II. Metodologia

II.1. Séries de Compostos e Parâmetro Biológico Estudados

Neste trabalho foram selecionados da literatura original (French *et al.* 1974) (Du *et al.* 2002) (Chiyanzu *et al.* 2003) (Greenbaum *et al.* 2004) 90 compostos, incluindo-se 29 α-(N)-heterocíclica carboxaldeído tiossemicarbazona substituídas com atividade inibitória frente a ribonucleotídeo redutase (IRNR) de células H.Ep.-2 (French *et al.* 1974) (série I; compostos I.1. a I.29); 37 tiossemicarbazonas substituídas na cadeia lateral e no anel aromático (Du *et al.* 2002) (série II; compostos II.1 a II.37) e, 61 compostos estruturalmente diferentes (série III; II.1 a III.61) a saber, 37 tiossemicarbazonas selecionadas da série II; 16 derivados de isatinas (Chiyanzu *et al.* 2003); 8 tiossemicarbazonas (Greenbaum *et al.* 2004). As séries II e III apresentam atividade inibitória frente à cruzaína, uma cisteína protease do *T.cruzi*.

II.1.1. Critérios de Seleção

Todas as séries estudadas foram selecionadas da literatura, utilizando os seguintes critérios:

i) Os experimentos foram realizados pelo mesmo grupo de pesquisa;

ii) Seguiram-se os mesmos protocolos experimentais para cada uma das medidas de atividade biológica e,

iii) Para a série I selecionou-se apenas os compostos derivados da tiossemicarbazona substituídos no anel piridínico, eliminando-se desta forma os compostos substituídos no fragmento da tiossemicarbazona;

iv) Para as séries II e III, os compostos que apresentaram valores de IC_{50} maior ou igual a 10 μM. foram considerados inativos (DU *et al*., 2002)

v) Os intervalos de variação nos valores de pIC_{50} foi de no mínimo 2 unidades.

II.1.2. Série I

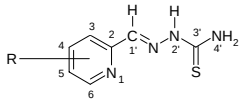
Considerando-se a similaridade estrutural com a série II objeto principal desta dissertação e detalhada no item II.1.3, a metodologia proposta neste trabalho foi aplicada para a série de 29 derivados da 2-formilpiridina tiosemicarbazonas anteriormente já estudada no grupo. Estes foram selecionados da literatura (French *et al.* 1974) e minimizados por Hamilton Ishiki, em sua tese de doutorado em fase final de redação. Estes compostos apresentam atividade inibitória frente a ribonucleotídeo redutase (IRNR) de células H.Ep.-2, determinadas em termos de da concentração molar necessária para produzir 50% da inibição máxima (IC_{50}), com erro dentro dos limites de ± 10% e, considerados como parâmetro biológico. Os valores de pIC_{50} variaram de 6,66 (composto 4) a 4,62 (composto 14), ou seja, ΔpIC_{50}=2,04. As estruturas e os valores correspondentes de pIC_{50} estão apresentados na tabela II.1.2.1.

Tabela II.1.2.1. Valores da atividade inibitória da ribonucleotídeo redutase (pIC_{50})[a] da série de tiossemicarbazonas substituídas[b], Série I, selecionadas da literatura[c].

Formatado

Formatado

Formatado



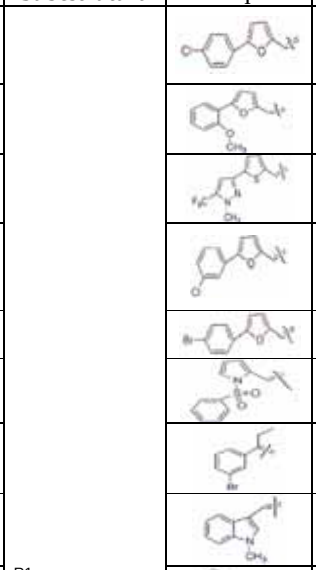
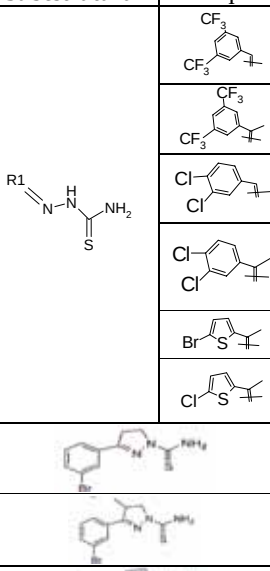
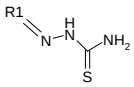
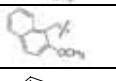
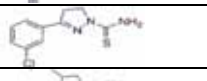
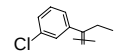
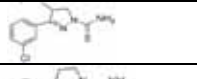
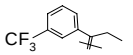
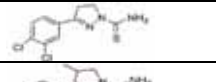
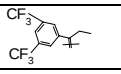
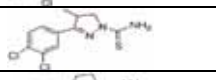
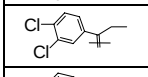
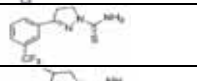
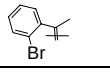
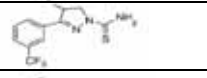
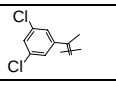
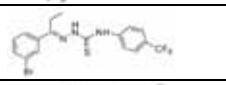
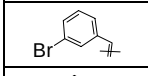
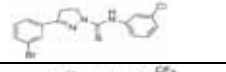
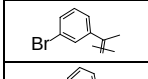
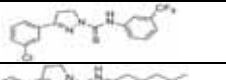
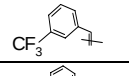
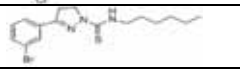
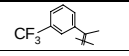
Nº do composto		R	pIC_{50}
1	*11*[b]	3-CH_3	6,59
2	*12*	4-CH_3	6,57
3	*13*	5-CH_3	6,51
4	*14*	5-C_2H_5	6,66
5	*15*	6-CH_3	5,11
6	*17*	3-OCH_3	5,89
7	*18*	3-OC_2H_5	6,04
8	*19*	3-$OOCCH_3$	5,44
9	*20*	3-F	5,42
10	*22*	5-F	5,92
11	*23*	5-Cl	6,25
12	*24*	5-Br	6,30
13	*25*	5-I	6,39
14	*26*	5-CF_3	5,62
15	*28*	5-OCH_3	5,92
16	*29*	5-OCF_3	5,60
17	*30*	5-OC_2H_5	6,07
18	*32*	5-$OC_2H_4N(CH_3)_2$	4,62
19	*33*	5-$O(C_2H_4O)_2C_2H_5$	5,69
20	*34*	5-$OOCCH_3$	5,44
21	*35*	5-$OOCC_2H_5$	5,28
22	*36*	5-n-$OOCC_3H_7$	5,17
23	*39*	5-$OOCCH_2OCH_3$	5,30
24	*40*	5-$OOCCH_2OC_2H_5$	5,25
25	*41*	5-$OOCCH_2N(CH_3)_2$	5,24
26	*42*	5-$OOCCH_2OC_6H_5$	4,89
27	*49*	5-$NHCOCH_3$	5,92
28	*50*	5-$N(CH_3)_2$	6,40
29	*61*	3,5-$(OC_2H_5)_2$	5,68

[a] logaritmo negativo da concentração (em M) para produzir 50% da inibição máxima da RNR;
[b] os números sublinhados em itálicos correspondem aos apresentados na literatura original;
[c] (French *et al.* 1974);

II.1.3. Série II

Numa segunda etapa, obedecendo-se aos critérios de seleção de séries, citados na seção II.1.2., foram utilizadas 37 tiosemicarbazonas substituídas na cadeia lateral e no anel aromático com atividade frente a cruzaína do *T. cruzi*, expressos por pIC_{50} (concentração em M). Estas foram selecionadas da literatura (Du *et al.* 2002) de uma série original de 58 compostos (tabelas II.1.3.1) sendo excluídos 21, estes foram considerados inativos (DU *et al*., 2002), pois apresentam valores de IC_{50} maior ou igual a 10 μM. Os correspondentes erros associados não foram citados. Os valores de pIC_{50} variaram de 7,70 (compostos 12 e 13) a 5,15 (composto 22), ou seja, ΔpIC_{50} =2,55. As estruturas e os valores correspondentes de pIC_{50} estão apresentados na tabela II.1.3.1.

