Quantitative structure-property relationship (QSPR) study of n-octanol-water partition coefficients (logPo/w) of fatty acids using multiple linear regression (MLR)

Sadeghali BAVAFA*[a], Mona MAHBOUBI[b], Reza BEHJATMANESH ARDAKANI[b] and Farzane FARAJIAN MASHHADI[c].

[a] *International University of Chabahar (IUC), Chabahar, Iran*
[b] *Chemistry Department, Payame Noor University, Teheran 19395-4697, Islamic Republic of Iran*
[c] *Department of Pharmacology, University of Medical Sciences, Zahedan, Iran*

Abstract The training set of 20 fatty acids with regularly distributed logPo/w values was used to assess the predictive ability of the QSPR/QSAR models produced in the regression. All the structures studied in this work were optimized by using B3LYP method in conjunction with 6-31G* basis set. Statistical characteristics of the best model are the following: n = 20, R^2=0.999, R^2_{CV} = 0.997, F =2938, standard error (SE) = 0.148 and Durbin-Watson (DW) =2.606

Keywords: QSPR, Partition coefficients (logPo/w), Fatty acids, Multiple linear regression (MLR).

1. Introduction

In the fields of organic and medicinal chemistry, a partition (P) is the ratio of concentrations of a compound in the two phases of a mixture of two immiscible solvents at equilibrium [1]. Normally one of the solvents chosen is water while the second is hydrophobic such as octanol [2]. Hence both the partition and distribution coefficient are measures of how hydrophilic (water loving) or hydrophobic (water fearing) a chemical substance is. Partition coefficients are useful for example in estimating distribution of drugs within the body. Hydrophobic drugs with high partition coefficients are preferentially distributed in hydrophobic compartments such as lipid bilayers of cells while hydrophilic drugs (low partition coefficients) preferentially are found in hydrophilic compartments such as blood serum. The logarithm of this coefficient, logPo/w, has been shown to be one of the key parameters in quantitative structure–activity/property relationship (QSAR/QSPR) studies. There are some reports about the applications of MLR [3-6] and artificial neural network [7-10] modeling to predict the n-octanol/water partition coefficient of organic compounds. In chemistry, especially biochemistry, a fatty acid is a carboxylic acid often with a long unbranched aliphatic tail (chain), which is either saturated or unsaturated (ω-3 and ω-6). Fatty acids are aliphatic monocarboxylic acids derived from, or contained in esterified form in an animal or vegetable fat, oil, or wax. Unsaturated fatty acids are of similar form, except that one or more alkenyl functional groups exist along the chain, with each alkene substituting a single-bonded "$-CH_2-CH_2-$" part of the chain with a double-bonded "-CH=CH-" portion (that is, a carbon double-bonded to another carbon) [11]. n−6 fatty acids (popularly referred to as ω−6 fatty acids or omega-6 fatty acids) are a family of unsaturated fatty acids which have in common a final carbon–carbon double bond in the n−6 position; that is, the sixth bond from the end of the fatty acid. Some medical research suggests that excessive levels of n−6 fatty acids, relative to n−3 fatty acids, may increase the probability of a number of diseases and depression [12-13].

ISSN-1223-7221

Table 1. Experimental values of logPo/w for fatty acids.

