. E, ainda mais, o desenvolvimento computacional possibilitou realizar mais rapidamente cálculos que geram as estruturas em 3 dimensões (Sadowski & Gasteiger 1993). Conseqüentemente, encontram-se na literatura inúmeros descritores moleculares, como por exemplo, índices topológicos como também índices que codificam as informações geométricas em 3D da molécula. (Consonni *et al.* 2002a).

Ao mesmo tempo, observa-se na literatura, uma procura crescente (Todeschini & Gramatica 1997a; Consonni *et al.* 2002a) tanto de descritores moleculares que sejam, no entanto, validados bem como de métodos de seleção (Baroni *et al.* 1993) (Kubinyi 1994; Golbraikh & Tropsha 2002; Gasteiger *et al.* 2003) visando representar significativamente as informações relacionadas às propriedades físico-químicas e/ou à atividade biológica contidas nas séries de compostos estudadas.

Entre os programas existentes para cálculos de descritores moleculares (Xtsar, AMPAC, Molconnz, CODESSA), cita-se o programa DRAGON, versão 3.0 (Todeschini, *et al.* 2002) que calcula até 1497 descritores. Todos estes descritores são facilmente e, rapidamente calculados , portanto sendo apropriados para análise de *QSAR* e análises de

similaridade/diversidade de extensos bancos de dados (Consonni *et al.* 2002a). A grande maioria dos descritores DRAGON (Topológicos, Geométricos, BCUT, Autocorrelação 2D, Índices de Carga Topológica Galvez, WHIM, GETAWAY, RDF, 3D-MoRSE entre outros} são holísticos (Guha *et al.* 2004), portanto utilizados para classificar séries de dados em termos de características globais. (tabela A1 do Apêndice)

I.4.1. Descritores GETAWAY

Sigla utilizada para *Geometric Topology and Atom Weights Assembly.* Estes descritores são calculados a partir de uma matriz de influência molecular *MIM* (H) (equação I.4.1.1), que é calculada utilizando a matriz de coordenadas dos átomos (M) em relação ao centro da molécula com geometria em 3 dimensões, como definida no item I.4.1.2. Descritores WHIM. Na matriz de influência molecular (H), as linhas representam os átomos (inclusive o Hidrogênio) e as colunas as coordenadas x, y e z de cada átomo de uma estrutura molecular em 3 dimensões. A matriz de influência molecular é simétrica *A* X *A*, onde *A* representa o número de átomos.

Os elementos diagonais (h_{ii}) da matriz de influência molecular, denominados *leverages*, representam cada átomo na determinação da forma molecular. O valor da somatória dos elementos diagonais pode ser 1, 2 ou 3, para moléculas lineares, planares e em 3 dimensões, respectivamente. Os átomos presentes na periferia da molécula apresentam maiores valores de *leverage* que os localizados no centro. Átomos maiores também apresentam maiores valores de *levarage* que átomos menores. Moléculas esféricas apresentam átomos com menores valores de *levarage* que moléculas lineares. Para série de moléculas com aproximadamente a mesma conformação, o maior valor de *leverage* decresce com o aumento do número de átomos na molécula. Os valores de *leverage* dependem da geometria da molécula e são sensíveis à mudança conformacional e ao comprimento das ligações e portanto a sua multiplicidade.

Os elementos (h_{ij}) fora da diagonal representam os graus de acessibilidade do átomo *j* para interagir com o átomo *i*, e valor da somatória destes elementos é sempre 0. Valores negativos destes elementos significam que os átomos ocupam posições opostas em relação ao centro da molécula.

Os descritores calculados a partir da matriz de influência molecular (H), denominados descritores H-GETAWAY, podem ser ponderados pelas propriedades atômicas como massa atômica, polarizabilidade, volume de van der Waals e eletronegatividade, respectivamente.

$$H = M \bullet (M^T \bullet M)^{-1} \bullet M^T \qquad \text{Equação I.4.1.1}$$

Os descritores $H_k(w)$ (equação I.4.1.2) estão entre os descritores obtidos através da matriz de influência molecular (H). Nesta equação k é a distância topológica fixada, w_i e w_j são as propriedades atômicas respectivamente dos átomos i e j, d_{ij} é a distância topológica entre os átomos i e j, h_{ij} são os elementos fora da diagonal da matriz de influência molecular e representam o grau de acessibilidade entre os átomos i e j. $\delta(k;d_{ij};h_{ij})$ é a função delta de Dirac definida na equação I.4.1.3.

$$H_k(w) = \sum_{i=1}^{A-1} \sum_{j>i} h_{ij} w_i w_j \delta(k; d_{ij}; h_{ij}) \qquad \text{Equação I.4.1.2.}$$

$$\delta(k; d_{ij}; h_{ij}) = \begin{cases} 1 \text{ se } \mathrm{d_{ij}} = \mathrm{k} \text{ e } \mathrm{h_{ij}} > 0 \\ 0 \text{ se } \mathrm{d_{ij}} = \mathrm{k} \text{ ou } \mathrm{h_{ij}} \leq 0 \end{cases} \qquad \text{Equação I.4.1.3.}$$

Os descritores $H_k(w)$ são descritores de autocorrelação, onde são considerados apenas os valores das propriedades dos átomos que estejam numa distância topológica igual a determinada (k) e apresentem valor de valores de acessibilidade positivos (h_{ij}), pois este valor positivo significa que há uma chance de interagir entre estes átomos. Como todos os descritores de autocorrelação, os descritores $H_k(w)$ são utilizados para verificar similaridade/dissimilaridade numa série de compostos (Consonni *et al.* 2002a; Consonni *et al.* 2002b).

A partir da matriz de influência molecular (H), criou-se uma nova matriz R denominada matriz de influência/distância. A matriz R (equação I.4.1.4) utiliza os valores de *leverages* h_{ii}, h_{jj} (elementos diagonais da matriz de influência molecular - H) de dois

átomos *i* e *j* quaisquer da molécula e a distância geométrica entre estes r_{ij}. Os elementos diagonais da matriz R apresentam valor 0 (zero), enquanto cada elemento não diagonal é calculado pela razão dos valores da média geométrica dos elementos diagonais da matriz H com a distância geométrica entre os dois átomos (Consonni *et al.* 2002a).

$$\left[R\right]_{ij} = \left[\frac{\sqrt{h_{ii} \bullet h_{ij}}}{r_{ij}}\right]_{ij} \quad i \neq j \qquad \text{Equação I.4.1.4}$$

Os maiores valores dos elementos da matriz R derivam dos átomos mais externos (mais altos *levarages*) e simultaneamente próximos um do outro no espaço molécula (uma pequena distância interatômica).

A somatória das linhas da matrix de influência/distância codifica alguma informação útil que poderia ser relacionada à presença de substituintes ou de fragmentos na molécula. Os autores (Consonni *et al.* 2002a) observaram que valores altos das somatórias das linhas correspondem a átomos terminais que estão localizados a outros átomos terminais como aqueles presentes nos substituintes de uma molécula.

Os descritores calculados a partir da matriz de influência/distância (R), denominados descritores R-GETAWAY, podem ser ponderados pelas propriedades atômicas como massa atômica, polarizabilidade, volume de van der Waals e eletronegatividade.

Os descritores $R_k(w)$ (equação I.4.1.5) estão entre os descritores obtidos através da matriz de influência/distância (R). Nesta equação *k* é a distância topológica fixada, w_i e w_j são as propriedades atômicas respectivamente dos átomos *i* e *j*, d_{ij} é a distância topológica entre os átomos *i* e *j*, h_{ii} e h_{jj} elementos da diagonal da matriz de influência molecular, representam a influência do átomo na forma da molécula, r_{ij} distância geométrica entre os átomos *i* e *j*, e $\delta(k;d_{ij})$ é a função delta de Dirac definida na equação I.4.1.6.

$$R_k(w) = \sum_{i=1}^{A-1} \sum\nolimits_{j>i} \frac{\sqrt{h_{ii} h_{jj}}}{r_{ij}} w_i w_j \delta(k; d_{ij}) \qquad \text{Equação I.4.1.5}$$

$$\delta(k, dij) = \begin{Bmatrix} 1 \text{ se dij} = \text{ k} \\ 0 \text{ se dij} \neq k \end{Bmatrix} \qquad \text{Equação I.4.1.6}$$

Os descritores $R_k(w)$ são descritores de autocorrelação, onde é considerado apenas os valores das propriedades dos átomos que estejam numa distância topológica igual a determinada (k). Como todos os descritores de autocorrelação, os descritores $R_k(w)$ são utilizados para verificar similaridade/dissimilaridade numa série de compostos (Consonni *et al.* 2002a;Consonni *et al.* 2002b).

A classe GETAWAY apresenta um total de 197 descritores (Consonni *et al.* 2002b; Consonni *et al.* 2002a).

I.4.2. Descritores WHIM

Sigla utilizada para *Weighted Holistic Invariant Molecular*. São descritores baseados na análise de componentes principais (PCA) (Wold *et al.* 1987) aplicadas à uma matriz de coordenadas dos átomos de uma molécula em relação ao centro da molécula com geometria em 3 dimensões (matriz molecular). Nesta matriz as linhas representam os átomos, portanto uma molécula (com **n** átomos) gera uma matriz com n linhas e três colunas representando as coordenadas x, y, z. Além da matriz molecular, é definida uma matriz diagonal n x n, onde os elementos da diagonal principal contêm os valores de uma propriedade atômica (sem nenhuma propriedade – valores unitários, massa atômica, volume de van der Walls, eletronegatividade, polarizabilidade, ou estado eletrotopológico (Kier *et al.* 1991).

Uma matriz de covariância ponderada (3X3 - invariância com relação à translação e rotação) é obtida através dos dados das duas matrizes (matriz molecular e a matriz com os valores de propriedade atômica), através da equação I.4.2.1, semelhante ao cálculo do momento de dipolo. Nesta equação, n é o número de átomos, w_i é a propriedade atômica do átomo i, q_{ij} e q_{ik} são respectivamente os valores das coordenadas j (j = 1, 2 e 3) e k do átomo i, $\overline{q}_j$ e $\overline{q}_k$ são respectivamente os valores das médias dos valores da coordenada j e k.

$$s_{jk} = \frac{\sum_{i=1}^{n} w_i (q_{ij} - \overline{q}_j)(q_{ik} - \overline{q}_k)}{\sum_{i=1}^{n} w_i} \qquad \text{Equação I.4.2.1}$$

A análise de componentes principais (PCA) é executada sobre a matriz de covariância, obtendo 3 autovalores (λ_1, λ_2 e λ_3) e a matriz de autovetores. As coordenadas dos átomos são projetadas em cada componente principal t_m (m=1,2 e 3), gerando uma nova matriz de coordenadas (matriz T - invariância com relação à translação e rotação). Finalmente os descritores são calculados a partir dos dados desta matriz (Belvisi *et al.* 1994).

Os descritores WHIM são construídos de forma que tentem capturar as informações relevantes em 3 dimensões com relação, respectivamente ao tamanho, forma, simetria e distribuição dos átomos numa molécula independente da referência de coordenadas. Portanto, a abordagem WHIM pode ser definida como uma procura generalizada dos eixos principais com respeito a uma propriedade molecular definida.

Os descritores WHIM são divididos em dois tipos de descritores: direcionais e não direcionais.

Os descritores direcionais são divididos em 4 tipos relacionados, respectivamente ao tamanho, ao formato, à simetria da molécula e à distribuição dos átomos (acessibilidade entre os mesmos)

Os descritores relacionados ao tamanho da molécula são definidos diretamente pelos autovalores λ_1, λ_2 e λ_3. Os descritores relacionados ao formato da molécula são obtidos pela equação I.4.2.2, onde ϑ_m (m = 1, 2 e 3) são os autovalores proporcionais calculados a partir dos valores dos autovalores (λ_1, λ_2 e λ_3). Como $\vartheta_1 + \vartheta_2 + \vartheta_3 = 1$, Só dois descritores são independentes.

$$\vartheta_m = \frac{\lambda_m}{\sum_m \lambda_m} \qquad \text{Equação I.4.2.2}$$

Os descritores relacionados à simetria (γ_1, γ_2 e γ_3) são obtidos através das equações I.4.2.3 e I.4.2.4. Nestas, n_s é a soma de todos os grupos de átomos que apresentem os mesmos autovalores, com sinais opostos, presentes no mesmo componente *m*, na é o número de átomos os quais seus apresentem autovalores opostos simétricos presentes no mesmo componente. $0 < \gamma \leq 1$

$$\gamma'_m = -\left[\frac{n_s}{n}\log_2\frac{n_s}{n} + n_a\left(\frac{1}{n}\log_2\frac{1}{n}\right)\right] \qquad \text{Equação I.4.2.3}$$

$$\gamma_m = \frac{1}{1+\gamma'_m} \qquad 0 < \gamma \leq 1 \qquad \text{Equação I.4.2.4}$$

O quarto tipo de descritor (η_m) relacionado à acessibilidade dos átomos, é calculado a partir da inversa da kurtosis k_m (equações I.4.2.5. e I.4.2.6.). Onde t_{im} é o valor da projeção do átomo i no eixo principal t_m.

$$k_m = \frac{\sum_i t_{im}{}^4}{\lambda_m{}^2 n} \qquad \text{Equação I.4.2.5}$$

$$\eta_m = \frac{1}{k_m} \qquad \text{Equação I.4.2.6}$$

O grupo de descritores η_m, pode ser interpretado como a quantidade de espação não preenchido por átomo projetado. Quanto menor for o valor da kurtosis, maior será o valor de η_m, portanto maior o espaço projetado não preenchido.