Tabela II.1.3.1. Valores da atividade inibitória frente à cruzaína (pIC_{50})[a] da série de tiossemicarbazonas substituídas, série II, selecionadas da literatura(DU *et al*., 2002).

Nº	*ID*[b]	Composto		pIC_{50}	Nº	*ID*[c]	Composto		pIC_{50}
		Subestrutura	R_1				Subestrutura	R_1	
1	a1b			6,52	20	a3e			6,49
2	a1c			6,55	21	a3f			7,22
3	a1d			5,15	22	a3g			6,34
4	a1e			6,85	23	a3h			7,30
5	a1f			6,55	24	a3i			5,42
6	a1g			5,55	25	a3j			5,72
7	a1i			7,00	26	a4a			6,70
8	a1m			5,55	27	a4b			7,10
9	a1n			6,25	28	a4c			6,57
10	a2a			6,66	29	a4d			6,64
11	a2b			7,30	30	a4e			7,22
12	a2h			7,70	31	a4f			7,15
13	a2i			7,70	32	a4g			7,10
14	a2l			5,41	33	a4h			7,40
15	a2o			6,25	34	a5c			5,80
16	a3a			5,89	35	a5d			6,00
17	a3b			6,70	36	a5e			5,77
18	a3c			6,32	37	a5f			5,85
19	a3d			6,77					

[a] logaritmo negativo daconcentração (em M) para produzir 50% da inibição máxima frente à cruzaína;

[c] identificação do composto é o mesma utilizada pela literatura original, sendo que a primeira letra (*a*{DU *et al*., 2002}) foi incluída para para indicar o artigo da literatura a que se refere;

Formatado

II.1.4. Série III

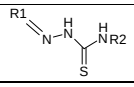
De modo análogo, obedeceu-se aos critérios de seleção de séries citados anteriormente. A série III (n=61) é constituída por: compostos II.1 a II.37 (n=37) da série II; compostos III.38 a III.53 (n=16) selecionados de uma série de 34 compostos, da literatura original (Chiyanzu *et al.* 2003). Adicionalmente, incluiu-se os compostos III.54 a III.61 (n=8) selecionados de uma série de 19 compostos, da literatura original (Greenbaum *et al.* 2004). Os compostos com valores $IC_{50} \geq 10$ μM foram considerados inativos (Chiyanzu *et al.* 2003) (Greenbaum *et al.* 2004) e excluídos da análise.

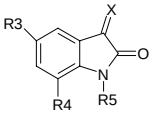
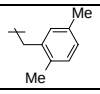
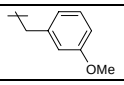
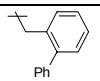
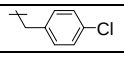
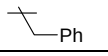
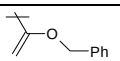
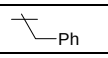
Excluído: s IIIx e IIIy

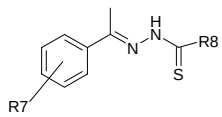
Excluído: IIIz

Os compostos da série III possuem os valores de pIC_{50} frente à cruzaína do *T. cruzi*. Os correspondentes erros associados não foram citados. Os valores de pIC_{50} variaram de 7,70 (compostos 12 e 13) a 4,05 (composto 51), ou seja, ΔpIC_{50} =2,55. As estruturas e os valores correspondentes de pIC_{50} estão apresentados na tabela II.1.4.1

Tabela II.1.4.1. Valores da atividade inibitória frente à cruzaína (pIC50)[a] da série de tiossemicarbazonas substituídas selecionadas da literatura (Chiyanzu et al., 2003, Greenbaum et al., 2004).

Nº	Composto	$_pIC_{50}$
	Subestrutura	
1 a 37		**Estruturas, *ID* e Valores Idênticos a Série II**

Nº	*ID*[b]	Composto					pIC_{50}
		Subestrutura	**R_3**	**R_4**	**R_5**	**X**	
38	*b2a*		H	H	H	N-NHC(S)NH_2	5,10
39	*b2b*		Me	H			4,46
40	*b2c*		F	H			4,52
41	*b2d*		Cl	H			4,68
42	*b2e*		Br	H			4,70
43	*b2f*		I	H			5,05
44	*b2g*		NO2	H			4,52
45	*b2h*		Me	Me			4,80
46	*b2i*		Cl	Me			4,98
47	b3e		Me	H		O	4,10
48	*b3g*		Me				5,70
49	*b3h*		Me				5,55
50	*b3i*		Cl				5,22
51	*b3k*		Cl				4,05
52	*b3l*		Cl				4,10
53	*b3m*		I				4,05

Nº	*ID*[b]	Composto			pIC_{50}
		Subestrutura	**R_7**	**R_8**	
54	*c1b*		3'-C_6H_5	NH_2	5,10
55	*c1c*		4'- C_6H_5	NH_2	5,30
56	*c2d*		3'- OC_6H_5	NH_2	5,00
57	*c2e*		4'-OC_6H_5	NH_2	5,00
58	*c3i*		3'-CH(Me)NNHC(S)NH_2	NH_2	5,30
59	*c4c*		3'-Br		5,30
60	*c4d*		3'-Br	NEt_2	5,30
61	*c4e*		3'-Br		5,30

[a] logaritmo negativo da concentração (em M) para produzir 50% da inibição máxima frente à cruzaína;

[b] identificação do composto é o mesma utilizada pela literatura original, exceto as primeiras letras, respectivamente: a (DU *et al*., 2002), *b* (Chiyanzu et al., 2003)e *c* (Greenbaum et al., 2004) que foram utilizadas para especificar o artigo que foi retirado.

Excluído: subsérie IIIx,

Formatado

II.2. Programas e Recursos Computacionais Utilizados

As estruturas das moléculas foram escritas e minimizadas utilizando os seguintes programas computacionais:

- SYBYL 6.9, Tripos Inc (Sybyl program 2000) utilizado para escrever as estruturas das moléculas e,
- CORINA 3.0 (Schonberger *et al.* 2000) utilizado para gerar as estruturas 3D (em três dimensões) das moléculas (Sadowski & Gasteiger 1993);

Os descritores moleculares foram calculados utilizando o programa DRAGON 3.0 web (Todeschini, *et al.* 2003).

As análises *PLS* foram feitas utilizando os seguintes programas computacionais:

- ANALYZE, (Embrechts, M. J. 2001) para os dados da Série I. A visualização gráfica dos resultados das análises PLS foi realizada através de interface gráfica obtida por modificações feitas no programa pelo grupo utilizando linguagem *Visual Basic* (VB).
- SYBYL, versão 6.9, Tripos Inc, (Sybyl program 2000) para os dados das séries II e III. Os valores dos coeficientes estatísticos (r^2, Q^2) das análises PLS e a seleção automática dos descritores mais significativos, ou seja, maiores módulos dos valores dos coeficientes dos descritores nos modelos de *PLS*, foram obtidos por modificações introduzidas no programa pelo grupo utilizando linguagem de programação SPL.