NO.	**common name**	**LogP.** ***Exp***[a]	**NO.**	**common name**	**LogP.** ***Exp***[a]
1	Propionic acid[S]	0.33	**26**	Triacontanoic acid[S]	13.84
2	Butanoic acid[S]	0.79	**27**	henatriacontylic[S]	No. Exp
3	Valeric acid[S]	1.39	**28**	lacceric acid[S]	No. Exp
4	Heptanoic acid[S]	2.42	**29**	psyllic acid[S]	No. Exp
5	caprylic acid[S]	3.05	**30**	geddic acid[S]	No. Exp
6	pelargonic acid[S]	3.42	**31**	cerpolastic acid[S]	No. Exp
7	capric acid[S]	4.09	**32**	Oleic acid[U]	7.73
8	undecylic acid[S]	4.42	**33**	Erucic acid[U]	9.69
9	lauric acid[S]	4.6	**34**	Alpha-Linolenic acid[U]	6.46
10	Tridecanoic acid[S]	5.49	**35**	eicosatrienoic acid[U]	No. Exp
11	myristic acid[S]	6.11	**36**	Eicosatetraenoic acid[U]	No. Exp
12	pentadecylic acid[S]	6.47	**37**	Eicosapentaenoic acid[U]	No. Exp
13	Palmitic acid[S]	7.17	**38**	Docosapentaenoic acid[U]	No. Exp
14	Heptadecanoic acid[S]	7.45	**39**	Docosahexaenoic acid[U]	No. Exp
15	Stearic acid[S]	8.23	**40**	tetracosahexaenoic acid[U]	No. Exp
16	nonadecylic acid[S]	8.44	**41**	Linoleic acid[U]	7.05
17	arachidic acid[S]	9.29	**42**	gamma-linolenic acid[U]	No. Exp
18	heneicosylic acid[S]	No. Exp	**43**	Eicosadienoic acid[U]	6.251
19	Docosanoic acid[S]	9.91	**44**	Dihomo-gamma-linolenic acid[U]	No. Exp
20	Tricosanoic acid[S]	No. Exp	**45**	arachidonic acid[U]	6.98
21	lignoceric acid[S]	No. Exp	**46**	docosadienoic acid[U]	No. Exp
22	cerotic acid[S]	No. Exp	**47**	adrenic acid[U]	No. Exp
23	heptacosylic acid[S]	No. Exp	**48**	calendic acid[U]	No. Exp
24	montan wax[S]	No. Exp	**49**	Palmitoleic acid[U]	6.75
25	Hexanoic acid[S]	1.84			

S: Saturated, U: Unsaturated, Exp: Experimental, a: No unit

2. Experimental

In this paper, we design a QSPR model for some fatty acids by using quantum chemical and structural descriptors. **Table 1** shows the name of different compounds taken for this study. This table contains 31 saturated and 18 unsaturated fatty acids. List of descriptors is shown in **Table 2**. Except 6 structural descriptors containing Mm, MV, NH, NC, NSB and NDB, all other descriptors are taken from the results of quantum chemical calculations.

Gaussian 2003 (GW03)TM program package [14] has been used for calculation of quantum chemical descriptors. To do this, at first, all molecules are drawn in GaussviewTM version 3 and they model builded. As a second step, these structures are saved in Gaussian job function 'gjf' format. Then, these input 'gjf' files are opened in the GW03 program. Results of calculation are from using two keywords FOPT and FREQ. FOPT, that is full optimization, was carried out by the level B3LYP that is a kind of Density Function Theory (DFT) method. 6-31G* basis set was used during all calculations. To obtain statistical mechanical (LOG10(Q), S, Cv) and thermochemical (Hf,E^0, E, H and G) descriptors and to be sure that the optimized structures are all in minimum point of potential energy surface, frequency analysis has been used. NIMAG=0 shows that the number of imaginary frequencies are equal to zero and that the structure is really a stationary minimum point and not a transition state. All calculations have been done by a single processor Pentium 4 computer. Descriptors from number 9 to 22, except 19, were taken from NBO analysis. We have used NBO version 3.1 that is called by POP=NBO in the GW03 program [15] All statistical analyses were performed using SPSS version 16 program [16] Physicochemical properties activitie of fatty acids, such as n-octanol/water partition coefficient (logPo/w) play a major role in determining the distribution of fatty acids. Numerical data on the octanol/water partition coefficient (logPo/w) are taken from Ref 17-18.

In QSPR, molecular descriptors (X) are correlated with one or more response variable (y).
If it is assumed that the relationship is well represented by a model that is linear in the regressed variables, a suitable model may be as follows:

Table 2. Symbols and definitions of the molecular descriptors used in the present study.

Nr.	Descriptor	Interpretation
1	LOG10(Q)	Partition function
2	S	Entropy
3	CV	constant volume molar heat capacity
4	E^0	sum of electronic and zero-point Energies
5	E	sum of electronic and thermal Energies
6	H	Sum of electronic and thermal Enthalpies
7	G	sum of electronic and thermal Free Energies
8	Mm	Molecular mass
9	E_{HOMO}	energy of the highest occupied molecular orbital
10	E_{LUMO}	energy of the lowest unoccupied molecular orbital
11	μ	chemical potential
12	η	chemical hardness
13	ω	electrophilicity
14	Q^-	The largest negative atomic charge on an atom
15	Q^+	The largest positive atomic charge on an atom
16	QO	Sum of absolute values of atomic charge on oxygen
17	QC	Sum of absolute values of atomic charge on carbon
18	QH	Sum of absolute values of atomic charge on hydrogen
19	Hf	Heat of formation
20	Core	Core-Core repulsion
21	Valence	Valence
22	Rydberg	Rydberg
23	MV	Molar Volume
24	NH	number of hydrogen
25	NC	number of carbon
26	NSB	number of single bond
27	NDB	number of double bond

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + e \qquad (1)$$

In Eq. (1) the b's are unknown constants called regression coefficients and the objective of regression analysis is to estimate these coefficients.