Os descritores não direcionais WHIM são diretamente derivados dos descritores direcionais, não dependendo dos eixos principais t_m. Os descritores T, A e V representam respectivamente às contribuições linear, quadrática e completa para o tamanho da molécula (equações I.4.2.7 a I.4.2.9). O formato da molécula, a simetria da molécula e sua densidade são representados respectivamente por K, G, D (equação I.4.2.10 a I.4.2.12).

$$T = \lambda_1 + \lambda_2 + \lambda_3 \qquad \text{Equação I.4.2.7}$$

$$A = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3 \qquad \text{Equação I.4.2.8}$$

$$V = T + A + \lambda_1\lambda_2\lambda_3 \qquad \text{Equação I.4.2.9}$$

$$K = \frac{\sum_m \left| \frac{\lambda_m}{\sum_m \lambda_m} - \frac{1}{3} \right|}{\frac{4}{3}} \qquad 0 \le K \le 1 \qquad \text{Equação I.4.2.10}$$

$$G = (\gamma_1\gamma_2\gamma_3)^{1/3} \qquad \text{Equação I.4.2.11}$$

$$D = \eta_1 + \eta_2 + \eta_3 \qquad \text{Equação I.4.2.12}$$

Esta classe apresenta 99 descritores (Todeschini & Gramatica 1997a) (Todeschini & Gramatica 1997b);

I.4.3. Descritores RDF

Sigla utilizada para *Radial Function Distribution*. São obtidos através da função (equação I.4.3.1) de distribuição radial calculada sobre as distâncias interatômicas de uma molécula. A função pode ser interpretada como sendo a distribuição de probabilidade para encontrar um átomo em um volume esférico de raio de valor r. (Hemmer *et al*. 1999).

Na equação I.4.3.1, **N** é o número de átomos da molécula, *f* é um faro de escalonamento, A_i e A_j são propriedades dos átomos (massa atômica, eletronegatividade, volume de van der Waals e pela polarizabilidade) *i* e *j* respectivamente. No termo exponencial da equação, r_{ij} é a distância entre os átomos *i* e *j*, *B* é um parâmetro de

aplainamento (que define a distribuição de probabilidade das distâncias individuais), e *r* é o raio pré-definido. Quanto maior o valor de *B*, maior é a influência da diferença das distâncias nos valores de *g(r)*.

Esta classe de descritores apresenta algumas características em comum com a classe de descritores 3D MoRSE desenvolvida pelo mesmo grupo de pesquisa (Schuur *et al.* 1996) (descrita no item I.4.4. *Descritores 3D-MoRSE*). Estas características são:

- independência da quantidade dos valores do número de átomos, ou seja, do tamanho da molécula;
- exatidão relativa ao arranjo em 3 dimensões dos átomos;
- invariância com relação à translação e rotação da molécula inteira;

$$g(r) = f\sum_{i}^{N-1}\sum_{j>i}^{N} A_i A_j e^{-B(r-r_{ij})^2} \qquad \text{Equação I.4.3.1}$$

Esta classe apresenta 150 descritores (Hemmer *et al.* 1999);

I.4.4. Descritores 3D-MoRSE

Os descritores 3D MoRSE (*Molecule Representation of Structure based on Electron diffraction)*. Estes descritores refletem a distribuição em três dimensões de diferentes propriedades moleculares e, expressam informações sobre a ramificação das moléculas.

Os descritores são obtidos através da somatória dos produtos de cada uma das propriedades atômicas, a saber ou entre elas: massa, eletronegatividade, volume de van der Waals e, polarizabilidade. A função de cálculo (equação I.4.4.1), deriva daquela utilizada determinação da estrutura molecular através das medidas de difração eletrônica. Devido a característica desta função, o número de valores obtidos independem do tamanho da molécula. Nesta função A_i e A_j são os valores das diferentes propriedades dos átomos *i* e *j*, r_{ij} é a distância interatômica entre os respectivos átomos, e s é um fator que divide a função

em 32 valores. Para o cálculo do descritor Mor07m, a propriedade utilizada para A_i e A_j é a massa atômica e o valor de *s* é 7 Å^{-1}. Esta classe apresenta 160 descritores.(Schuur *et al.* 1996) (Gasteiger *et al.* 1996)

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{sin(sr_{ij})}{sr_{ij}} \qquad \text{Equação I.4.4.1}$$

Onde: s = 1,2,...32 Å^{-1}

I.4.5. Descritores de Autocorrelação 2D

Os descritores de autocorrelação 2D podem ser definidos como relação entre valores de uma única variável entre os átomos (considerando a distância topológica entre estes) de uma molécula representada em 2 dimensões.

Os descritores de autocorrelação derivam de funções matemáticas que foram utilizadas principalmente para estudos estatísticos geográficos (Moran 1950) (Geary 1954). Os descritores de autocorrelação gerados pelo programa DRAGON são: ATS (descritor de autocorrelação de uma estrutura topológica Broto-Moreau), MATS (Moran autocorrelation), GATS (Geary autocorrelation).

Os descritores ATS, são derivados da função matemática (equação I.4.5.1**)**, onde *f(x)* é a medida de uma propriedade associada a cada ponto do segmento AB, *f(x+t)* é a medida da mesma propriedade em um ponto diferente de *f(x)*. Portanto a função *F(t)* é a descrição da mesma propriedade, porém com uma precisão menor. Contudo a *F(t)* tem uma vantagem de independer de um referencial externo, já que t é uma variável interna e permanece inalterada quando é a função f(x) é transladada ao longo do eixo x. A autocorrelação também é utilizada no tratamento de sinais elétricos como a eletroencefalografia (Moreau & Broto 1980).

$$F(t) = \int_{AB} f(x) f(x+t) dx \qquad \text{Equação I.4.5.1}$$

A equação equação I.4.5.1 função é adaptada para a forma vetorial considerando as distâncias topológicas entre os átomos (*i* e *j*) de uma molécula representada em 2 (equação I.4.5.2).

$$S^2 = \sum_i f^2(i) + \sum_{i \neq j} 2 f(i) f(j) \qquad \text{Equação I.4.5.2}$$

O primeiro termo da equação I.4.5.2 é o primeiro componente do vetor de autocorrelação, o qual é associado a um valor de distância topológica igual a 0. O segundo termo pode ser dividido em diversas somatórias parciais contendo pares de átomos separados com o mesmo valor de distância topológica. Estas somas parciais são os outros componentes do vetor de autocorrelação (Broto *et al.* 1984).

Os descritores ATS obtidos pelo programa DRAGON utilizam o segundo termo da equação I.4.5.2, portanto é obtido para os átomos com distâncias topológicas maiores ou iguais a 1 (equação I.4.5.3). Nesta equação k é um valor de distância topológica pré-determinada, Né o número de átomos na molécula, *Ai* e *Aj* são propriedades atômicas (massa atômica, o volume de van der Waals, a polarizabilidade ou a eletronegatividade) dos átomos *i* e *j* que estejam a uma distância topológica k, e δ é a função delta de Dirac definida na equação I.4.5.4. (Broto *et al.* 1984; Consonni *et al.* 2002a).

$$ATS_k = \sum_{i=1}^{N-1} \sum_{j>i} A_i A_j \delta(k, d_{ij}) \qquad \text{Equação I.4.5.3}$$

$$\delta(k, dij) = \begin{Bmatrix} 1 \text{ se dij} = \text{k} \\ 0 \text{ se dij} \neq k \end{Bmatrix} \qquad \text{Equação I.4.5.4}$$

O descritor de autocorrelação Moran (MATS - equação I.4.5.5), um dos mais antigos descritores de autocorrelação, compara o valor de uma variável de um vértice (átomo), com todos os outros vértices, que estejam numa separados por um valor de distância topológica k. Na equação I.4.5.5 x_i e x_j são as propriedades atômicas dos átomos *i* e *j* respectivamente e $\bar{x}$ é a média dos valores da propriedade atômicas dos átomos.

Valores altos deste descritor indicam uma autocorrelação positiva, valores negativos indicam uma autocorrelação negativa (Moran 1950).

$$MATS_k = \frac{\sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x})\delta(k, d_{ij})}{\sum_i (x_i - \bar{x})^2} \qquad \text{Equação I.4.5.5}$$

O descritor de autocorrelação Geary (GATS - equação I.4.5.6), é semelhante ao descritor MATS, porém a interação não é calculada pelo produto dos desvios da média, mas pelos desvios dos valores da propriedade atômica de um vértice (átomo) com o de outro vértice. Valores maiores deste descritor indicam uma autocorrelação inversa, valores próximos de 0 indicam uma autocorrelação positiva (Geary 1954). Os descritores MATS fornecem valores mais representativos globalmente, enquanto o GATS é mais sensível a diferença de valores de propriedades de átomos vizinhos.

$$GATS_k = \frac{\sum_i \sum_j (x_i - x_j)\delta(k, d_{ij})}{\sum_i (x_i - \bar{x})^2} \qquad \text{Equação I.4.5.6}$$

Esta classe apresenta 96 descritores..

I.4.6. Descritores Geométricos Calculados pelo Programa DRAGON

São diversos descritores baseados na distância geométrica entre os átomos. Alguns destes descritores calculam a soma geométrica entre os átomos de nitrogênio, átomos de oxigênio, entre os átomos de enxofre,....

Além destes alguns descritores são baseados na *layer distance matrix* (LM3D) (equação I.4.6.2), a qual é obtida através da matriz de distância geométrica (equação I.4.1.6.1). (Diudea *et al.* 1995). Esta classe apresenta 70 descritores

$$m_i = \sum_{i=1}^{N} d_{ij} \qquad \text{Equação I.4.6.1}$$

$$lm_{ik} = \sum_{u=1}^{N} m_i \delta(k, d_{ij}) \qquad \text{Equação I.4.6.2}$$

$$\delta(k, dij) = \begin{Bmatrix} 1 \text{ se dij} = \text{k} \\ 0 \text{ se dij} \neq k \end{Bmatrix} \qquad \text{Equação I.4.6.3}$$

I.4.7. Descritores Topológicos Calculados pelo Programa DRAGON

A necessidade de usar descritores topológicos originou do fato que propriedades físico-químicas podem ser expressas em números e, portanto têm uma possibilidade numérica de se fazer comparações e correlações. Em contraste as estruturas químicas são entidades discretas, portanto é preciso que se traduzam estas estruturas em números com o objetivo de avaliar o grau de similaridade/dissimilaridade e fazer correlações com diversas propriedades físico-químicas. As estruturas em 3 dimensões das moléculas dependem da sua topologia, ou seja, das posições individuais dos átomos e das ligações entre eles (Hansch *et al.* 1990).

Os descritores topológicos, comumente (Balaban & Devillers 1999) chamados de índices topológicos (*TI*), são calculados baseados na matriz de adjacência e/ou na matriz de distância topológica de uma molécula representada em 2 dimensões. Nas representações das moléculas em 2 dimensões, os átomos e as ligações correspondentes são representados como vértices e arestas, respectivamente (Balaban & Devillers 1999).

Quando dois átomos (vértices) estão ligados (vizinhos) por uma ligação covalente (aresta), sua distância topológica é definida como 1(Balaban & Devillers 1999) e estes átomos são adjacentes. As distâncias e as adjacências entre dois átomos numa molécula representada em 2 dimensões são as menores possíveis. Exemplificando, para a representação da molécula do 1-metil-2-propil-ciclobutano (figura I.4.7.1), as matrizes de adjacência e de distância são apresentadas nas figuras, respectivamente, I.4.7.2 e I.4.7.3.

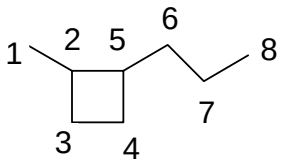


Figura I.4.7.1. Representação em 2 dimensões da estrutura molecular do 1-metil-2-propil-ciclobutano.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0	1	0	0	0	0	0	0
2	1	0	1	0	1	0	0	0
3	0	1	0	1	0	0	0	0
4	0	0	1	0	1	0	0	0
5	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	1	0
7	0	0	0	0	0	1	0	1
8	0	0	0	0	0	0	1	0

Figura I.4.7.2. Matriz de adjacência da molécula do 1-metil-2-propil-ciclobutano. Os átomos foram numerados como atribuído na figura I.4.7.1.

	1	2	3	4	5	6	7	8
1	0	1	2	3	2	3	4	5
2	1	0	1	2	1	2	3	4
3	2	1	0	1	2	3	4	5
4	3	2	1	0	1	2	3	4
5	2	1	2	1	0	1	2	3
6	3	2	3	2	1	0	1	2
7	4	3	4	3	2	1	0	1
8	5	4	5	4	3	2	1	0

Figura I.4.7.3. Matriz de distâncias topológicas da molécula do 1-metil-2-propil-ciclobutano. Os átomos foram numerados como atribuído na figura I.4.7.1.

Os descritores moleculares são regularmente criticados na literatura sobre *QSAR*. Algumas das principais críticas (Balaban & Devillers 1999) dos descritores topológicos são:

- O significado físico-químico pouco claro;

- Probabilidade de correlação ao usar um grande número de descritores altamente intercorrelacionados como por exemplo conectividade normal e conectividade de valência;
- O índice de degeneração de certos descritores topológicos pode ser alto;

Algumas das vantagens dos descritores topológicos que os fazem ser largamente utilizados nos estudos de QSAR e QSPR são:

- Os descritores topológicos podem ser calculados para todas as moléculas existentes;
- A obtenção dos valores dos descritores topológicos é relativamente rápida utilizando os computadores hoje existentes;
- O cálculo de diferentes descritores de uma mesma molécula permite uma abordagem multivariada (Balaban & Devillers 1999);

Há uma extensa quantidade de diferentes descritores topológicos presentes no programa DRAGON (266 descritores), como por exemplos:

a) O índice topológico CIC_k (equação I.4.7.2) que como o índice IC_k (índice de informação das moléculas) (equação I.4.7.1), considera os átomos de hidrogênio nas moléculas. Os valores iguais a ordem zero representa grupos de átomos isolados em classes equivalentes e a ordem 1 denota pares de átomos ligados covalentemente, agrupados em ordem de equivalência (de acordo com a natureza dos átomos, e a multiplicidade da ligação). Para uma série de n vértices, estes são considerados equivalentes se representam o mesmo elemento químico, e possuem as mesmas características estruturais com os seus vizinhos de ordem k. Se há diferentes classes diferentes classificados na ordem k, estes elementos são numerados sucessivamente p_i (i = 1, 2, 3, ..., r), onde r é o número total de diferentes elementos classificados na mesma ordem (Balaban & Devillers 1999).

$$IC_k = -\sum_{i=1}^{r} p_i \log_2 pi \qquad \text{Equação I.4.7.1.}$$

$$CIC_k = \log_2 n - IC_k$$ Equação I.4.7.2.

$$p_i = \frac{n_i}{n}$$ Equação I.4.7.3.