Para a obtenção das equações de *QSAR* clássico, modelos lineares, foi utilizado o programa BILIN (KUBINYI 1995).

Todos os cálculos foram feitos utilizando-se **u**m PC com processador Intel Pentium II de 350 MHz, e memória RAM de 64 MB exceto para aqueles utilizando o programa SYBYL. Estes últimos foram feitos em uma estação gráfica Silicon Graphics Power

Challenge 10000, com 4 processadores de 194 MHz MIPS R10000, e memória RAM de 512 Mb.

II.3. Obtenção da Geometria das Moléculas

II.3.1. Série I

As geometrias dos compostos da série I já haviam sido previamente minimizadas por Hamilton Ishiki em sua tese de doutorado, através de análise conformacional utilizando a rotina "*systematic search*" implementada no programa Sybyl v. 6.9 (Sybyl Tripos Associates Inc). Em seguida, as geometrias de menor energia, obtidas através da busca sistemática, foram otimizadas com o emprego do método hamiltoniano AM1 (Dewar *et al.* 1985, 1990) utilizando o programa semiempírico MOPAC 6.0 (Stewart 1990), implementado no programa Sybyl. O critério da norma do gradiente foi reduzido para 0.0 com o emprego da palavra chave GNORM=0.0 e, a otimização do critério de convergência e os cálculos auto consistentes foram incrementados em 100 vezes com a utilização da palavra chave PRECISE. A palavra chave MMOK foi utilizada para permitir as correções de mecânica molecular das ligações peptídicas dos compostos. Os compostos foram completamente otimizados sem restrições de geometria com o emprego do algoritmo de Baker (Baker 1986). Neste estudo foram calculadas os valores das cargas Gasteiger-Marsili para os compostos (Gasteiger & Marsili 1981).

II.3.2. Séries II e III

As moléculas das séries II e III foram desenhadas empregando o programa Sybyl v. 6.9 (Sybyl program 2000). A seguir, utilizou-se o programa CORINA 3.0 (Sadowski & Gasteiger 1993, Schonberger *et al.* 2000) para gerar as coordenadas em 3 dimensões de cada uma das moléculas e, utilizadas diretamente como arquivo de entrada para a obtenção dos descritores pelo programa DRAGON. Neste estudo foram calculadas os valores das cargas Gasteiger-Marsili para os compostos (Gasteiger & Marsili 1981).

II.4. Obtenção dos Parâmetros Estruturais

Para a obtenção de todos os parâmetros estruturais para as três séries, foi utilizado o programa DRAGON. Os arquivos de entrada para as moléculas de cada série foram selecionados na opção *"Calculate Descriptors"*. Na opção *"Descriptor Selection"* selecionou-se todos os descritores possíveis. A obtenção dos parâmetros foi rápida, ou seja, em torno de 4 minutos para 61 moléculas, utilizando-se um PC com processador Pentium II (350 MHz) com 64 Mb de memória RAM com sistema operacional Windows 98.

Após a obtenção dos descritores DRAGON, um arquivo contendo todos os descritores para cada série foi salvo no formato de arquivo texto. A seguir, foram adicionados os valores correspondentes de pIC_{50} na última coluna do arquivo.

II.5. Critérios Utilizados para a Seleção de Variáveis

II.5.1. Séries de Treinamento e de Teste

Para a seleção de variáveis, obtidos os valores das variáveis independentes (descritores) através do programa DRAGON, cada Série (I, II e III) foi subdividida em duas subséries, denominadas de série de treinamento e de teste. Este processo de divisão de cada Série (I, II, III) foi repetido três vezes. Assim para cada Série (I, II e III) foram geradas três séries de treinamento com as suas respectivas séries de teste. O processo de divisão das Séries (I, II, e III) está detalhado a seguir.

II.5.1.1. Série I

As três séries de treinamento, denominadas Séries IA IB e IC, e suas respectivas séries de teste mostradas na tabela II.5.1.1.1, foram divididas de forma que todas as seis séries (de teste e de treinamento) formadas abrangessem toda a faixa de valores de pIC_{50}. As séries de teste, no entanto, não mantiveram obrigatoriamente todos os compostos diferentes entre elas. O número de compostos de cada série de teste representa cerca de 20% do número de compostos, da correspondente série de treinamento**.** As distribuições

dos valores de atividade das séries de treinamento e de teste são mostradas nas figuras II.5.1.1.1, II.5.1.1.2 e II.5.1.1.3.

Tabela II.5.1.1.1. Números dos compostos das séries de treinamento e de teste (IA, IB e IC respectivamente) constituídas a partir dos 29 IRNR[a] selecionados da literatura[b] para análise *PLS*.

Séries I A		**Séries I B**		**Séries I C**	
Treinamento[c]	**Teste[c]**	**Treinamento[c]**	**Teste[c]**	**Treinamento[c]**	**Teste[c]**
2; 3; 4; 5; 7; 8;	1;	1; 2; 4; 5; 6; 8;	3;	1; 2; 3; 4; 5; 6;	7;
9; 10; 11; 13;	6;	10; 11; 12; 13;	7;	8; 9; 10; 13;	11;
14; 15; 16; 18;	12;	14; 15; 16; 17;	9;	14; 15; 16; 18;	12;
19; 21; 23; 24;	17;	18; 19; 20; 22;	21;	19; 20; 21; 22;	17;
25; 26; 27; 28;	20;	23; 25; 27; 28;	24;	23; 24; 26; 28;	25;
29;	22;	29;	26;	29;	27;

[a] inibidores da ribonucleotídeo redutase;
[b] (French *et al.* 1974);
[c] os números dos compostos correspondem aos apresentados na tabela II.1.2.1;

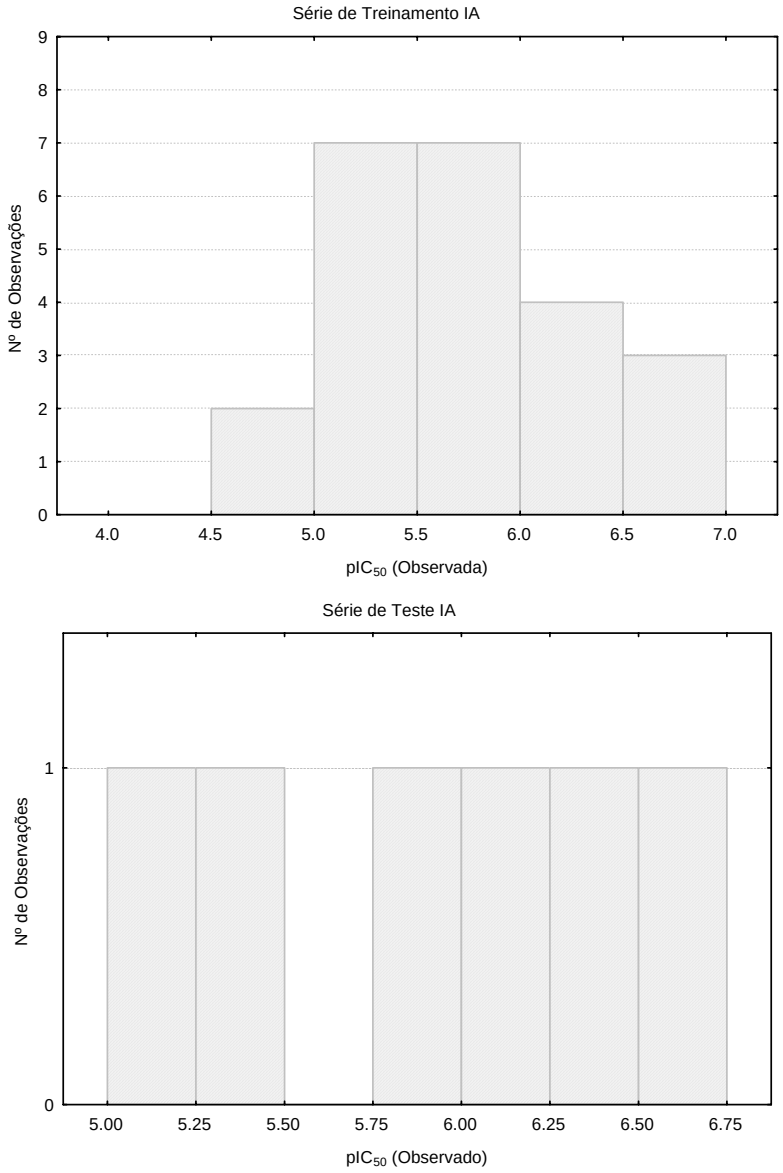


Figura II.5.1.1.1. Histogramas da distribuição dos valores de atividade pIC_{50} respectivamente das séries de treinamento e de teste IA.