The statistical parameters used to assess the quality of the models are the Prediction Error Sum of Squares (PRESS) of validation (Eq. (2)) and The leave-one-out (LOO) cross-validation correlation coefficient or cross-validated explained variance (R_{cv}^{2}[19-20]).

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (2)$$

$$R_{cv}^2 = 1 - \left(\frac{PRESS}{\sum_{i=1}^{n} (y_i - \bar{y})} \right) \qquad (3)$$

In these equations, n is the number of compounds used for cross validation, y_i is the experimental value of the physicochemical property for the ith sample and $\hat{y}_i$ is the value predicted by the model built without sample i: PRESS is the prediction error sum of squares for all samples included in the model. The correlation between the variables in the model was estimated by the variance inflation factor (VIF). VIF is equal to 1/ (1-r^2), in which r is the correlation coefficient of multiple regressions between one variable and the others in the equation. If value of VIF_j is over 10, there is a high correlation between the variable x_j and others, and the regression model is not a stable one.

3. Results and discussion

In the present study, the QSPR model is generated using a training set of 20 molecules. The training set of 20 molecules (**Table 3**) with regularly distributed logP*o/w* values is used to assess the predictive ability of the QSPR models produced in the regression.

The statistical processing to obtain the QSPR model is carried out by using the stepwise multiple linear regression that is based on the forward-selection and backward-elimination methods, where the independent variables are individually added or deleted from the model at each step of the regression depending on three criteria: Prediction Error Sum of Squares, Standard Error of Validation and standard correlation coefficient variables are selected to enter or to remove until the 'best' model is obtained. The result shows that logP*o/w* is highly dependent on the NSB and E_{LUMO}. Unstandard equation from stepwise MLR calculations is as follows:

$$LogPo/w=-1.485+0.174NSB+18.918E_{LUMO} \qquad (4)$$

Sig = 0.000, R^2 = 0.999, R^2_{cv} = 0.997, F = 2938, standard error (SE) = 0.148 and Durbin-Watson (DW) =2.606

Without standardization of above equation, it is not possible to discuss about importance of variables in the prediction model. Following equation is obtained after standardization:

$LogPo/w = 1.012NSB+0.116E_{LUMO}$ (standardized coefficient) (5)

In equation (4) the coefficient of E_{LUMO} is greater than the coefficient of NSB, but this is a wrong conclusion, if we say that E_{LUMO} variable is more important than NSB. Contrary to the equation (4), the corrected standardized coefficients show that NSB is much more important than E_{LUMO}. Moreover, it is expected that in a series of fatty acids of varying chain length, logP*o/w* will increase gradually with increase of chain length. This is reflected in the use of NSB parameter as one the descriptors in Eq. (5) unfortunately, in some papers in this subject, some authors compare importance of variables in an unstandardized equation that in a certain condition (such as here) may be lead to a wrong result. **Fig 1a** shows relation between experimental and predicted logP*o/w* values. Correlation coefficient (R^2) for this curve is equal to 0.997.

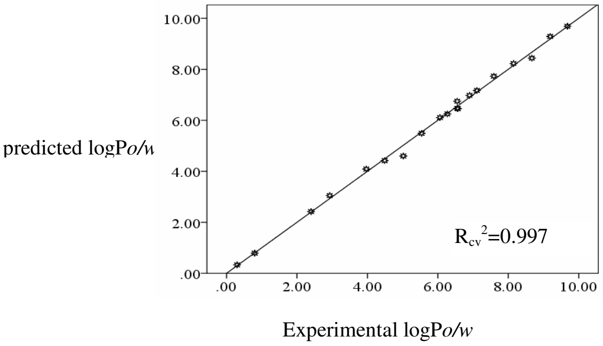

Fig 1a. Experimental versus predicted values of logP*o/w* with NSB and LUMO as descriptor.

Table 3. Experimental and Predicted values of logP*o/w* for fatty acids (training set).