Onde: n_i - número de átomos de mesmo elemento com a mesma vizinhança de ordem k;

n – número total de átomos;

Pela equação I.4.7.1, verifica-se que quanto maior a diversidade entre os vértices de mesma ordem k, maior será o valor do índice de informação das moléculas (IC_k). Através da equação I.4.7.2 verifica-se que quanto maior o IC_k, menor será CIC_k, portanto quanto maior a diversidade entre os vértices de mesma ordem k, menor será o CIC_k.

b) O descritor PJI_2 é calculado a partir do raio (R) e do diâmetro D generalizados (equação I.4.7.4). O raio e o diâmetro são calculados a partir dos pontos extremos e centrais de uma molécula em duas dimensões. Todas as distâncias topológicas dos átomos (vértices) de uma molécula representada em 2 dimensões são calculadas com relação a todos os outros átomos (vértices) desta. O átomo que apresentar o maior valor de distância topológica com o átomo mais distante será considerado como ponto extremo e seu valor de distância topológica será o diâmetro generalizado (D). Conseqüentemente o átomo que apresentar o menor valor de distância topológica com o átomo mais distante será considerado como ponto central e seu valor de distância topológica será o raio generalizado (R). O ponto extremo e o centro da molécula não precisam ser únicos (Petitjean 1992).

O descritor PJI_2 pode ser interpretado com uma medida de balanço entre uma molécula cíclica e uma acíclica. Um valor de PIJ2 igual a 0, indica uma molécula estritamente cíclica, quanto maior o valor de PJI2, maior será o caráter acíclico do formato da molécula (Petitjean 1992).

$$PJI_2 = \frac{(D - R)}{R}$$ Equação I.4.7.4

I.4.8. Descritores BCUT

Sigla utilizada para os descritores propostos por Burden (B), validados pelo *Chemical Abstracts Service* (CAS) *Registry* e ampliados na Universidade do Texas (UT). Os descritores BCUT são calculados através dos autovalores obtidos da matriz de adjacência (exemplo: figura I.4.7.2) com elementos nulos da diagonal substituídos por alguma propriedade atômica (massa atômica, volume de van de Waals, eletronegatividade, e polarizabilidade) (Burden 1997).

A essência da obtenção dos descritores é resolver a equação de autovalor (equação I.4.8.1).

$$[B][V]=[V][e] \qquad \text{Equação I.4.8.1}$$

Na equação I.4.8.1, [V] é a matriz de autovetores, [e] é uma matriz diagonal de autovalores, e [B] é uma matriz de conectividade com as seguintes características (BURDEN 1989):

- Os elementos diagonais dos átomos são valores de alguma propriedade atômica (massa atômica, volume de van de Waals, eletronegatividade, e polarizabilidade);
- Os valores dos elementos não diagonais dependem da ligação existente entre os átomos i e j. O valor é um para uma ligação simples, 0,2 para uma ligação dupla, 0,3 para uma ligação tripla e 0,15 para uma ligação aromática;
- Os valores dos elementos não diagonais das matrizes dos átomos terminais (aqueles com apenas uma conexão) são aumentados por 0,01;
- Todos os outros elementos não diagonais recebem valor 0,001;

Considerando-se que esta classe de descritores depende das propriedades atômicas, pode-se aplicar esta em estudos de QSAR e QSPR (Pearlman & Smith 1999; Burden 1997) inclusive para moléculas isotopológicas, ou seja, com a mesma conectividade.

Esta classe apresenta 64 descritores ;

I.4.9. Grupos Funcionais do Programa DRAGON

Coletânea de fragmentos moleculares, contendo poucos átomos. Como por exemplo, números de carbonos primário, secundário, terciário, quaternário; de anéis aromáticos substituídos ou não-substituídos; de cetonas alifáticas ou aromáticas. Esta classe apresenta 121 descritores (Todeschini & Consonni 2000);

I.4.10. Descritores de Átomo Centrado

Os descritores de átomo centrado, identificam diversas seqüências de átomos como fragmentos e, verificado que estes fragmentos (ou seja sua estrutura química) se correlacionam com a atividade biológica (Ghose *et al.* 1988). Estes fragmentos estão classificados a partir de um átomo central, portanto estes fragmentos classificam o átomo de acordo com sua vizinhança (dependem dos átomos aos quais o átomo central está ligado e, dos tipos de ligações envolvidas: simples, dupla, tripla, aromática). Como por exemplo: CRn, número de carbonos (sp^3) ligados respectivamente à uma, a duas, a três ou, a quatro cadeias alifáticas; CXn, número de carbonos (sp^3) ligados a um, a dois, a três ou a quatro halogênios e, =CXn, número de carbonos (sp^2) ligados a um, a dois, a três ou a quatro halogênios. Esta classe apresenta 120 descritores (Viswanadhan *et al.* 1989b);

I.4.11. Descritores Constitucionais Calculados pelo Programa DRAGON

São descritores independentes da conectividade e conformação moleculares. Alguns exemplos desta classe de descritores são: tipos de átomos e de ligações, peso molecular, e somatória do volume atômico de van der Waals. Esta classe de descritores não consegue distinguir a maioria dos isômeros moleculares e as moléculas similares. Esta classe apresenta 47 descritores;

I.4.12. Propriedades Moleculares Calculadas pelo Programa DRAGON

Para se calcular a refratividade molar Ghose-Crippen (Viswanadhan *et al.* 1989a), primeiramente a molécula é dividida em fragmentos de átomo centrado (item I.4.10. Descritores de Átomo Centrado), e a seguir, a partir destes fragmentos é calculado a refratividade molar através da equação I.4.12.1. Onde n_i é o número de átomos de um determinado tipo *i*, classificado de acordo com o item I.4.10. e a_i é a refratividade molar do átomo *i*.

$$MR = \sum n_i a_i \qquad \text{Equação I.4.12.1}$$

O cálculo da área de superfície polar é tradicionalmente calculado gerando-se uma estrutura em 3 dimensões da molécula para a seguir identificar os átomos polares e finalmente determinando a o valor área de superfície polar.

O cálculo baseado nos fragmentos é obtido através da divisão da molécula em fragmentos, e a partir destes, identifica-se os fragmentos polares para finalmente obter o valor da área de superfície polar através da seguinte Equação I.4.12.2. Onde n_i é o número de átomos de um determinado tipo *i*, e a_i é a contribuição do átomo *i* para a área de superfície polar (Ertl *et al.* 2000).

$$PSA = \sum n_i a_i \qquad \text{Equação I.4.12.2}$$

O valor do coeficiente de partição octanol-água é obtido através da contribuição do número de átomos e de fragmentos (Moriguchi *et al.* 1992).

I.4.13. *Molecular Walk Counts*

Em uma molécula representada em 2 dimensões, pode-se considerar os átomos como vértices e as ligações entre estes como arestas. A partir de um vértice, ou átomo, pode-se calcular quantos caminhos existem para uma determinada distância topológica através da equação I.4.13.1. Nesta equação *i* é o vértice (átomo), *e* é a distância topológica

fixada, (ai,u)^e são os elementos da matriz A^e, e A é a matriz de adjacência como mostrada na figura I.4.7.2 (Rucker & Rucker 1993).

$$w_i = \sum_{u \in V(g)} (a_i, _u)^e \qquad \text{Equação I.4.13.1}$$

Os descritores são derivados somas dos diferentes caminhos de distância topológica *e* para uma estrutura molecular projetada num plano. Esta classe apresenta 21 descritores (Diudea *et al.* 1994);

I.4.14. Índices de Carga Topológica Galvez

Os descritores são calculados através dos elementos de uma matriz T quadrada (onde as linhas e colunas representam os átomos) obtida a partir da matriz de adjacência e da matriz do quadrado do inverso das distâncias topológicas (equação I.4.14.1).

$$T = A \times D^* \qquad \text{Equação I.4.14.1}$$

Os elementos da matriz T são denominados termos de carga (*CTij*), e os cálculo do descritor de índice de carga topológica G_k é realizada como mostrado na equação I.4.14.2, onde *i* e *j* representam dois átomos da molécula, k uma distância topológica fixada, e $\delta(k, dij)$ é a função delta de Dirac como mostrado na equação I.4.14.3.

.

$$G_k = \sum_{i=1,\, j=i+j}^{i=N-1,\, j=N} \left[CT_{ij} \right] \delta(k, d_{ij}) \qquad \text{Equação I.4.14.2}$$

$$\delta(k, dij) = \begin{Bmatrix} 1 \text{ se dij} = \text{k} \\ 0 \text{ se dij} \neq k \end{Bmatrix} \qquad \text{Equação I.4.14.3}$$

Os descritores G_k avaliam a o total de carga transferida entre os átomos a uma distância topológica fixada *k*. Os descritores J_k são calculados a partir dos descritores G_k (equação I.4.14.4).

$$J_k = \frac{G_k}{N-1} \qquad \text{Equação I.4.14.4}$$

Nesta equação N é o número de átomos ou vértice, e Jk representa o valor médio de transferência de carga para cada ligação, desde que o número de o número de arestas (ligações) em uma molécula acíclica é N-1.

Esta classe apresenta 21 descritores (Galvez *et al.* 1994);

I.4.15. Descritores de Carga Calculados pelo Programa DRAGON

São descritores calculados através dos valores das cargas atômicas da estrutura molecular com geometria otimizada. As cargas atômicas das moléculas são determinantes para as interações eletrostáticas. As cargas ou densidades eletrônicas são importantes em muitas reações químicas e propriedades físico-químicas de compostos.

São descritas na literatura (Karelson *et al.* 1996) diferentes definições de carga atômica, como a análise populacional de Mülliken, para a distribuição de cargas numa molécula. Os valores de cargas calculados por diversos métodos semi-empíricos não são coerentes entre si (Karelson *et al.* 1996). Todavia são fáceis de serem obtidos e fornecem ao menos uma informação qualitativa de distribuição de cargas numa molécula.

Mais recentemente foi verificado que a carga atômica pode ser medida e contribui para diversas propriedades como por exemplo: todos os momentos molecular, a polarizabilidade molecular, as intensidades de absorção Raman e infravermelho (Bader & Matta 2004).

Os valores das cargas atômicas parciais são usadas como índices de reatividade química. Várias somas das cargas parciais como também o quadrado destas somas são usadas para descrever interações intermoleculares e interações entre o solvente e o soluto (Karelson *et al.* 1996).

Esta classe apresenta 14 descritores como por exemplo: valores máximos das cargas positiva e negativa; valores totais de cargas positiva e negativa; de carga absoluta; da carga ao quadrado

I.4.16. Índices de Aromaticidade Calculados pelo Programa DRAGON

As propriedades aromáticas, são aquelas propriedades que tornam o benzeno distinto dos hidrocarbonetos alifáticos. Do ponto de vista experimental, as moléculas aromáticas são aquelas que apresentam alto grau de insaturação mas dificilmente participam nas reações de adição. Do ponto de vista teórico, a regra de Hückel que indica que uma molécula com nuvens cíclicas de (4n+2) elétrons é um composto aromático e com um sistema de 4n elétrons, não o é.

Alguns índices baseados que utilizam a distância geométrica entre os átomos com ligações de caráter aromático são utilizados para descrever os valores do grau de aromáticidade de uma molécula. Descritores que podem descrever os valores do grau de aromaticidade de uma molécula. Esta classe apresenta 4 descritores (Jug 1983);

I.4.17. *Randic Molecular Profiles*

É uma série de números que representam o perfil molecular calculados a partir da matriz de distância geométrica.. Estes descritores são calculados a partir de uma matriz de distância geométrica, na qual os elementos d_{ij}, representam a distância geométrica entre os átomos i e j. Os elementos das linhas ou colunas são somados para diversos expoentes e (e = 1, 2, 3...) (equação I.4.17.1) e a seguir é obtida a média da soma das linhas/colunas com N átomos (equação I.4.17.2).

$$ {}^{e}R_i = \sum_{i=1}^{N} \left(d_{ij}\right)^{e} \qquad \text{Equação I.4.17.1} $$

$$ {}^{e}R = \frac{\sum_{i=1}^{N} {}^{e}R_i}{N} \qquad \text{Equação I.4.17.2} $$

Esta classe apresenta 41 descritores (Randic 1995);

I.4.18. Descritores Empíricos Calculados pelo Programa DRAGON

Índice de insaturação; fator hidrofílico; razão aromática. Esta classe apresenta 3 descritores;

I.5. Relações Quantitativas entre Estrutura Química e Atividade Biológica: *QSAR*

A abordagem extratermodinâmica ou de Hansch-Fujita (Hansch & Fujita 1964) considera que a atividade biológica de um composto, resultado da(s) interação(ões) deste com a(s) diferentes biofase(s) pode ser expressa pela contribuição dos propriedades físico-químicas ou estruturais, identificadas como sendo responsáveis pela atividade ou resposta biológica.