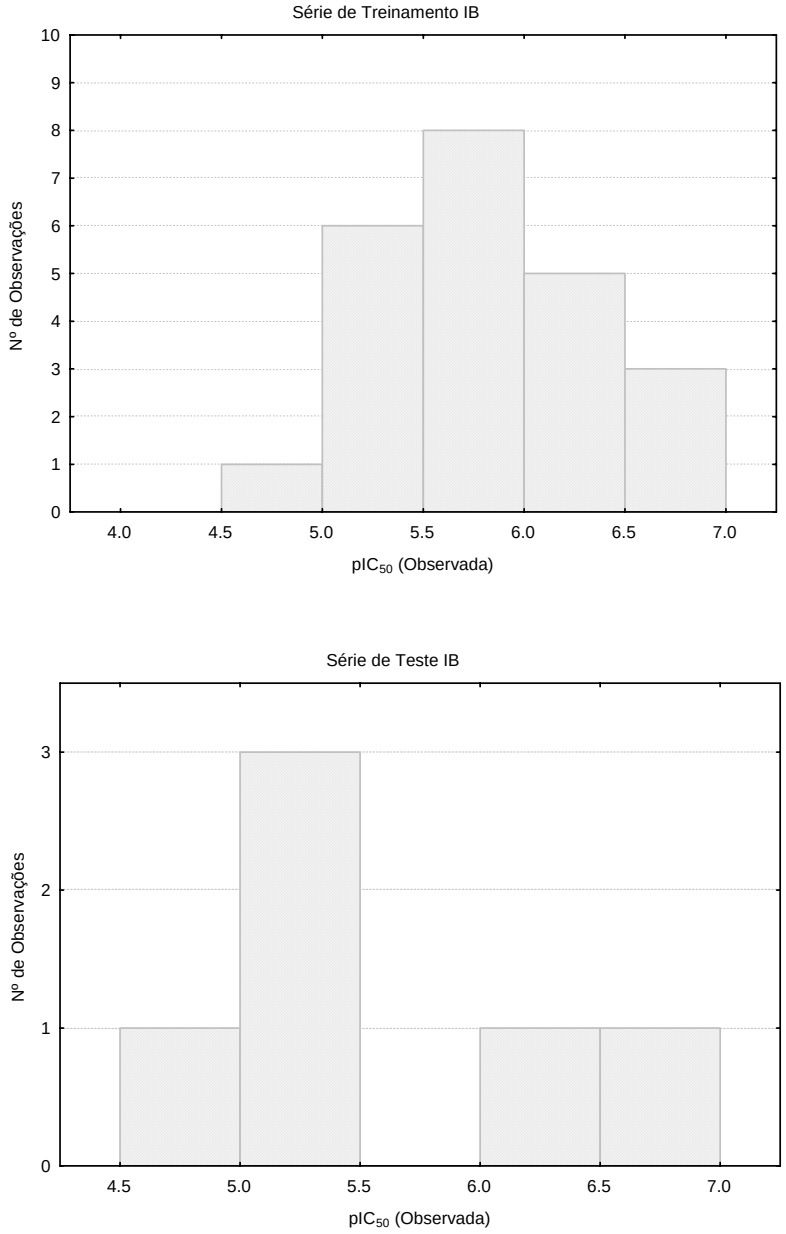


Figura II.5.1.1.2. Histogramas da distribuição dos valores de atividade pIC_{50} respectivamente das séries de treinamento e de teste IB.

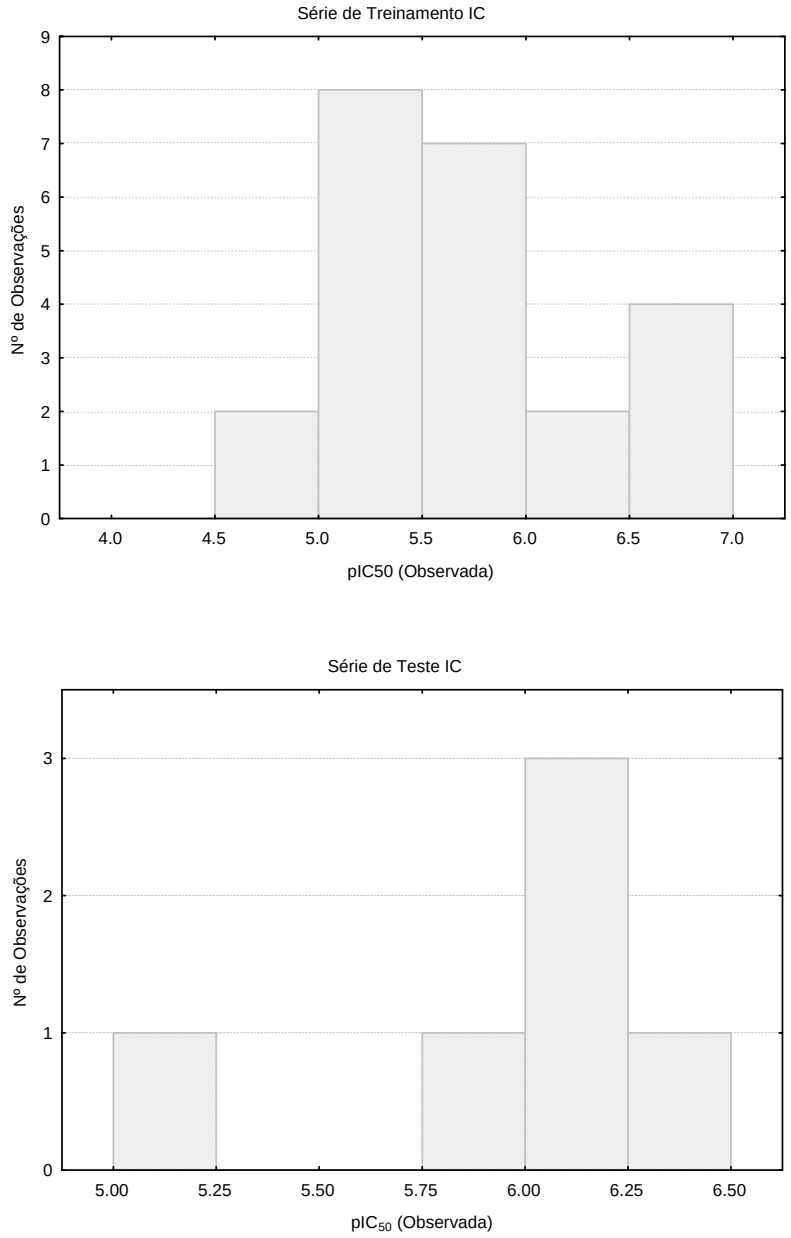


Figura II.5.1.1.3. Histogramas da distribuição dos valores de atividade pIC_{50} respectivamente das séries de treinamento e de teste IC.