Nr.	**Common name**	**LogP. *Exp*^{a}**	**LogP. *Pred*^{a}**	**NSB**	**E_{LUMO}(ev)**	**Residual**
1	Propionic acid^{S}	0.33	0.3	9	0.0116	0.03
2	Butanoic acid^{S}	0.79	0.8	12	0.01039	-0.01
3	Heptanoic acid^{S}	2.42	2.4	21	0.01243	0.02
4	caprylic acid^{S}	3.05	2.93	24	0.01248	0.12
5	capric acid^{S}	4.09	3.97	30	0.01254	0.12
6	undecylic acid^{S}	4.42	4.49	33	0.01256	-0.07
7	lauric acid^{S}	4.6	5.02	36	0.01257	-0.42
8	Tridecanoic acid^{S}	5.49	5.54	39	0.01258	-0.05
9	myristic acid^{S}	6.11	6.06	42	0.01259	0.05
10	pentadecylic acid^{S}	6.47	6.55	45	0.01082	-0.08
11	Palmitic acid^{S}	7.17	7.11	48	0.01263	0.06
12	Stearic acid^{S}	8.23	8.15	54	0.01259	0.08
13	nonadecylic acid^{S}	8.44	8.67	57	0.01261	-0.23
14	arachidic acid^{S}	9.29	9.19	60	0.01261	0.1
15	arachidonic acid^{U}	6.98	6.9	48	0.00162	0.08
16	Erucic acid^{U}	9.69	9.68	63	0.01094	0.01
17	Palmitoleic acid^{U}	6.75	6.55	45	0.01066	0.2
18	Oleic acid^{U}	7.73	7.59	51	0.01076	0.14
19	Alpha-Linolenic acid^{U}	6.46	6.57	45	0.01193	-0.11
20	Eicosadienoic acid^{U}	6.251	6.27	51	-0.05921	-0.02

a: No unit, S: Saturated, U: Unsaturated, Exp: Experimental Pred: Predicted, NSB: Number of Single Bond

Table 4. Experimental and Predicted values of logP*o/w* for fatty acids (test set).

test set	R_{cv}^2 = 0.999	F=19758.215	standard error (SE)=0.079	Durbin-Watson (DW)=1.461		
NO.	**common name**	**LogP. *Exp*^{a}**	**LogP. *Pred*^{a}**	**NSB**	**E_{LUMO}(ev)**	**Residual**
1	Valeric acid^S	1.39	1.32	15	0.0106	0.07
2	Hexanoic acid^S	1.84	1.88	18	0.0123	-0.04
3	pelargonic acid^S	3.42	3.45	27	0.0125	-0.03
4	Heptadecanoic acid^S	7.45	7.59	51	0.0108	-0.14
5	Triacontanoic acid^S	13.84	14.41	90	0.0126	-0.57
6	Docosanoic acid^U	9.91	10.24	66	0.0126	-0.33
7	Linoleic acid^U	7.05	7.1	48	0.0121	-0.05

a: No unit, S: Saturated, U: Unsaturated,Exp: Experimental, Pred: Predicted

The agreement observed between the predicted and experimental logP*o/w* values in **Fig. 1b** confirms a good predictive ability of MLR modeling.

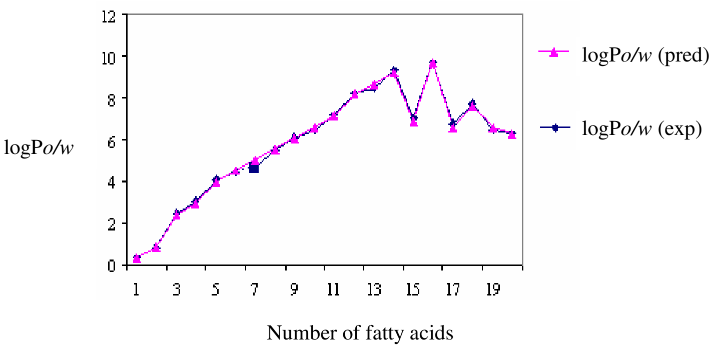

Fig. 1b. Plots of experimental and predicted logPo/w values versus sample number in the training set.

Table 4 shows the results of prediction of the model for five saturated and two unsaturated fatty acids as a test set. R_{cv}^2, F, standard error (SE), Durbin-Watson (DW) and residual value show that the model predictions are very good.

Multicollinearity between the descriptors of the Eq. (4) were checked by calculating their variation inflation factors (VIF) to evaluate the correlation value between independent variables in the equation. The self-correlation coefficients of the independent variables in Eq. (4) are listed in **Table 5**.