Desta forma, a análise de Hansch (Hansch & Fujita 1964) correlaciona os valores das atividades biológicas com as propriedades físico-químicas através de regressão linear; linear-múltipla (Barros Neto *et al.* 2002), representadas pelos modelos descritos a seguir.

Modelo Linear

No modelo linear a atividade biológica é expressa pela somatória das contribuições individuais de parâmetros físico-químicos ou estruturais, entre eles: hidrofóbicos (ou lipofílicos) (Dearden & Bresnen 1988; Kubinyi 1993), eletrônicos (Hammett 1937, 1970; HANSCH & Leo 1995), estéricos (Taft 1971) e de dispersão (Dearden *et al.* 1991) relativos ao substituinte ou à molécula toda. Este modelo pode ser representado pela equação I.5.1:

$$\mathbf{\log 1/C = a\pi + b\sigma + cE_s + r} \qquad \text{Equação I.5.1}$$

Onde:
log 1/C é a resposta biológica do composto;
π - parâmetro de hidrofobicidade de Hansch-Fujita;
σ - parâmetro eletrônico de Hammett;
E_s – parâmetro estérico de Taft;
a, b, c – mede a contribuição de cada parâmetro sobre a resposta biológica (valor do coeficiente determinado pela análise de regressão);
C – Valor da concentração molar do composto que produz um determinado efeito biológico;
r – constante.

Como exemplo, descreve-se que a atividade anti-adrenérgica (ED_{50} – concentração molar para obter 50% efeito desejado) uma série de 22 α-bromo-fenetilaminas. (Graham & Karrar 1963; HANSCH & Lien 1968; Kubinyi, 1993a) pode ser descrita pelo modelo linear representado na equação I.5.2,

$$\log 1/ED_{50} = 1{,}151\ (\pm 0{,}19)\pi - 1{,}464\ (\pm 0{,}38)\ \sigma^{+} + 7{,}817\ (\pm 0{,}19) \qquad \text{Equação I.5.2}$$
$$(n=22;\ r=0{,}945;\ s=0{,}238;\ F=78{,}63)$$

Onde: π - parâmetro de hidrofobicidade de Hansch-Fujita;
σ^{+} - parâmetro eletrônico de Hammett;
$\log 1/ED_{50}$ é a resposta biológica do composto**.**

<u>Modelo Parabólico</u>

O primeiro modelo matemático descrevendo a dependência não-linear da atividade biológica em função dos parâmetros físico-químicos foi proposto por Hansch-Fujita (Hansch & Fujita 1964). Eles se basearam nos conceitos de absorção e de distribuição de compostos no sistema biológico.

Observaram que a atividade biológica apresentava, para alguns sistemas, uma dependência não-linear com a lipossolubilidade dos compostos destes sistemas, representada por uma parábola e, expresso pela equação I.5.3.

$$\mathbf{\log 1/C = a\pi^2 + b\pi + \rho\sigma + cEs + r} \qquad \text{Equação I.5.3}$$

Onde:
log 1/C é a resposta biológica do composto;
π - parâmetro hidrofóbico de Hansch-Fujita;
σ - parâmetro eletrônico de Hammett;
E_s – parâmetro estérico de Taft;
a, b, ρ, c – mede a influência de cada parâmetro sobre a resposta biológica (coeficiente determinado pela análise de regressão);
C – Valor da concentração molar do composto que produz um determinado efeito biológico;
r – constante.

Como exemplo, descreve-se que a atividade espasmolítica (valores de atividade relativa ao efeito produzido pelo ciclandelato – RA = 100%) de uma série de 11 ésteres derivados do ácido mandélico (Kubinyi 1993a) que pode ser descrita pelo modelo parabólico representado na equação I.5.4,

$$\log RA = -0{,}189\ (\pm 0{,}09)(\log P)^2 + 1{,}556\ (\pm 0{,}56) - 1{,}438 \qquad \text{Equação I.5.4}$$

$$\text{valor ótimo } \log P = 4{,}15$$

$$(n = 11;\ r = 0{,}915;\ s = 0{,}298)$$

Onde: P - coeficiente de partição;
valor ótimo de logP – valor de logP para se obter o maior valor de log RA;
log RA é a resposta biológica do composto**.**

<u>Modelo Bilinear</u>

Embora o modelo parabólico descreva os fenômenos ocorridos no sistema biológico, observou-se que tanto o ramo ascendente como o descendente se assemelhavam ao modelo linear (Kubinyi & Kehrhahn 1978). Verificou-se que o modelo matemático deveria ser composto de uma combinação entre os modelos linear e o parabólico, denominado modelo bilinear (Kubinyi 1993a).

O modelo bilinear representado pela equação I.5.5, está fundamentado na probabilidade de compostos biologicamente ativos atingirem seu sítio de ação, considerando-se um sistema multicompartimentado (Kubinyi 1993a).

Assim, a discrepância observada entre os modelos linear e parabólico poderia ser atribuída a diversos fatores, entre outros: cinética do transporte da droga, distribuição da

droga em diferentes compartimentos do sistema biológico, espaço limitado para as interações de grupos hidrofóbicos no sítio de ação, efeitos alostéricos, formação de micelas, princípio de ocupação mínima do receptor (Hansch *et al.* 1990; Abd 1998).

$$\log 1/C = a \log P + b \log(\beta P + 1) + c\,\sigma + d\,E_s + r \qquad \text{Equação I.5.5}$$

Onde:
log 1/C é a resposta biológica do composto;
P – coeficiente de partição;
β - relação entre volumes das fases orgânica e aquosa;
σ - constante eletrônica de Hammett;
E_s – parâmetro estérico de Taft;
a, b, c – mede a influência de cada parâmetro sobre a resposta biológica (coeficiente determinado pela análise de regressão);
C – Valor da concentração molar do composto que produz um determinado efeito biológico;
r – constante.

Como exemplo, descreve-se que a atividade antifúngica (IC_{50}) de uma série de 15 aminas alifáticas avaliada contra *Rhinocladium beurmanni* (Unger 1984)pode ser descrita pelo modelo bilinear representado pela equação I.5.6.

$$\log 1/C_{50} = 0{,}944\ (\pm 0{,}01)\log P - 2{,}347\ (\pm 0{,}05)\ \log(\beta P+1) - 0{,}053\ (\pm 0{,}05) \qquad \text{Equação I.5.6}$$

$\log \beta = -5{,}787$ valor ótimo de $\log P = 5{,}62$

(n =15; r = 1,000; s=0,031; F=7945)

Onde: P – coeficiente de partição;
β - relação entre volumes das fases orgânica e aquosa;
valor ótimo de logP – valor de logP para se obter o maior valor de $\log 1/IC_{50}$;
$\log 1/IC_{50}$ é a resposta biológica do composto**.**

I.5.1. Modelo de Free-Wilson e Abordagem Mista

A abordagem de Free-Wilson (Free & Wilson 1964; Kubinyi 1988) é um verdadeiro modelo de relação estrutura-atividade. Uma Variável Indicadora é gerada para cada característica estrutural que difere de um composto referência, escolhido arbitrariamente. Esta variável assume valores iguais a 1 e/ou a 0, indicando a presença/ausência desse substituinte ou dessa característica estrutural. Os valores dos coeficientes de regressão obtidos para cada variável indicadora são as contribuições relativas de cada elemento

estrutural correspondente, para a atividade biológica. “Modelo matemático”, “modelo de aditividade” e “abordagem de novo” são sinônimos para o método de Free-Wilson (Kubinyi 1993a).

O modelo de Free-Wilson pode ser expresso pela equação I.5.1.1, na qual a_{ij} é a contribuição do substituinte X_i na posição **j** e, μ é o valor de atividade biológica (teórico) de um composto da série (referência).

$$\mathbf{\log\ 1/C = \sum a_{ij} + \mu} \qquad \text{Equação I.5.1.1}$$

Onde: $\mathbf{a_{ij}}$ - contribuição do substituinte X_i na posição j;
μ - valor de atividade biológica de um composto da série (referência);

A abordagem de Free-Wilson aplicada a uma série de 22 α-bromo-fenetilaminas com valores de atividade anti-adrenérgica (Graham & Karrar 1963) gera a equação equação I.5.1.2.

$$\log 1/ED_{50} = -0{,}301(\pm0{,}50)\,[m\text{-}F] + 0{,}207(\pm0{,}50)\,[m\text{-}Cl] + 0{,}434\,(\pm0{,}27)\,[m\text{-}Br] + \qquad \text{Equação I.5.1.2.}$$
$$+ 0{,}579\,(\pm0{,}50)\,[m\text{-}I] + 0{,}454(\pm0{,}27)\,[m\text{-}Me] + 0{,}340(\pm0{,}30)\,[p\text{-}F] +$$
$$+0{,}768(\pm0{,}30)\,[p\text{-}Cl] + 1{,}020(\pm0{,}30)\,[p\text{-}Br] + 1{,}429(\pm0{,}50)\,[p\text{-}I] +$$
$$+1{,}256(\pm0{,}33)\,[p\text{-}Me] + 7{,}821(\pm0{,}27)$$
$$(n = 22;\ r = 0{,}969;\ s = 0{,}194;\ F = 16{,}99)$$

Onde: [m-F] – indica a presença ou a ausência do átomo de flúor na posição meta;
[m-Cl] – indica a presença ou a ausência do átomo de cloro na posição meta;
[m-Br] – indica a presença ou a ausência do átomo de bromo na posição meta;
[m-I] – indica a presença ou a ausência do átomo de iodo na posição meta;
[m-Me] – indica a presença ou a ausência do grupo metil na posição meta;
[p-F] – indica a presença ou a ausência do átomo de flúor na posição para;
[p-Cl] – indica a presença ou a ausência do átomo de cloro na posição para;
[p-Br] – indica a presença ou a ausência do átomo de bromo na posição para;
[p-I] – indica a presença ou a ausência do átomo de iodo na posição para;
[p-Me] – indica a presença ou a ausência do grupo metil na posição para.

A análise de Hansch e o método de Free-Wilson diferem em suas aplicações, porém eles são estritamente relacionados (Kubinyi 1990).

Devido as relações entre as análises de Hansch e o modelo de Free Wilson, variáveis indicadoras foram incluídas nas análises de Hansch (Martin & Lynn 1971). Os dois modelos podem ser combinados em uma abordagem mista, em uma forma linear ou

não linear (equação I.5.1.3) o que oferece vantagens sobre as duas abordagens e aumenta a aplicabilidade das relações quantitativas entre estrutura-atividade.

$$\log 1/C = k_1 \Phi_1 + k_2 \Phi_2 + ... + k_n \Phi_n + \sum a_i + c = \sum k_j \Phi_j + \sum a_i + c \qquad \text{Equação I.5.1.3}$$

Onde: Φ_j - propriedades físico-quimicas
a_{ij} - contribuição do substituinte X_i na posição j;
c - constante;

Um exemplo de abordagem mista é representada pela equação I.5.1.4, onde os substituintes X da estrutura apresentada na figura I.5.1.1 são descritos por parâmetros físico-químicos, respectivamente π, σ, e E_S, enquanto que os substituintes Y são descritos por parâmetros de Free Wilson [I], para o grupo iodeto e [Me] para o grupo metila. (HANSCH *et al.* 1990;KUBINYI 1976).

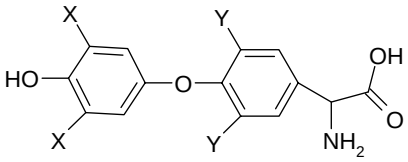


Figura I.5.1.1. Estrutura dos derivados de tironina.

$$\log A = 1{,}699(\pm 0{,}34)\ \pi_x - 2{,}059(\pm 0{,}70)\ \sigma_x + 1{,}713(\pm 0{,}39)\ E_s' + 0{,}234(\pm 0{,}20)\ [I] - 0{,}532(\pm 0{,}26)\ [Me] - 1{,}792$$

(n = 25; r = 0,943; s = 0,170; F=30,37) Equação I.5.1.4

Onde: π - parâmetro hidrofóbico de Hansch-Fujita dos substituintes na posição X;
σ - parâmetro eletrônico de Hammett dos substituintes na posição X;
E_s' – parâmetro estérico de Taft dos substituintes na posição X;
[m-I] – indica a presença ou a ausência do átomo de iodo na posição Y;
[m-Me] – indica a presença ou a ausência do grupo metil na posição Y.

I.6. Recomendações Para Obtenção de Modelos de QSAR

Na literatura, Unger e Hansch (Unger & Hansch 1973) formularam regras para a obtenção de modelos válidos que, devido a sua validade geral foram revistos anteriormente (Kubinyi 1993a), apresentados no grupo (Pires 1998) e, sumarizados, a seguir:

- Seleção de variáveis independentes. Uma grande faixa de diferentes parâmetros dever ser testados, como por exemplo, log P ou π, σ, MR e parâmetros estéricos. Parâmetros obtidos por cálculos de orbitais moleculares e Variáveis Indicadoras não devem ser superestimados. Topliss verificou que as chances de obter um coeficiente de correlação maior que 0,9 não só aumentavam com o número de variáveis incluídas na equação como também com o número de variáveis das quais as diferentes combinações são selecionadas (Topliss & Costello 1972; Topliss & Edwards 1979).
- Os parâmetros selecionados para a "melhor equação" devem apresentar correlação não significativa (ou seja, o coeficiente de intercorrelação r deve ser menor do que 0,6 – 0,7. Exceções são combinações dos termos lineares e quadráticos como log P e $(\log P)^2$, os quais são geralmente altamente intercorrelacionados, com $r > 0,9$) (Kubinyi 1993a).
- Justificativa da escolha das variáveis independentes. Todos os parâmetros devem ser validados pela análise de regressão que é um procedimento estatístico adequado. A melhor equação é geralmente aquela com o mais baixo desvio padrão e, com todos os termos significantes (indicado pelos intervalos de confiança de 95% ou pelo teste seqüencial F). Alternativamente, a equação com o maior valor de F pode ser selecionada como a melhor equação.
- Havendo mais de uma equação (aproximadamente) com mesmo nível de confiança estatística, a mais simples com menos parâmetros deve ser escolhida.
- Número de termos. Deve-se ter no mínimo 5 a 6 dados (compostos) por variável para se evitar correlações ao acaso. (Esta regra se aplica apenas a conjuntos de dados de tamanho intermediário; para conjuntos de dados pequenos, mais parâmetros podem ser permitidos se eles estiverem

baseados em um modelo razoável. Para grandes conjuntos de dados, por exemplo n > 30, essa recomendação leva a equações as quais incluem muitas variáveis).