II.5.1.2. Série II

Para a Série II, analogamente à Série I, foram constituídas três diferentes séries de treinamento denominadas respectivamente, IIA, IIB e IIC e suas correspondentes séries de teste (tabela II.5.1.2.1), fazendo-se com que ambas abrangessem toda a faixa de valores de pIC_{50}. Neste caso, todas as séries de teste não possuíam nenhum composto idêntico, ou seja , as séries de teste (representando em torno de 20% da série de treinamento) são totalmente diferentes, quanto a natureza dos compostos. Adotamos este critério para a Série II pela presença de um número maior de compostos nesta, em relação a Série I. As distribuições dos valores de atividade das séries de treinamento e de teste são mostradas nas figuras II.5.1.2.1, II.5.1.2.2 e II.5.1.2.3.

Tabela II.5.1.2.1. Números dos compostos das séries de treinamento e de teste (IIA, IIB e IIC respectivamente) constituídas a partir das 37 tiossemicarbazonas substituídas (Série II) com atividade inibitória frente a cruzaína selecionadas da literatura[a] para análise *PLS*.

Série II A		**Série II B**		**Série II C**	
Treinamento[b]	**Teste**[b]	**Treinamento**[b]	**Teste**[b]	**Treinamento**[b]	**Teste**[b]
1; 3; 4; 5; 7; 8;	2;	1; 2; 3; 5; 6; 7;	4;	1; 2; 4; 5; 6;	3;
9; 10; 11; 12;	6;	8; 9; 11; 12;	10;	7; 8; 10; 11; 13;	9;
14; 15; 16; 19;	13;	13; 14; 15; 17;	16;	14; 15; 16; 17;	12;
20; 21; 22; 23;	17;	18; 19; 20; 21;	23;	18; 19; 21; 22;	20;
24; 25; 26; 27;	18;	22; 24; 25; 26;	27;	23; 24; 26; 27;	25;
28; 29; 30; 31;	32;	28; 29; 30; 32;	31;	28; 29; 31; 32;	30;
33; 35; 36; 37;	34;	33; 34; 35; 37;	36;	33; 34; 36; 37;	35;

[a] (Du *et al*. 2002);
[b] os números dos compostos correspondem aos apresentados na tabela II.1.3.1;

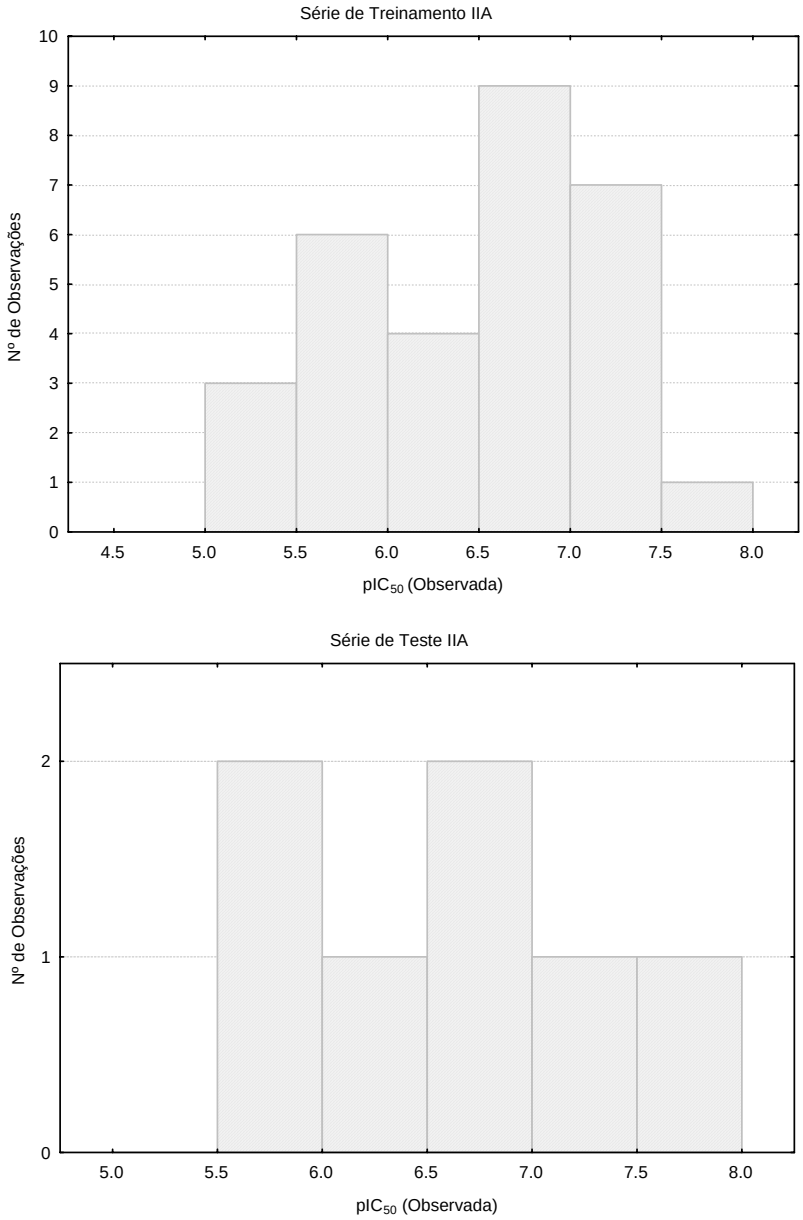


Figura II.5.1.2.1. Histogramas da distribuição dos valores de atividade pIC_{50} respectivamente das séries de treinamento e de teste IIA.