- Modelo Qualitativo. É mais importante ter um modelo qualitativo o qual seja consistente com o processo físico-orgânico-químico do processo em consideração.

Considerando-se as equações de correlação, com o aumento de variáveis, a regressão linear múltipla (MLR), válida para o modelo de Hansch, tornou-se inadequada. Assim, nos casos onde há mais variáveis que amostras, aumentam-se as chances de se propor um falso modelo com um bom ajuste, ou seja, não se obtém uma correlação verdadeira, mas sim apenas uma coincidência (Topliss & Costello 1972; Topliss & Edwards 1979).

As regras e as condições foram, então, revistas e reformuladas (Kubinyi 1994) para se obter correlações válidas. De acordo com as novas propostas, recomenda-se (Kubinyi 1994) que

- variáveis significativas devem ser selecionadas;
- a correlação e cada individual termo no modelo de regressão deve ser justificado por parâmetros estatísticos apropriados;
- para resultados semelhantes, o modelo com um menor número de variáveis deve ser escolhido;
- não muitas variáveis devem ser incluídas no modelo final;

Além destas recomendações, outros procedimentos de validação foram propostos (Lindgren *et al.* 1991; Tropsha *et al.* 2003), como por exemplo: Validação cruzada (*cross-validation*) (Lindgren *et al.* 1991; Cramer *et al.* 1988), teste externo (Golbraikh & Tropsha 2002) e, embaralhar os valores da variável dependente.

Ao longo dos anos, vários parâmetros estatísticos foram desenvolvidos (Kubinyi 1994). Dentre eles, citam-se aqueles para justificar: a correlação, os termos individuais presentes, a robustez, e, a predição da equação de regressão.

Tabela 1.6.1. Alguns parâmetros estatísticos selecionados para avaliar a validade estatística das correlações/modelos gerados.

Expressão Matemática		Equação	Literatura
$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	- SS_T é a soma total dos quadrados; - y_i é o valor da variável dependente observado; - $\bar{y}$ é o valor médio da variável dependente observado na série;	Equação. I.6.1	(Barros Neto *et al.* 2002)
$SS_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	- SS_R é a soma dos quadrados dos resíduos; - y_i é o valor da variável dependente observado; - $\hat{y}_i$ é o valor calculado da variável dependente através do modelo de regressão;	Equação. I.6.2	
$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i^{*})^2$	- PRESS é a soma dos quadrados dos erros residuais de predição; - y_i é o valor observado da variável dependente do composto da série de treinamento, o qual não participou da equação de regressão; - $\hat{y}_i^{*}$ é o valor calculado da variável dependente através do modelo de regressão do respectivo composto;	Equação. I.6.3	(Lindgren *et al.* 1991) (Baroni *et al.* 1993)
$RMSE = \sqrt{\frac{SS_R}{n-p-1}}$	- *RMSE* é a raiz da média quadrática dos erros; - SS_R é a soma dos quadrados dos resíduos; - n é o número de amostras; - p é o número de variáveis;	Equação. I.6.4	(Livingston e 1995) (Kubinyi 1994)
$SEP_{cv} = \sqrt{\frac{PRESS}{n}}$	- SEP_{cv} é a raiz da média quadrática dos erros de predição do *cross-validation;* - PRESS é a soma dos quadrados doserros residuais de predição; - n é o número de amostras;	Equação. I.6.5	(Baroni *et al.* 1993)
$SEP = \sqrt{\frac{SS_R^{*}}{n'}}$	- *SEP* é a raiz da média quadrática dos erros de predição*;* - SS_R^{*} é a soma dos quadrados dos resíduos da série de teste ; - n' é o número de amostras da série de teste;	Equação. I.6.6	(Livingston e 1995)
$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_{pi} - \hat{y}_{pi})^2}{\sum_{i=1}^{n} (y_{pi} - \bar{y}_p)^2}$	- r^2 é o coeficiente de correlação entre os y calculados e os y observados; - y_{pi} é o valor calculado da variável dependente do composto através do modelo; - $\hat{y}_{pi}$ é o valor calculado da variável dependente através da equação da reta de ajuste entre os valores observados e os valores calculados de pIC50; - $\bar{y}_p$ é a média dos valores calculados de y pelo modelo;	Equação I.6.7	(Barros Neto *et al.* 2002) (Livingston e 1995)
$Q^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$	- Q^2 é o coeficiente de predição; - y_i é o valor observado da variável dependente do composto da série de teste; - $\hat{y}_i$ é o valor calculado através do modelo de regressão do respectivo composto; - $\bar{y}$ é a média dos valores observados da variável dependente na série de teste;	Equação I.6.8	(Baroni *et al.* 1993)
$Q_{cv}^{\ 2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i^{*})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$	- $Q_{cv}^{\ 2}$ é o coeficiente de predição pelo método de *full cross-validation*; - y_i é o valor observado da variável dependente do composto da série de treinamento o qual não participou da equação de regressão; - $\hat{y}_i^{*}$ é o valor da calculado variável dependente através do modelo de regressão do respectivo composto; - $\bar{y}$ é a média dos valores observados da variável dependente ;	Equação I.6.9	(Cramer *et al.* 1988) (Baroni *et al.* 1993)

Alguns parâmetros desenvolvidos recentemente, estão apresentados nesta dissertação, no item *Metodologias Recentes* (item I.9).

I.7. Relações Quantitativas entre Estrutura química e Atividade Biológica em Três Dimensões: *QSAR 3D*

Os métodos desenvolvidos em *QSAR* e em *Modelagem Molecular* são atualmente aplicados, *simultaneamente*, para descrever de modo quantitativo as interações entre o composto e o sistema biológico, considerando seus aspectos tridimensionais. Esta abordagem é denominada *QSAR-3D* e complementa a primeira (Kim & Martin 1991; Kim 1992). A abordagem *QSAR-3D* considera os aspectos estereoquímicos e topológicos das interações ligante-alvo(receptor) biológico. Estes contribuem e, na verdade, muitas vezes, são os fatores determinantes da resposta/atividade biológica

As primeiras abordagens propostas visando o planejamento de ligantes assistido por computador, quando a estrutura do receptor é conhecida, foram apresentadas nos programas DOCK e GROW (Shoichet *et al.* 1993; Moon & Howe 1991). Na sua versão original executa, em um banco de dados de estruturas 3D, busca por ligantes que se ajustam a uma cavidade somente considerando a por complementariedade geométrica. Por outro lado, o programa GROW inicia com um único fragmento-semente, por exemplo, um grupo amida, capaz de interagir com o sítio ativo e, então diferentes aminoácidos, em diferentes conformações, são adicionados a este fragmento. Apenas os melhores candidatos são selecionados e o processo é repetido várias vezes, até que determinado peptídeo de certo tamanho tenha sido gerado.

No programa **LUDI** (Bohm 1992), após a definição da região do sítio ativo pelo usuário, identifica-se automaticamente todos os grupamentos doadores e aceptores de ligação de hidrogênio, bem como as áreas hidrofóbicas, alifáticas e aromáticas. O programa gera, a seguir, vetores e regiões no espaço nas quais grupos complementares do ligante devem ser colocados. Numa próxima etapa, **LUDI** executa uma busca por potenciais ligantes em bancos de estruturas 3D de moléculas pequenas e médias. Cada candidato é, a seguir, testado em todas as diferentes orientações possíveis e modos de interações. Novas versões do programa já consideram a flexibilidade conformacional dos ligantes em

potencial. Depois de uma avaliação aproximada, contando o número de interações e, considerando-se as superposições desfavoráveis da superfícies de Van der Waals, os candidatos remanescentes são priorizados por uma função simples, mas eficiente. Esta função estima a energia de interação com base na energia de formação de ligações de hidrogênio, áreas de contato hidrofóbico e o número de ligações passíveis de rotação do ligante (Bohm 1994).

A análise de forma molecular (*MSA*) é um formalismo designado para ser de utilidade prática no planejamento de fármacos com auxílio de computador (*CADD, Computer-Aided Drug Design*) (Kubinyi 1993b), quando a estrutura do receptor não é conhecida. A primeira etapa em *MSA* envolve a análise conformacional de cada composto investigado. A conformação ativa (bioativa) de um ligante é geralmente entendida como sendo a conformação adotada pelo ligante ao se ligar ao receptor. No entanto, a conformação bioativa, como qualquer outra propriedade estrutural importante de um composto biologicamente ativo em *QSAR*, corresponde a conformação envolvida na etapa limitante e, portanto, que está controlando o aparecimento da resposta biológica. Freqüentemente, esta etapa limitante é a formação do complexo ligante-receptor, mas poderia também ser, por exemplo, um passo de ativação/desativação metabólica, um passo de transporte entre membranas, entre outros (Lopez *et al.* 1990; Kubinyi et al. 1998).

Considerando-se que *MSA* foi proposta para casos em que a geometria do receptor não é conhecida, toda informação sobre a conformação bioativa deve provir das atividades biológicas observadas e, das correspondentes propriedades conformacionais computadas para os ligantes. As etapas seguintes envolvem a seleção do composto de referência e, a superposição ou alinhamento das moléculas**.**

Várias abordagens para o alinhamento das estruturas visando comparar similaridades moleculares têm sido descritas (Klebe & Abraham 1993). Algumas abordagens baseiam-se nas informações estruturais, que são diretamente retiradas das disposições atômicas. Desta forma, todas as moléculas que possuem grupos funcionais semelhantes (grupos farmacofóricos), podem formar um arranjo tridimensional semelhante (arranjo farmacofórico).

O método mais intuitivo de alinhamento de moléculas segue um ajustamento interativo de ângulos de torsão com subseqüente superposição por método de mínimos

quadrados de átomos-chave nos grupos farmacofóricos, utilizando-se computação gráfica. Neste método, pode haver, no entanto, a predisposição do modelador na seleção dos grupos farmacofóricos e no que considera ajuste razoável entre as moléculas.

Mais recentemente, algumas abordagens consideram as superfícies de potencial para determinar orientação comum para comparação de moléculas (Klebe & Abraham 1993).

I.7.1. Exemplos de Abordagens QSAR Envolvendo Muitas Variáveis

Análise Comparativa dos Campos Moleculares *(CoMFA, Comparative Molecular Field Analysis)*

A abordagem CoMFA correlaciona a atividade biológica com propriedades estruturais em três dimensões representadas pelos campos estéricos e eletrostáticos dos compostos, bem como por propriedades. E, desde sua publicação (Cramer *et al.* 1988) a análise comparativa de campos moleculares (***CoMFA***) rapidamente se tornou uma das ferramentas em *QSAR-3D* mais poderosas e mais utilizadas (Clark & Cramer 1993)

No procedimento CoMFA padrão, série de moléculas são selecionadas de acordo com a atividade e modo de ação. As coordenadas tridimensionais destas moléculas são utilizadas nos cálculos de suas propriedades estruturais e eletrônicas, de acordo com métodos computacionais de vários níveis de sofisticação. A seguir, estas moléculas são sobrepostas em suas supostas conformações bioativas, obtendo-se um padrão único de alinhamento (Kubinyi *et al.* 1998). A partir daí, um retículo tridimensional é colocado sobre as moléculas sobrepostas, de forma a estabelecer uma caixa regular de vários Ângstrons a mais que o volume das estruturas sobrepostas. Em seguida, os campos estéricos e eletrostáticos ao redor destas moléculas são avaliados através de átomos de teste (*probe atoms*) no retículo retangular de pontos. Átomos de teste ou grupos são dispostos alternadamente nas extremidades do retículo para avaliar as interações de cada campo. Assim, um carbono neutro quantifica interações de Van der Walls, uma dada carga avalia as interações eletrostáticas, enquanto doadores ou receptores de átomos de hidrogênio avaliam as possibilidades destes tipos de interações. Este procedimento é realizado para

cada molécula e em cada ponto da grade, onde as funções matemáticas utilizadas são os potenciais respectivamente de Lennard-Jones para as interações de Van der Waals e o de Coulomb para as interações eletrostáticas.

A partir da série original de compostos formam-se dois sub-grupos, quais sejam: um *grupo de treinamento* do modelo CoMFA e um *grupo de moléculas para avaliar as predições* do modelo gerado. O resultado da avaliação dos campos em cada ponto do retículo é colocado em uma tabela. A tabela *CoMFA-QSAR* gerada é analisada pelo método *PLS* (***P***artial ***L***east ***S***quares), que não é sensível a co-linearidade dos descritores de campo e fornece o modelo estatístico significativo.

Em *CoMFA avançado* o potencial lipofílico é também incorporado nos campos de CoMFA, para que se levem em conta também as interações hidrofóbicas (Gaillard *et al.* 1994), bem como os componentes entrópicos não adequadamente descritos pelos campos estéricos e eletrostáticos. Faz-se, a seguir a validação do modelo CoMFA gerado (Lopez *et al.* 1990).

A equação de *QSAR* obtida por CoMFA é usualmente resumida de modo gráfico (mapas de contorno 3D). Deste modo, os poliedros coloridos gerados delimitam volumes ao redor de pontos do retículo dos quais os coeficientes da equação de *QSAR* mostram, particularmente, alta associação entre as diferenças nas intensidades dos campos moleculares e a atividade biológica. Tipicamente, haverá dois níveis de contorno para cada campo molecular. Estes realçam as regiões de maior associação, as mais positivas e aquelas mais negativas. Estes mapas são, freqüentemente, bastante úteis na sugestão de novos compostos com probabilidade de ter maiores valores de atividade biológica.

Análise Comparativa de Índices de Similaridades Moleculares (CoMSIA, *Comparative Molecular Similarity Indices Analysis)*

A análise comparativa dos índices de similaridade molecular (*CoMSIA. Comparative Molecular Similarity Indices Analysis*) foi introduzida,mais recentemente, como uma aproximação alternativa para se realizar os estudos de *QSAR-3D*. (Klebe *et al.* 1994).

Este método permite considerar várias outras propriedades físico-químicas. (Lopez *et al.* 1990). E, os mapas de contorno resultantes podem ser interpretados.

Este método corrige algumas deficiências inerentes devido às funções potenciais de Lennard-Jones e de Coulomb utilizados na versão original do CoMFA. Estes dois potenciais são semelhantes à superfície de Van der Waals e produzem singularidades nas posições atômicas. Como conseqüência, a energia potencial expressa nos pontos da grade nas proximidades da superfície sofrem mudanças drásticas. Esta região é justamente a que contém as descrições importantes para a análise de *QSAR*. Para permitir valores de energia inaceitavelmente grandes, a evolução dos potenciais é normalmente restrita a regiões fora da molécula e requer a definição de alguns valores de corte determinados arbitrariamente.

Em *CoMSIA*, a similaridade é expressa em termos de diferentes propriedades físico-químicas, a saber: ocupação estérica, cargas atômicas parciais, hidrofobicidade local e doadores e aceptores de ligação de hidrogênio. Utilizando-se um átomo de prova comum, pode-se calcular os índices de similaridade para um grupo de moléculas previamente alinhadas com espaçamento de grade regular. Uma das maiores vantagens do *CoMSIA* em relação ao *CoMFA* é a melhor habilidade para se visualizar e interpretar as correlações obtidas em termos das contribuições dos campos.

Aspectos teóricos, metodológicos, vantagens, limitações e aplicações recentes de sucesso são disponíveis na literatura, tanto para CoMFA como CoMSIA (Kubinyi *et al.* 1998b) (Kim *et al.* 1998). A maioria dos modelos *CoMFA* e *CoMSIA* foram aplicados para descrever as interações ligante-proteína, sendo estas descritas por constantes de afinidade ou constante de inibição. A aplicação de CoMFA para descrever dados de sistemas *in vivo* não são recomendadas.

I.8. *Data-Mining* (Gasteiger *et al.* 2003)

Decorrente deste desenvolvimento tecnológico nos últimos anos, o avanço na aquisição de dados para os sistemas tanto químicos como biológicos gerou um grande número de informações. Como consequência, nos últimos anos, procuram-se ferramentas, fundamentalmente matemáticas (Todeschini *et al.* 2004; Baroni *et al.* 1993; Kubinyi 1994), que permitam decodificar este volume imenso de informações, em termos estruturais e biológicos, ou seja, necessitou-se de criar um processo para analisar os dados e identificar/diferenciar as características e relações contidas neste. Estas abordagens, que se propõem extrair conhecimento de uma grande série de dados com o objetivo de fazer

predições de novos eventos é denominado na língua inglesa como *data mining* (Gasteiger *et al.* 2003).

Considerando-se a seleção de variáveis e, de compostos disponíveis em extensos bancos de dados, diversos algoritmos foram também utilizados e/ou desenvolvidos, como por exemplo o algoritmo genético para a primeira (Leardi *et al.* 1992; Leardi 1994). Estes procedimentos devem, em princípio, gerar modelos que se apliquem não somente à série de treinamento, ou seja, devem gerar modelos robustos. Pode-se citar como exemplo de sucesso, a regra de seleção de compostos proposta por Lipinski (Lipinski *et al.* 1997). Nesta regra, os compostos são selecionados considerando-se as faixas de variação das propriedades que são importantes para a farmacocinética do composto.

Pré-tratamento dos dados

O pré-tratamento de dados é recomendado ao se gerar um grande número de variáveis (Livingstone 1995), excluindo-se aquelas que não fornecem informações relevantes sobre o sistema, no entanto, contribuindo apenas, para aumentar a quantidade de dados e de ruídos a serem tratados.

Na literatura sugere-se (Livingstone 1995) que, uma maneira de se reduzir os dados é excluir as variáveis com valores constantes e aquelas com apenas um valor diferente na série. Tal situação ocorre quando há alguma propriedade mal escolhida para a série de compostos, ou seja, a variável é pouco representativa para aquela série. Atualmente, existem alguns pacotes de *softwares* que facilmente (com baixo custo computacional) identificam e/ou removem estas variáveis. Após a remoção destas, o escalonamento das variáveis e a matriz de correlação (análise da intercorrelação das variáveis restantes) podem ser então feitos (Livingstone 1995).

Deste modo, numa determinada série de dados, uma matriz de correlação pode ser construída entre cada par de variáveis. Em seguida, através da inspeção da matriz de correlação pode-se verificar e avaliar as características altamente correlacionadas, na série. A escolha do valor do nível máximo de corte entre as variáveis correlacionadas depende do método de análise aplicado a estas.

Alguns métodos, como por exemplo, a regressão linear múltipla – MLR, são sensíveis à presença de colinearidade na série de dados, podendo-se observar *overfit*. (ajuste em excesso). Considera-se que uma equação de regressão linear múltipla pode ser entendida como sendo uma série de variáveis, que explicam alguma ou toda variação da variável dependente (y). Assim sendo, se as variáveis independentes são correlacionadas em pares (apresentam colinearidade) ou em forma de combinações lineares (multicolinearidade), então diferentes combinações–podem explicar a mesma variação (grandeza e natureza) na variável dependente. A presença de duas variáveis colineares em uma equação pode gerar dados estatísticos de ajuste aparentemente válidos. O modelo gerado, porém, apresenta valores de coeficientes de regressão instáveis e, conseqüentemente acompanhados dos respectivos erros padrões altos (Livingstone 1995).

Efeito análogo de *overfit* pode ser observado ao se incluir muitas variáveis em uma equação de regressão. Desta forma, adiciona-se ruído ao modelo e, a equação resultante apresenta um bom ajuste apenas para as amostras aplicadas ao treinamento, apresentando um baixo poder de predição e de ajuste para outras amostras.

Métodos envolvendo escalonamento de variáveis tem como objetivo remover qualquer ponderação decorrente unicamente das unidades que são usadas para expressar uma variável, além de facilitar cálculos posteriores. Há diversos métodos de escalonamento (Geladi & Kowalski 1986), uma delas é denominada normalização. Nesta, o valor mínimo de uma variável é atribuído como zero e, os outros valores são subtraídos pelo menor valor original da variável e divididas pelo intervalo total (equação I.8.1).

$$X'_{ij} = \frac{X_{ij} - X_j(\min)}{X_j(\max) - X_j(\min)}$$

Equação I.8.1

Onde: X'_{ij} é o novo valor escalonado do composto *i* da variável *j*;

Estes valores estão escalonados entre zero e um. A principal desvantagem desta técnica de escalonamento é a sensibilidade aos *outliers*. Uma outra forma de escalonamento, que é menos sensível aos *outliers* é conhecido como *autoscaling* . Neste, a

valor da média é subtraída dos valores da variável e os valores resultantes são divididos pelo desvio padrão, (Equação I.8.2).

$$X'_{ij} = \frac{X_{ij} - \overline{X}_j}{s_j}$$

Onde: X'_{ij} é o novo valor escalonado do composto *i* da variável *j*; Equação I.8.2

$\overline{X}_j$ é o valor da média

s_j é o valor do desvio padrão;

Sendo s_j expresso pela equação Equação I.8.3.

$$s_j = \sqrt{\left(\sum_{i=1}^{n} \frac{\left(x_{ij} - \overline{x}_j\right)^2}{n-1} \right)}$$

Equação I.8.3

Onde: n é o número de amostras;

As variáveis autoescalonadas apresentam os valores da média igual a 0 e do desvio padrão igual a 1. Elas são menos sensíveis a valores extremos devido a sua média centrada. Foi verificado que o autoescalonamento fornecem bons resultados para as análises de *PLS* (Otto & Wegscheider 1985).

Um outro aspecto a ser considerado na manipulação e tratamento de grande número de dados se refere à homogeneidade na distribuição dos dados na população estudada e a presença de "*outliers*". Estes afetam as análises de regressão e, sua presença pode ocasionar erros na análise de regressão.

Obtenção de Modelos a Partir de um Grande Número de Dados

Modelos obtidos a partir de um grande número de dados, em geral apresentado como banco de dados, utilizam funções matemáticas, desenvolvidas e/ou aplicadas, inicialmente, em Quimiometria, como PCA (*Principal Component Analysis*) (Wold *et al.*

1987), *PLS* (*Partial Least Square*) (Geladi & Kowalski 1986; Baroni *et al*. 1993), KNN (*k-nearest-neighbour*) (Golbraikh & Tropsha 2002), redes neurais (Anzali *et al.* 1996).

Com relação ao número de informações utilizado para gerar o modelo, algumas das ferramentas desenvolvidas, como o PCA (Wold *et al.* 1987), diminuem a dimensionalidade dos dados, minimizando ao máximo a perda de informação estatisticamente significativas e, geram componentes ortogonais. A utilização do PCA na regressão linear múltipla resultou no procedimento denominado regressão de componentes principais, PCR (Geladi & Kowalski 1986), útil nos caso onde o número de variáveis é maior do que o de amostras. Uma outra técnica bastante utilizada na área de *QSAR* é o ***PLS – Partial Least Square***, (mínimos quadrados parciais). *PLS* (*Partial Least Square*, sigla em inglês para o método dos mínimos quadrados parciais) torna-se uma boa alternativa, por ser um método mais robusto do que o clássico *MLR* e o *PCR* (Clark & Cramer 1993).

Aplicando-se análises de regressão linear múltipla (*MLR*) para séries que apresentam um número muito superior de variáveis em relação de amostras (p>>n), aumentam-se as chances de se achar um falso modelo com um bom ajuste, ou seja, se obtém não uma correlação verdadeira, mas sim apenas uma coincidência (“correlação ao acaso”) (Topliss & Costello 1972; Topliss & Edwards 1979). Assim, esta chance aumenta, quanto maior o número de variáveis em relação ao número de amostras. Deve-se considerar que os resultados observados se referem a estudos realizados para números aleatórios (Topliss & Costello 1972; Topliss & Edwards 1979) enquanto para números reais, poder-se-ia observar comportamento diferente.

Estudos semelhantes utilizando, porém, *PLS* mostraram que a “correlação ao acaso” não é significativa para este método, pelo contrário, ao se aumentar o número de variáveis, os valores dos coeficientes de predição utilizando o *full cross-validation* diminuem (Clark & Cramer 1993). Neste estudo, também foram utilizados números randômicos, verificando-se que o valor máximo do coeficiente de predição ocorre quando os números de variáveis e de amostras são iguais. Estas duas características do *PLS* podem ser explicadas pelo comportamento do algoritmo. Assim, tanto no *PCA* (Wold *et al.* 1987) como no *PLS* (Geladi & Kowalski 1986; Wold *et al*. 2001) a operação fundamental é repetir a extração de componentes, ou seja, uma combinação linear de todas as variáveis sob consideração,

mantendo-se cada novo componente, ortogonal a qualquer outro componente previamente extraído.

No *PCA*, no entanto, cada novo componente apenas engloba a quantidade de variância das variáveis independentes remanescentes, enquanto no *PLS*, os componentes são extraídos ao mesmo tempo, de ambas séries de variáveis: dependente(s) e independentes. Este procedimento de extração de componentes é realizado, visando expressar a quantidade de variância comum em ambas séries de variáveis: dependente(s) e independentes. Assim, *PLS* tenta em uma única etapa correlacionar a(s) variável(is) dependente(s) com todas as variáveis independentes. Em contraste, a regressão multilinear, em cada etapa identifica uma variável independente que melhor correlaciona com a variável dependente.

O *PLS* como também o *PCR* (Geladi & Kowalski 1986; Wold *et al.* 2001), não são sensíveis à presença de colinearidade das variáveis independentes na série de dados. No entanto, para facilitar a interpretação e os cálculos no tratamento *PLS*, numa primeira aproximação recomenda-se (Livingstone 1995) retirar, já na série original e na medida do possível, as variáveis mais correlacionadas.

I.9. Metodologias Recentes

Apesar do conhecimento de diversos coeficientes estatísticos e de diversos métodos de seleção de modelos e, conseqüentemente das variáveis, ainda são encontrados estudos na literatura recente (Todeschini *et al.* 2004; Mattioni & Jurs 2002; Golbraikh & Tropsha 2002; Gasteiger *et al.* 2003), propondo-se novas ferramentas para avaliar e garantir a qualidade de predição do modelo bem como a elucidação de determinado mecanismo a partir do modelo gerado (Golbraikh & Tropsha 2002; Gasteiger *et al.* 2003). Esta necessidade aparece, pois encontram-se com freqüência modelos que apresentam bom ajuste mas baixo poder de predição. Estes são algumas vezes resultados de uma correlação ao acaso e geralmente apresentam características indesejáveis como multicolinearidade, *overfitting* e , inclusão de variáveis que são apenas "*ruídos*" (Todeschini *et al.* 2004).