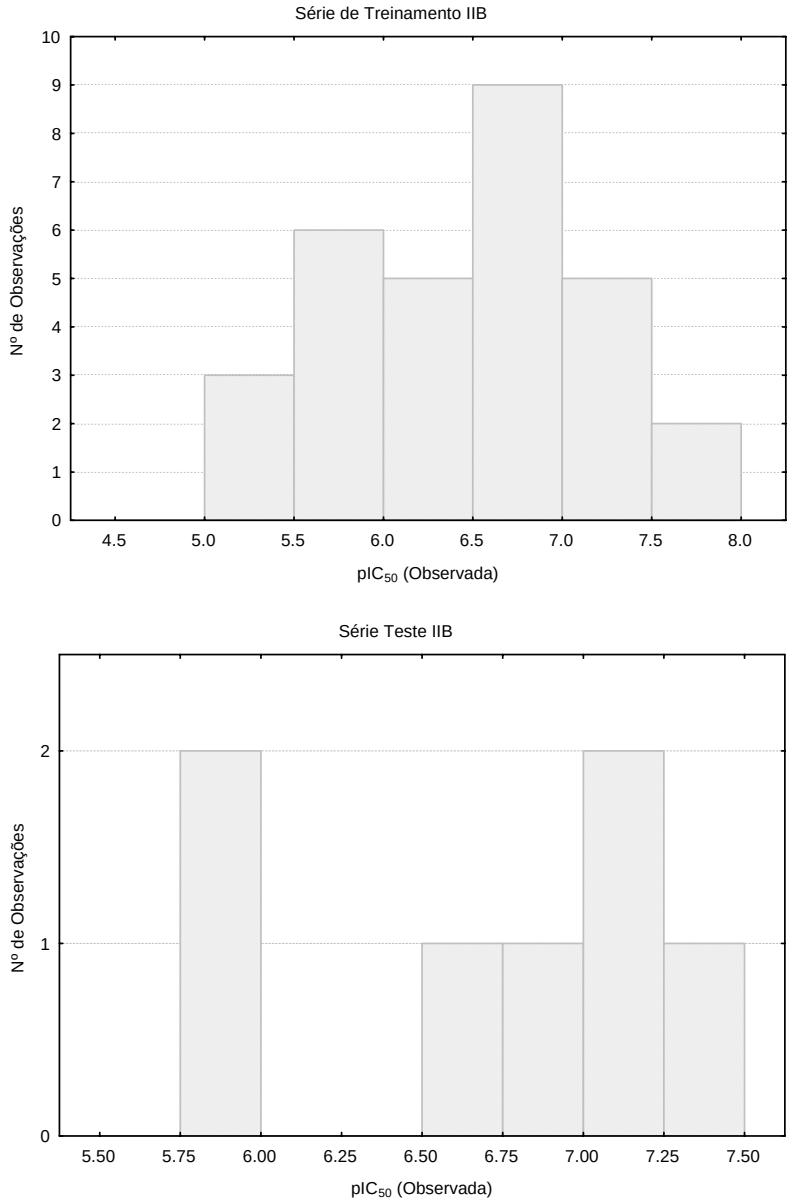


Figura II.5.1.2.2. Histogramas da distribuição dos valores de atividade pIC_{50} respectivamente das séries de treinamento e de teste IIB.

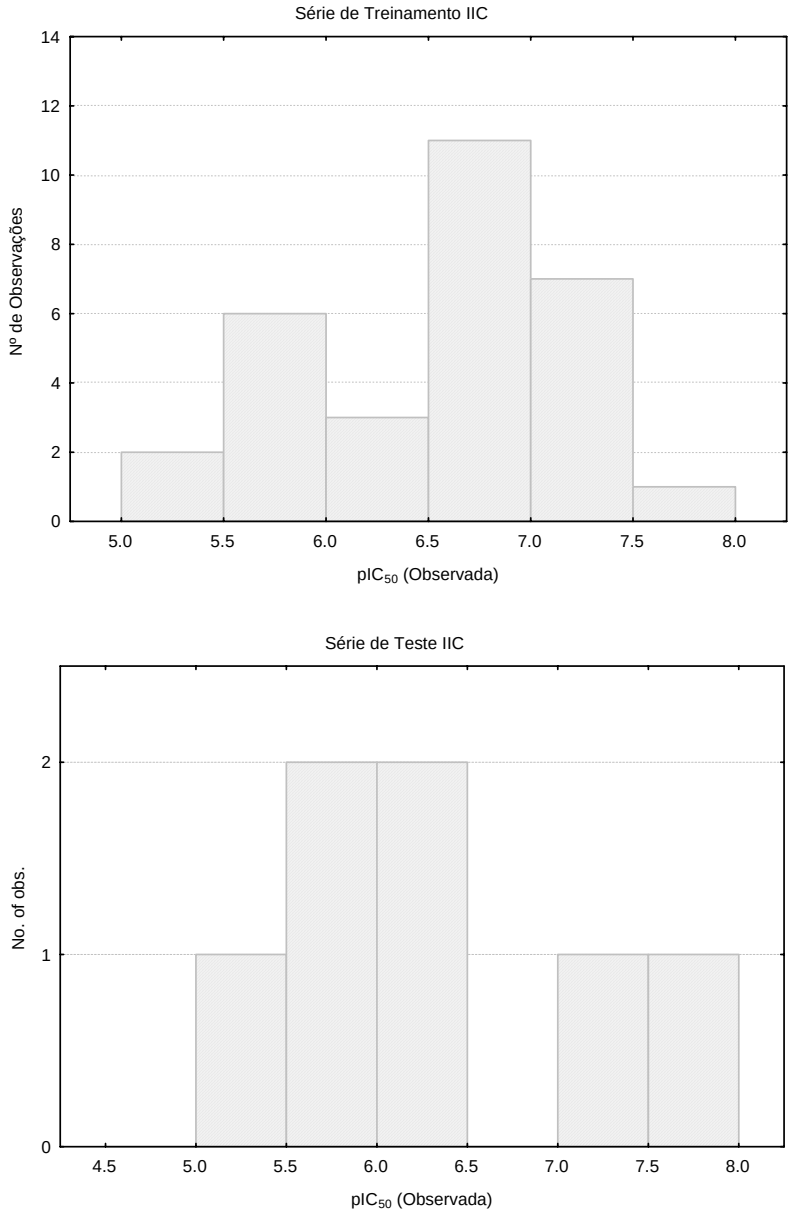


Figura II.5.1.2.3. Histogramas da distribuição dos valores de atividade pIC_{50} respectivamente das séries de treinamento e de teste IIC.

II.5.1.3. Série III

Para a série III, analogamente às séries I e II, foram constituídas três diferentes séries de treinamento, denominadas IIIA, IIIB e IIIC e suas respectivas séries de teste mostradas na tabela II.5.1.3.1, de forma que, ambas (teste e treinamento) abrangessem toda a faixa de valores de pIC_{50}. Para a série III, todas as séries de teste (representando em torno de 20% da série de treinamento) não possuíam nenhum composto idêntico, ou seja , são séries totalmente diferentes. As distribuições dos valores de atividade das séries de treinamento e as correspondentes de teste são mostradas nas figuras II.5.1.3.1, II.5.1.3.2 e II.5.1.3.3.

Tabela II.5.1.3.1.. Números dos compostos das séries de treinamento e de teste (IIA, IIB e IIC respectivamente) constituídas a partir das 37 tiossemicarbazonas substituídas (Série II) com atividade inibitória frente a cruzaína selecionadas da literatura[a] para análise *PLS*.

Série III A		**Série III B**		**Série III C**	
Treinamento[b]	**Teste**[b]	**Treinamento**[b]	**Teste**[b]	**Treinamento**[b]	**Teste**[b]
1; 3; 4; 5; 7; 8; 9;10;	2;	1; 2; 3; 5; 6; 7; 8; 9;	4;	1; 2; 4; 5; 6; 7; 8; 10;	3;
11; 12; 14; 15; 16;	6;	11; 12; 13; 14; 15;	10;	11; 13; 14; 15; 16;	9;
19; 20; 21; 22; 23;	13;	17; 18; 19; 20; 21;	16;	17; 18; 19; 21; 22;	12;
24; 25; 26; 27; 28;	17;	22; 24; 25; 26; 28;	23;	23; 24; 26; 27; 28;	20;
29; 30; 31; 33; 35;	18;	29; 30; 32; 33; 34;	27;	29; 31; 32; 33; 34;	25;
36; 37; 38; 40; 41;	32;	35; 37; 38; 39; 40;	31;	36; 37; 38; 39; 40;	30;
42; 43; 44; 45; 46;	34;	42; 43; 44; 45; 46;	36;	41; 42; 43; 44; 45;	35;
47; 48; 50; 51; 52;	39;	48; 49; 50; 51; 52;	41;	47; 49; 50; 51; 52;	46;
53; 54; 55; 57; 58;	49;	53; 54; 55; 56; 57;	47;	53; 55; 56; 57; 58;	48;
59; 60; 61;	56;	58; 60; 61;	59;	59; 60; 61;	54;

[a] (Du *et al*. 2002), (Chiyanzu et al., 2003), (Greenbaum et al., 2004);
[b] os números dos compostos correspondem aos apresentados na tabela II.1.3.1 e II.1.4.1;