I.9.1. Regra QUIK

A regra QUIK (Q^2 *Under Influence of K*) proposta em 1998 (Todeschini *et al.* 1999; Todeschini *et al.* 2004) é um simples critério que permite a rejeição de modelos com alta colinearidade, o que pode ocasionar uma correlação ao acaso (Topliss & Edwards 1979). A regra QUIK é baseada no índice de correlação K (Todeschini *et al.* 1999; Todeschini 1997) que mede a correlação total de uma série de variáveis expresso na equação I.9.1.1.

$$K = \frac{\sum_j \left| \left(\lambda_j / \sum_j \lambda_j \right) - (1/p) \right|}{2(p-1)/p} \qquad \text{Equação I.9.1.1.}$$

Onde: j = 1, ...,p e
$0 \leq K \leq 1$
λ_j são os auto-valores obtidos da matriz de correlação da série de dados de **X**(n,p);
n o número de objetos;
p o número de variáveis;

Essa regra é derivada da suposição evidente que a correlação total em uma série é dada pelas variáveis **X** independentes mais a variável dependente **Y** (K_{XY}), e esta deve ser sempre maior que a correlação medida apenas entre as variáveis independentes (K_X).

Desta forma, a regra QUIK determina que apenas modelos com correlação entre as variáveis independentes mais a variável dependente K_{XY} maior que a correlação entre as variáveis independentes K_X podem ser aceitos (equação I.9.1.2).

$$K_{XY} - K_X < \delta K \rightarrow \text{rejeite o modelo} \qquad \text{Equação I.9.1.2.}$$

Onde: δK é um limite definido (entre 0,01 a 0,05);

O δK pode ser zero se deseja um limite menos rigoroso. De qualquer forma limites menores que zero não são permitidos, ou seja, a diferença entre $K_{XY} - K_X$ não deve ser negativa.

A regra QUIK demonstrou-se eficiente em evitar modelos com multicolinearidade sem poder de predição. De outro lado essa regra não é eficiente para evitar variáveis que são apenas ruídos, desde que estas variáveis não são correlacionadas, portanto apresentando

um valor de K_X baixo. Nesse caso, mesmo uma baixa correlação entre a variável dependente com as variáveis independentes pode ser considerada significante através desta regra (Todeschini *et al.* 1999; Todeschini *et al.* 2004).

Adicionalmente a regra QUIK, propôs-se calcular o índice de degeneração multivariada *D* (equação I.9.1.3). Nesta equação S_R, *S*, e S^+ correspondem ao índice de entropia relativa multivariada, ao índice de entropia multivariada, e ao índice de entropia total multivariada (equações I.9.1.4 a I.9.1.6). Estes índices medem a variabilidade contida numa série de dados. Nestas equações, *n* é o número de amostras, *p* é o número de variáveis independentes, n_x é o número de valores iguais presentes na mesma variável, e *K* é o índice de correlação multivariada definida na equação Equação I.9.1.1. (Todeschini *et al.* 1999).

$$D = \frac{S^+ - S}{S^+} = 1 - S_R \qquad \text{Equação I.9.1.3}$$

$$S_R = \frac{S}{S^+} \qquad \text{Equação I.9.1.4}$$

$$S = \left[1 + (p-1)(1-K)\log_2 n\right] \times \frac{\sum_{j=1}^{p}\left(-\sum_x \frac{n_x}{n}\log_2 \frac{n_x}{n}\right)}{p} \qquad \text{Equação I.9.1.5}$$

$$S^+ = p \log_2 n \qquad \text{Equação I.9.1.6}$$

I.9.2. Regra do Q^2 Assintótico

Um modelo significantemente estatístico deve ter uma pequena diferença entre o valor do coeficiente de correlação (r^2) e a abilidade preditiva (Q_{cv}^2). De fato diferenças marcantes entre os valores r^2 e Q_{cv}^2 (Todeschini *et al.* 2004) podem ser devidos ao

overfitting (fornecendo altos valores de r^2) ou por algum caso não predito (fornecendo baixos valores de Q_{cv}^2).

Mattioni e Jurs (Mattioni & Jurs 2002) propuseram uma função a qual contabiliza o custo na seleção do modelo, expresso na equação I.9.2.1:

$$\cos t = rms_T + 0{,}4\left|rms_T - rms_{CV}\right| \qquad \text{Equação I.9.2.1}$$

Onde: rms_T é a raiz quadrada da média dos valores dos erros da série de treinamento;
rms_{CV} é a raiz quadrada da média dos valores dos erros da série de teste;
o valor 0.4 é um parâmetro empírico de ponderação da diferença entre a habilidade de ajuste e de predição;

Com o objetivo de se evitar este parâmetro empírico de ajuste, foi proposto um critério como uma regra de exclusão baseado no critério no comportamento assintótico do Q_{cv}^2. Foi demonstrado que o Q^2 (Alan J.Miller 1990) é relacionado assintóticamente ao coeficiente de correlação (r^2), desta forma um valor assintótico de Q^2 pode ser calculado pela equação I.9.2.2. expressa:

$$Q^2{}_{ASYM} = 1 - \left(1 - r^2\right) \times \left(\frac{n}{n - p'}\right)^2 \qquad \text{Equação I.9.2.2}$$

Onde: n é o número de objetos;
p' é o número de parâmetros do modelo;

A regra do Q assintótico é baseada na diferença entre o valor do coeficiente de predição Q^2_{cv} e o valor do Q^2_{ASYM} expresso na equação I.9.2.3:

$$\text{se} \quad Q^2{}_{cv} - Q^2{}_{ASYM} < \delta Q \longrightarrow \text{rejeite o modelo} \qquad \text{Equação I.9.2.3}$$

Onde: δQ é o valor limite determinado;

Os autores desta regra assumiram que um modelo com um valor de coeficiente de predição Q^2_{cv} menor que uma quantidade δQ do valor do coeficiente de predição assintótico

Q^2_{ASYM} deve ser rejeitado. Um limite simples δQ pode ser zero, um limite menos rigoroso pode ser –0,005, um limite mais rigoroso poderia ser 0.005 (Todeschini *et al.* 2004).

I.9.3. Regras baseadas nas funções R^P e R^N.

Os objetivos das duas regras apresentadas a seguir, são os de detectar *"overfitting"* devido presença de variáveis no modelo que estão explicando a mesma parte da variação da variável dependente e/ou devido a presença de variáveis no modelo que são apenas "ruídos"(Todeschini *et al.* 2004).

Ambas as regras estão baseadas no parâmetro M_j o qual é obtido através da equação I.5.2.3.1.

$$M_j = \frac{R_{jy}}{R} - \frac{1}{p} \qquad \text{Equação I.9.3.1}$$

Onde: $-\frac{1}{p} \le M_j \le \frac{p-1}{p}$;

p é o número de variáveis independentes presentes no modelo;
R_{jy} é o valor do coeficiente de correlação absoluta entre a variável independente j e a variável dependente y;
R é o valor do coeficiente de ajuste do model;

Nesta equação está implícito que se todas as variáveis independentes contribuírem na mesma proporção para explicar a variação contida na variável dependente, esta porção será de 1/p para a correlação múltipla R.

Cada contribuição R_{jy}/R do modelo é comparada com o valor 1/p, com o objetivo de se avaliar a contribuição de uma única variável no modelo. Os valores positivos de M_j são utilizados para o cálculo de R^P expresso na equação I.9.3.2, e os valores negativos são utilizados para o cálculo de R^N expresso na equação I.9.3.3.

$$R^P = \prod_{j=1}^{p^+} \left(1 - M_j \times \left(\frac{p}{p-1}\right)\right) \qquad \text{Equação I.9.3.2}$$

Onde: $M_j > 0$;
$0 \leq R^P \leq 1$;
R^P é calculado através das variáveis p^+, responsáveis pelas diferenças positivas M_j;
p é o número de variáveis independentes presentes no modelo;

$$R^N = \sum_{j=1}^{p^-} M_j \qquad \text{Equação I.9.3.3}$$

Onde: $M_j<0$;
$-1 < R^N \leq 0$;
R^N é calculado através das variáveis p^-, responsáveis pelas diferenças negativas M_j;
p é o número de variáveis independentes presentes no modelo;

Cada termo do produto de R^P representa o complemento de 1 de cada diferença positiva ($M_j>0$) escalonada para um valor máximo (p-1)/p. Dessa maneira é obtida uma espécie de penalidade para as variáveis presentes no modelo. O valor é baixo se a variável apresenta uma alta correlação absoluta com a resposta, caindo a zero quando o valor da correlação absoluta (R_{jy}) entre a variável independente e a dependente se iguala ao valor do coeficiente de ajuste do modelo (R) e com um número de parâmetros maior que 1. A função R^P é o produto destas penalidades. Um baixo valor de R^P, é decorrente de uma variável do modelo apresentar um valor de correlação absoluta muito próximo do valor do coeficiente de ajuste do modelo, portanto as outras variáveis presentes no modelo não são significativas desde que não contribuem para o aumento da correlação múltipla linear. Neste caso o modelo é demasiado complexo em relação a sua qualidade. Ao contrário se cada variável independente explicar uma fração 1/p do total do coeficiente de ajuste do modelo, o valor de R^P é igual a 1(Todeschini *et al.* 2004).

A regra validação de modelos através da função R^P é definida pela equação I.9.3.4.

$$R^P < t^p \rightarrow \text{rejeite o modelo} \qquad \text{Equação I.9.3.4}$$

Onde: t^p é um limite pré-definido de 0.01 a 0.1 dependendo dos dados. Um valor sugerido para t^p é 0.05. (Todeschini *et al.* 2004)

Suponha que um modelo apresente um valor de coeficiente de ajuste de R = 0,9, e que haja três variáveis independentes deste modelo as quais apresentem coeficientes de correlação absoluta com a variável dependente respectivamente de R_{1y} = 0,9, R_{2y} = 0,1 R_{3y} = 0,1, valor da função R^P será zero. Portanto o modelo seria rejeitado.

A função R^N expressa na equação I.9.3.3 é a soma das diferenças negativas M_j, obtida através das variáveis independentes as quais o valor da razão entre o valor de coeficiente de correlação absoluta e o valor do coeficiente de ajuste do modelo é igual a 1/p. A função R^N considera que um valor baixo de coeficiente de correlação absoluta da variável independente com a variável dependente pode ser um indício de uma variável não significante. A função RN indica o excesso de variáveis não significantes, e pode ser considerado como uma medida de *"overfitting"* devido a presença de variáveis que agregam apenas ruídos ao modelo (Todeschini *et al.* 2004).

Assumindo que em um modelo todas as variáveis apresentem um baixo valor de correlação absoluta com a variável dependente de ε, então o valor Mj de cada uma destas variáveis é expresso na equação I.9.3.5.

$$\frac{\varepsilon}{R} - \frac{1}{p} = \frac{p\varepsilon - R}{pR} \qquad \text{Equação I.9.3.5}$$

Onde: $\varepsilon << R$
ε é o valor de correlação absoluta entre a variável independente e a variável dependente;
R é o valor do coeficiente de ajuste do modelo;

O valor de ε pode ser alterado pelo usuário dependendo do conhecimento do ruído contido na variável dependente. Além disso, presume-se que não é permitida mais de uma variável que agregue somente ruído no modelo. Portanto o limite t^n para a função R^N pode ser estimado pela equação I.9.3.6:

$$t^N(\varepsilon) = \frac{p\varepsilon - R}{pR} \qquad \text{Equação I.9.3.6}$$

Onde: p é o número de variáveis no modelo;
ε é o valor determinado pelo usuário;
R é o valor do coeficiente de ajuste do modelo;

A escolha de aceitar eventualmente uma variável com baixo valor de correlação absoluta com a variável independente se deve a impossibilidade de saber se uma variável é apenas "ruído" ou se "explica" os resíduos do modelo. Por fim, regra validação de modelos através da função R^N é definida pela equação I.9.3.7.

se $R^N < t^N(\varepsilon) \rightarrow$ rejeite o modelo Equação I.9.3.7

Onde: $t^N(\varepsilon)$ é um limite pré-definido;

Ao contrário de R^N que só pode ser negativo, o valor do limite t^N pode ser positivo. Neste caso qualquer valor diferente de zero no R^N será rejeitado pela regra. Neste caso, a regra independe das correlações entre as variáveis independentes e a variável dependente, e o modelo deve ser rejeitado devido ao baixo valor do coeficiente de ajuste R com relação ao nível de ruído ε escolhido. Isto indica que a correlação entre as variáveis independentes e a variável dependente ocorreu ao acaso (Todeschini *et al.* 2004).

Aumentar os valore de ε, aumenta-se os valores do limite para a função R^N. Para um modelo com um coeficiente de ajuste de 0,6, e os valores de ε iguais a 0,01, a 0,05 e a 0,1 resultam respectivamente em valores de limite para a função de R^N de –0,317, de –0,250 e de –0,167 respectivamente. Com o valor de ε igual a 0, o valor limite para a função RN fica limitado a 1/p. Por exemplo: para um modelo com três variáveis, o valor de limite para a função de R^N é de –0,333 (Todeschini *et al.* 2004).

Alguns exemplos do comportamento das funções R^P e R^N são mostrados na tabela I.9.3.1.

Tabela I.9.3.1. Valores das funções R^P e R^N para alguns modelos teóricos com três variáveis independentes.