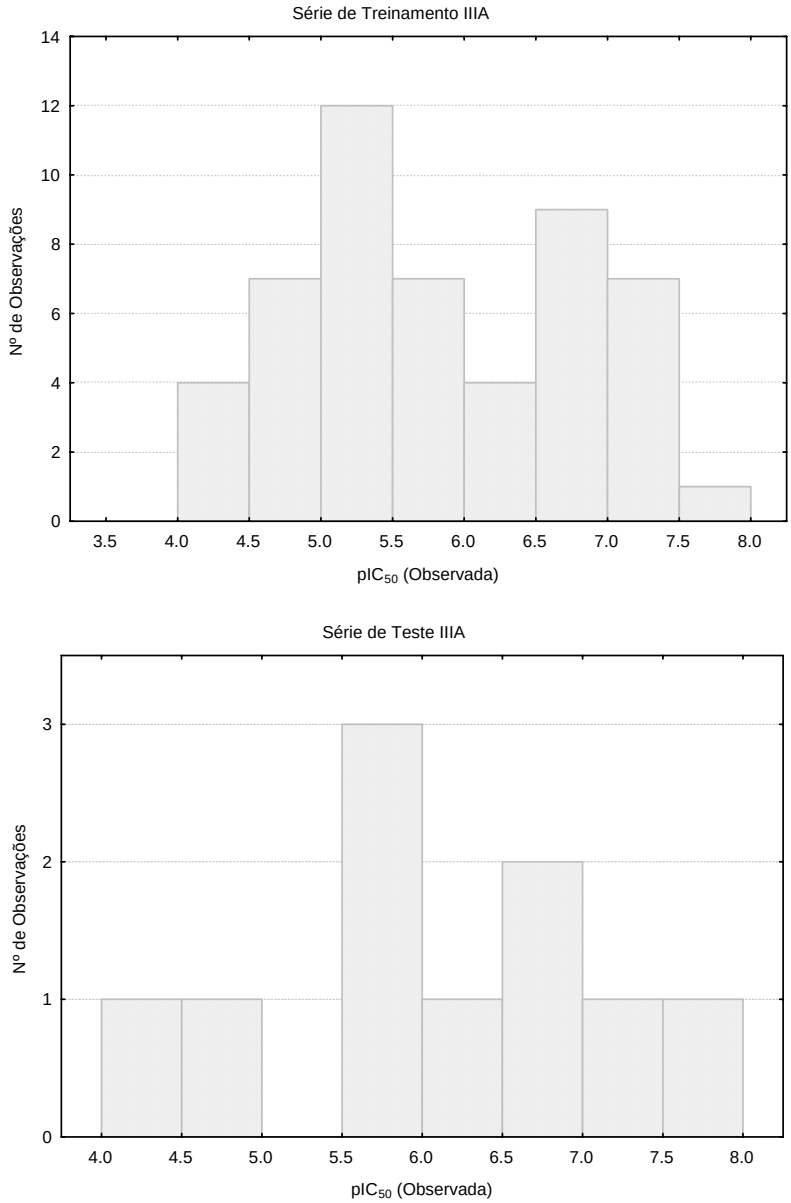


Figura II.5.1.3.1. Histogramas da distribuição dos valores de atividade pIC_{50} respectivamente das séries de treinamento e de teste IIIA.

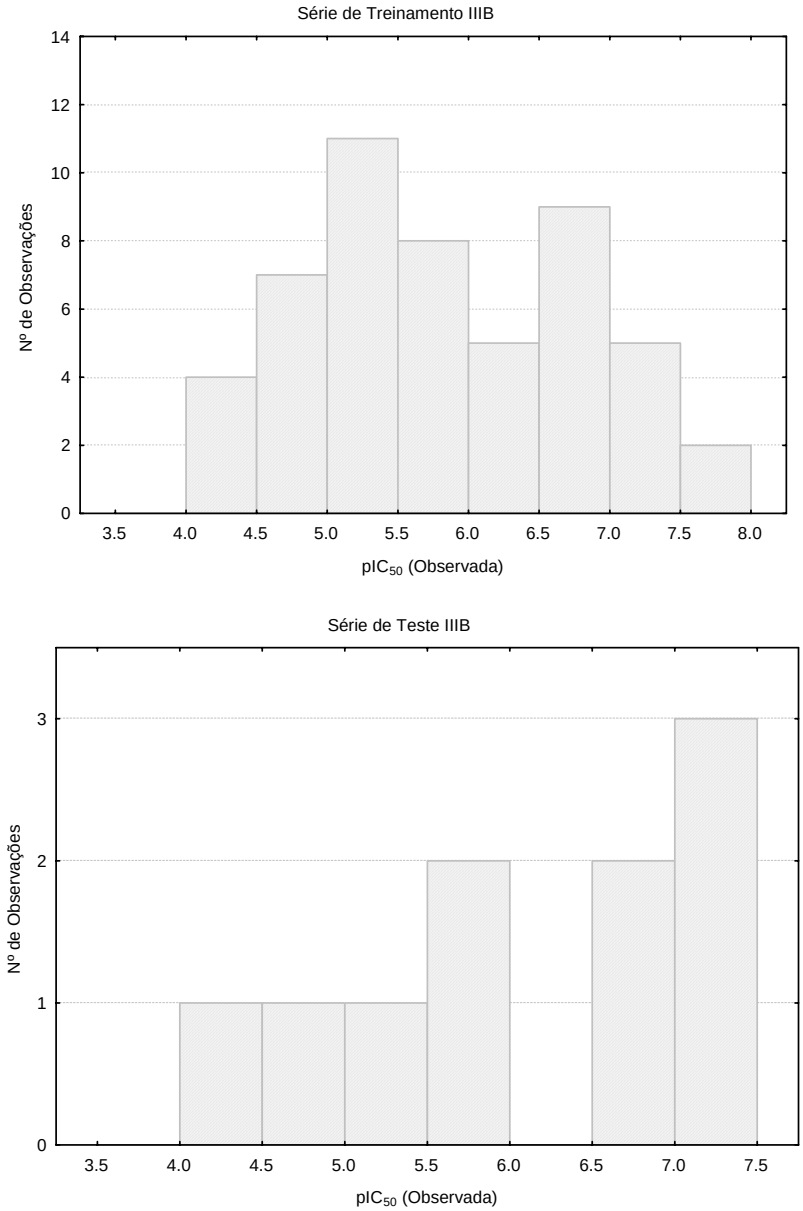


Figura II.5.1.3.2. Histogramas da distribuição dos valores de atividade pIC_{50} respectivamente das séries de treinamento e de teste IIIB.

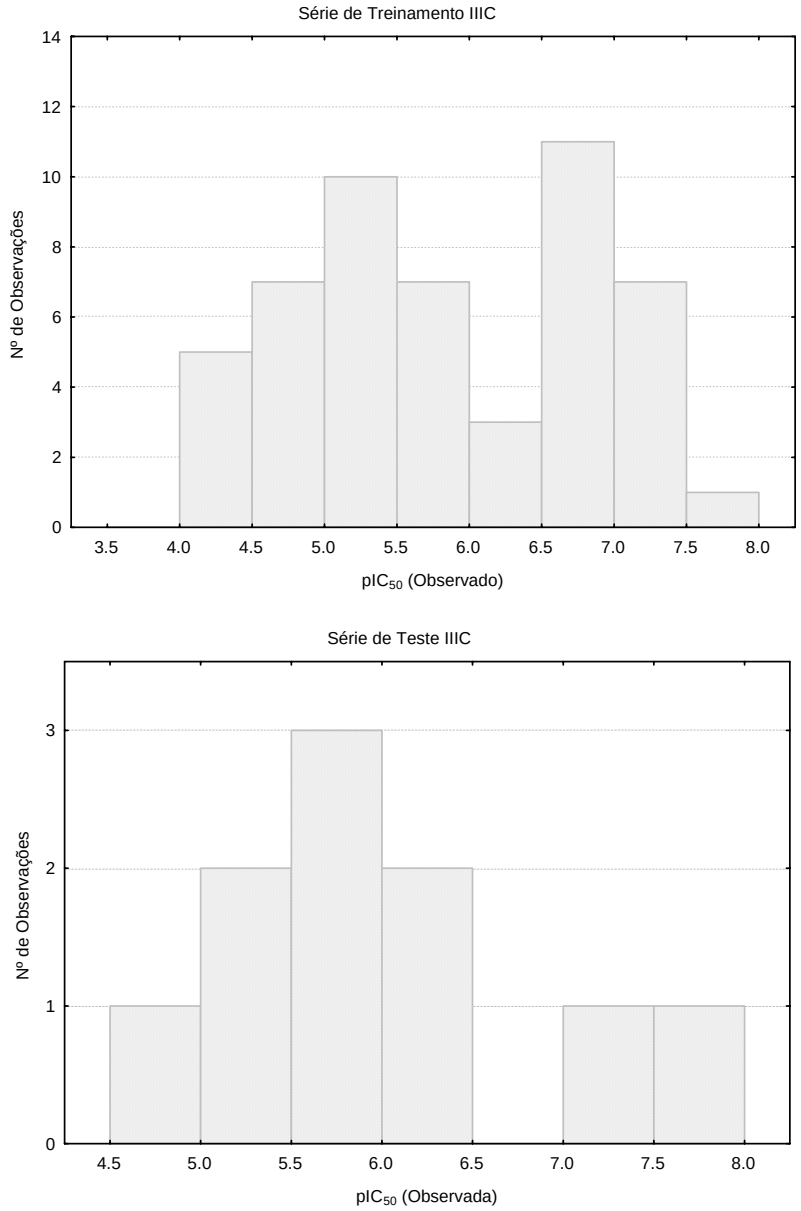


Figura II.5.1.3.3. Histogramas da distribuição dos valores de atividade pIC_{50} respectivamente das séries de treinamento e de teste IIIC.

II.5.2. Pré-tratamento de Dados

Considerando o apresentado e discutido (item I.8. *Data-Mining*), os critérios de pré-tratamento de dados utilizados para as três Séries I, II e III foram:

a. Retirada das variáveis que apresentavam valores iguais na série;
b. Retirada das variáveis que apresentavam apenas um valor diferente na série;
c. Retirada das variáveis que apresentavam valores além dos limites de 4 desvios padrão ($\pm 4\sigma$) em relação à média (Livingstone 1995). Nesta dissertação, daqui em diante este tratamento será denominado de 4σ, como por exemplo: uma variável que apresente um valor de média de 8,51 e um desvio padrão (σ) de 16,86, o valor do desvio padrão multiplicado por 4 é de 67,46, portanto o valor do limite inferior da série para esta variável é exatamente o valor da média subtraído 4 vezes o valor do desvio padrão (4σ), o resultado neste caso é –58,95, e o valor do limite superior é o valor da média somado 4σ (o resultado neste caso é 75,97), se houver na série o valor de 80 para esta variável ou –60, esta variável deve ser excluída;
d. Retirada das variáveis que apresentavam correlação maior que 0,95 com outras variáveis, sendo que é retirado o maior número de variáveis intercorrelacionadas permanecendo a variável independente que apresenta maior correlação com a variável dependente;

Para a Série II, foram feitos dois pré-tratamento de dados, um utilizando todos os passos citados anteriormente (a, b, c e d) e, um outro pré-tratamento sem aplicarmos o passo descrito no item c. (4σ). Portanto, para a Série II foi realizado duas vezes o processo de seleção das variáveis para compararmos a influência do pré-tratamento 4σ.

II.5.3. Seleção dos Descritores

II.5.3.1. Modelos *PLS*

Os descritores pré-selecionados foram submetidos à análise *PLS* (Geladi & Kowalski 1986) utilizando para a série de treinamento o método de *full cross-validation* (Clark & Cramer 1993).

Nas análises *PLS* foram gerados modelos com uma até quatro variáveis latentes. Para tanto utilizou se os programas ANALYZE para a Série I e o módulo QSAR do programa SYBYL respectivamente para as Séries II e III, com a finalidade de se extrair os descritores que possuíam as correlações mais significativas com os valores das atividades inibitórias correspondentes. Nesta etapa, foram selecionados, como já foi descrito no item II.5.1. *Divisão das Séries*, respectivamente três grupos de treinamento e de teste, visando abranger toda a faixa de valores de pIC_{50} respectivamente para as Séries I, II e III.

Para a seleção do melhor modelo em uma série, foram utilizados o seguinte critérios (Clark & Cramer 1993):

- Maior valor do coeficiente de predição gerado pelo método *full cross-validation* nas séries de treinamento (Q^2_{cv}), expresso pela equação I.6.9; (Cramer *et al.* 1988);

II.5.3.2 Critério de Validação de Modelos Proposto por Tropsha

Os critérios de validação de modelos desenvolvido por Alexander Tropsha (Golbraikh *et al.* 2001;Golbraikh & Tropsha 2002; Shen *et al.* 2003; Tropsha *et al.* 2003; Golbraikh & Tropsha 2003), descritos no item I.5.2.4, foram aplicados aos modelos *PLS* obtidos no item II.5.3.1.

II.5.3.3. Análise dos Valores de Freqüência de Presença

Após a escolha dos modelos com o número de variáveis latentes e variáveis originais obtidos, foram selecionados os descritores que estiveram presentes mais vezes entre os modelos selecionados.

Os valores da freqüência de presença de cada descritor nos três grupos de treinamento e de teste foram observados. Como critério, apenas os descritores que apresentaram um valor de freqüência maior que 65% nos três modelos foram selecionados como potencialmente válidos para a posterior análise de *QSAR* clássico.

II.5.4.Validação dos Critérios de Seleção de Variáveis

II.5.4.1 Análise do Modelo de *QSAR* Clássico Gerado

Os descritores selecionados pela análise de freqüência descrita no item II.5.3.3 foram validados analisando-se o modelo de *QSAR* clássico gerado utilizando estes descritores como parâmetros da equação de regressão, considerando-se os correspondentes valores dos parâmetros estatísticos. Foram gerados modelos com até 5 variáveis utilizando-se o programa BILIN (Kubinyi 1995). Os modelos com os maiores valores dos coeficientes de predição e estatística F foram selecionados.

II.5.4.2 Aplicação da Regra QUIK

A regra QUIK, explicada no item I.5.2.1, para validação de modelos gerados por MLR (regressão múltipla linear) (Todeschini *et al.* 1999;Todeschini *et al.* 2004) foi aplicada nos modelos de QSAR clássicos obtidos no item II.5.4.1.*Análise do Modelo de QSAR Clássico gerado*.

II.5.4.3 Aplicação da Regra do Q^2 Assintótico

A regra do Q^2 assintótico , explicada no item I.5.2.2, para validação de modelos gerados por MLR (regressão múltipla linear) (Todeschini *et al.* 2004) foi aplicada nos modelos de QSAR clássicos obtidos no item II.5.4.1.*Análise do Modelo de QSAR Clássico gerado*.

II.5.4.4 Aplicação das Regras baseadas nas funções R^P e R^N.

As regras baseadas nas funções R^P e R^N, explicadas no item I.5.2.3, para validação de modelos gerados por MLR (regressão múltipla linear) (Todeschini *et al.* 2004) foram aplicadas nos modelos de QSAR clássicos obtidos no item II.5.4.1. *Análise do Modelo de QSAR Clássico gerado*.