ID[a]	R_{1Y}[b]	R_{2Y}[c]	R_{3Y}[d]	R[e]	R^P[f]	R^N[g]	Modelo Aceito
1	0,90	0,90	0,90	0,90	**0**[h]	0	**Não**
2	0,90	0,10	0,10	0,90	**0**[h]	**-0,444**[i]	**Não**
3	0,89	0,50	0,10	0,90	**0,011**[h]	-0,222	**Não**
4	0,80	0,80	0,10	0,90	**0,028**[h]	-0,222	**Não**
5	0,80	0,70	0,10	0,90	0,056	-0,222	Sim
6	0,80	0,20	0,10	0,90	0,167	**-0,333**[i]	**Não**
7	0,60	0,40	0,10	0,90	0,417	-0,222	Sim
8	0,60	0,40	0	0,90	0,417	**-0,333**[i]	**Não**
9	0,50	0,30	0,10	0,90	0,667	-0,222	Sim
10	0,40	0,40	0,10	0,90	0,694	-0,222	Sim
11	0,30	0,30	0,30	0,90	1	0	Sim

[a] Número de identificação do modelo;
[b] Valor da correlação absoluta entre a variável independente 1 e a variável dependente;
[c] Valor da correlação absoluta entre a variável independente 2 e a variável dependente;
[d] Valor da correlação absoluta entre a variável independente 3 e a variável dependente;
[e] Valor do coeficiente de ajuste do modelo;
[f] Valor da função RP obtido;
[g] Valor da função RN obtido;
[h] Valores em negrito por serem menores do valor do limite de 0,05 estabelecido para a função R^P;
[i] Valores em negrito serem menores do valor do limite de -0,261 (ε = 0,01) estabelecido para a função R^N;

I.9.4. Critério de Validação de Modelos Desenvolvido por Tropsha

A validação de modelos de QSAR (Lindgren *et al.* 1991) é realizada através da predição tanto interna (utilizando os compostos da série de treinamento) como também com a validação externa (utilizando compostos aos quais não participaram da série de treinamento).

Na literatura, Tropsha estudando os modelos de QSAR obtidos por Cramer para série de esteróides (Cramer *et al.* 1988) observou (Golbraikh & Tropsha 2002; Novellino *et al.* 1995) que altos valores do coeficiente de correlação de predição interna (Q^2_{cv})

independe do valor do coeficiente de ajuste de predição externa (R^2). Como pode ser observado na figura I.9.4.1 não se observa correlação entre os valores de Q^2_{cv} e de R^2.

Considerando-se que a um alto coeficiente de predição interno $Q^2_{cv} > 0,5$, não está, automaticamente garantido, um alto poder de predição externo, foi proposto e aplicado por grupo de pesquisadores (Golbraikh & Tropsha 2002; Golbraikh & Tropsha 2002; Golbraikh & Tropsha 2003; Golbraikh *et al.* 2003; Tropsha *et al.* 2003) uma regra baseada na utilização de série externa para a validação do modelo.

Esta regra assume que os valores do coeficiente de correlação (R) obtidos entre os valores preditos e os correspondentes observados, para a série de teste externa, devem ser maiores que 0,6. Adicionalmente, foi mostrado que a observação de um alto valor do coeficiente de regressão externo não é suficiente para que um modelo apresente uma predição acurada (Golbraikh & Tropsha 2002).

A partir dos valores observados *versus* os correspondentes preditos construíram-se gráficos apresentados nas figuras I.9.4.1b e I.9.4.1d. E, generalizando, estes podem ser expressos pela *e*quação de regressão linear I.9.4.1.

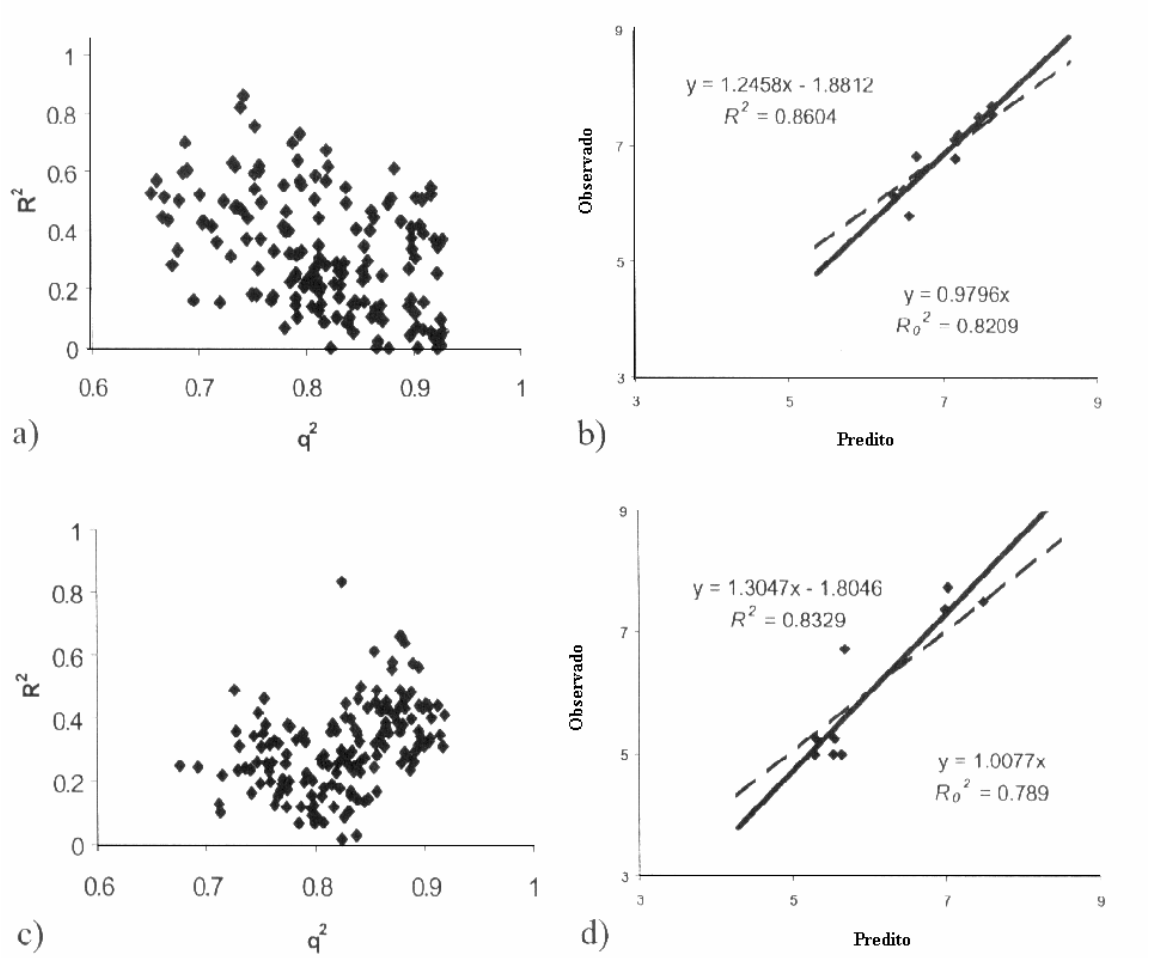


Figura I.9.4.1- Valores do coeficiente de correlação de ajuste dos valores preditos da série de teste (R^2) em função dos valores de coeficiente predição interna (q^2), apresentados nos gráficos **a** e **c; V**alores observados em função dos correspondentes preditos da série de teste gráficos **b** e **d**, considerados no trabalho original de validação de modelos, proposto por Tropsha (Golbraikh & Tropsha 2002).

Numa equação de regressão, para um modelo "ideal" observa-se um coeficiente angular (*a- e*quação I.9.4.2) com valor igual a 1 e um coeficiente linear (*b- e*quação I.9.4.3), com valor igual a 0.

Por outro lado, para que um modelo de QSAR estatisticamente significativo, seja considerado válido deve-se observar valores próximos de 1, para os valores tanto do coeficiente de ajuste externo bem como para os dos coeficientes angulares das equações de regressão das retas que passam pela origem.

$$y^r = a\tilde{y} + b \qquad \text{Equação I.9.4.1}$$

Onde: y^r é o valor da atividade biológica obtido pela equação de regressão;
$\tilde{y}$ é o valor de atividade biológica predito pelo modelo obtido através da série de treinamento;
a é o valor do coeficiente angular da reta;
b é o o valor do parâmetro constante da equação de regressão;

Os valores do coeficientes *a* e *b* são calculados através das equações I.9.4.2. e I.9.4.3.

$$a = \frac{\sum (y_i - \bar{y}) \times (\tilde{y}_i - \bar{\tilde{y}})}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \qquad \text{Equação I.9.4.2}$$

Onde: y_i é o valor de atividade biológica observado;
$\tilde{y}_i$ é o valor de atividade biológica predito pelo modelo obtido através da série de treinamento;
$\bar{y}$ é a média dos valores de atividade biológica observado;
$\bar{\tilde{y}}$ é a média dos valores de atividade biológica preditos pelo modelo obtido através da série de treinamento;

$$b = \bar{y} - a\bar{\tilde{y}} \qquad \text{Equação I.9.4.3}$$

Onde: $\bar{y}$ é a média dos valores de atividade biológica observado;
$\bar{\tilde{y}}$ é a média dos valores de atividade biológica preditos pelo modelo obtido através da série de treinamento;

Desta forma, nas equações de regressão, das retas que passam pela origem e, expressas respectivamente pelas equações I.9.4.4. e I.9.4.5. um dos dois termos k ou k' deve apresentar valores próximos de 1,

$$y^{r0} = k\tilde{y} \qquad \text{Equação I.9.4.4}$$

Onde: $\tilde{y}$ é o valor de atividade biológica predito pelo modelo obtido através da série de treinamento;
k é o coeficiente angular da equação de regressão;

$$\tilde{y}^{ro} = k' y \qquad \text{Equação I.9.4.5}$$

Onde: y é o valor de atividade biológica observado;

k' é o coeficiente angular da euqação de regressão;

Adicionalmente, invertendo-se os eixos correspondentes aos valores observados e preditos (Figura I.9.4.2) pode-se observar que tanto os valores dos coeficientes estatísticos bem como os das equações de regressão geradas são diferentes entre si. (Golbraikh & Tropsha 2002).

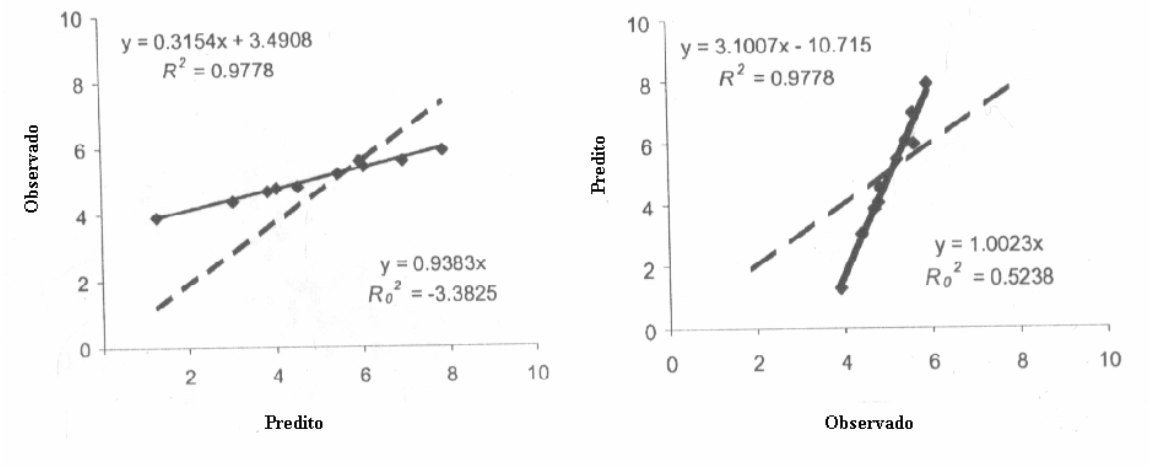


Figura I.9.4.2. Gráfico entre os valores observados e preditos e o correspondente construído invertendo-se os eixos, para uma série de teste.

Verifica-se na figura I.9.4.2. uma grande diferença entre os valores do coeficiente de regressão R_0^2 e $R_0^{2'}$ respectivamente das retas $y^{r0} = k\tilde{y}$ (gráfico dos valores observados em função dos valores preditos do modelo obtido através da série de treinamento) e $\tilde{y}^{ro} = k'y$ (gráfico dos valores observados em função dos valores preditos do modelo obtido através da série de treinamento) que passam pela origem. Deste modo, a regra proposta por Tropsha impõe que além de apresentarem valores do coeficiente de correlação de predição interna, maiores que 0,5 (Q^2_{cv}>0,5) e do coeficiente de ajuste de predição externo maiores que 0,6 (R^2> 0,6), os valores dos coeficientes de ajuste de predição externo das retas que passam pela origem dos eixos de coordenadas (R_0^2 ou $R'_0{}^2$) devem ser próximos ao valor do melhor coeficiente de ajuste de regressão (R^2). Este critério pode ser aplicado, considerando-se valores do limite máximo atribuídos aos desvios de R_0^2 e de $R'_0{}^2$, calculados em relação a R^2 e, expressos pelas equações I.9.4.6 e I.9.4.7. Adicionalmente,

através das equações I.9.4.8. e I.9.4.9. atribuí-se valor limite para os coeficientes angulares k ou k', ou seja, assuma valor próximo de 1.

$$\left[\frac{\left(R^2 - {R_0}^2\right)}{R^2}\right] < 0{,}1 \qquad \text{Equação I.9.4.6}$$

ou

$$\left[\frac{\left(R^2 - {R'_0}^2\right)}{R^2}\right] < 0{,}1 \qquad \text{Equação I.9.4.7}$$

$$0{,}85 \leq k \leq 1{,}15 \qquad \text{Equação I.9.4.8}$$

ou

$$0{,}85 \leq k' \leq 1{,}15 \qquad \text{Equação I.9.4.9}$$

Apesar de todos os índices de validação citados anteriormente, os gráficos dos valores dos resíduos em função dos valores preditos continuam sendo de extrema importância para a análise dos modelos gerados, verificando o ajuste linear do modelo e/ou a presença de *outliers*.