Table 5. Self-correlation coefficient of independent variables in Eq. (4)

Equation	Variable	VIF
logP*o/w* = -1.485+0.174NSB+18.918E_{LUMO}	NSB	1.031
	E_{LUMO}	1.031

The table shows that the VIF values for Eq. (4) are all less than 2.0, and no intercorrelation exists for the selected variables. The residuals of the MLR calculated values of the logP*o/w* are plotted against the experimental values in **Fig. 1c.**

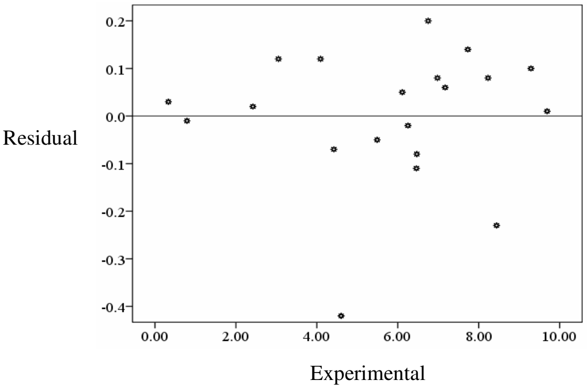

Fig. 1c. Plot of predicted logPo/w against the experimental logPo/w values.

The propagation of the residuals on both sides of the zero line indicates that no systematic error exists in the development of the MLR. One of the important characteristics of MLR models is the distribution of errors. For a good MLR model the distribution should be normal. The shape of the histogram should approximately follow the shape of the normal curve. This is shown in **Fig. 1d.**

There is not logP*o/w* information for some important saturated and unsaturated fatty acids in the literature

Table 6 shows the predictions of the model for these compounds.

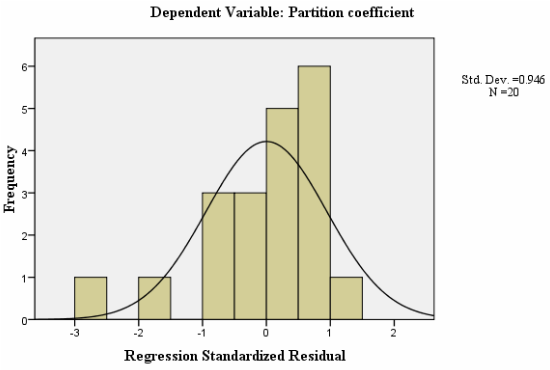

Fig. 1d. Frequency of faty acids and descriptor's range for NSB and LUMO.

Table 6. Predicted logPo/w for some important saturated and unsaturated fatty acids.

N	Common name	LogP. *Exp*[a]	LogP. *Pred*[a]	NSB	E_{LUMO}(ev)
1	heneicosylic acid[S]	No. Exp	9.72	63	0.0126
2	Tricosanoic acid[S]	No. Exp	10.76	69	0.0126
3	lignoceric acid[S]	No. Exp	11.28	72	0.0126
4	cerotic acid[S]	No. Exp	12.29	78	0.0108
5	heptacosylic acid[S]	No. Exp	12.85	81	0.0126
6	montan wax[S]	No. Exp	13.37	84	0.0126
7	nonadecylic acid[S]	No. Exp	13.89	87	0.0126
8	lacceric acid[S]	No. Exp	15.46	96	0.0126
9	psyllic acid[S]	No. Exp	15.98	99	0.0126
10	geddic acid[S]	No. Exp	16.50	102	0.0126
11	cerpolastic acid[S]	No. Exp	17.02	105	0.0126
12	eicosatrienoic acid[S]	No. Exp	6.05	51	-0.0710
13	Eicosatetraenoic acid[U]	No. Exp	5.38	48	-0.0786
14	Eicosapentaenoic acid[U]	No. Exp	4.76	45	-0.0840
15	Docosapentaenoic acid[U]	No. Exp	5.80	51	-0.0840
16	Docosahexaenoic acid[U]	No. Exp	5.20	48	-0.0879
17	tetracosahexaenoic acid[U]	No. Exp	6.08	53	-0.0879
18	gamma-linolenic acid[U]	No. Exp	6.48	45	0.0072
19	Dihomo-gamma-linolenic acid[U]	No. Exp	7.62	51	0.0122
20	docosadienoic acid[U]	No. Exp	7.83	60	-0.0592
21	adrenic acid[U]	No. Exp	8.14	54	0.0120
22	calendic acid[U]	No. Exp	5.48	44	-0.0364

4. Conclusion

The success of any QSPR model depends on the selection of appropriate descriptors.
Exploring the usefulness of descriptors, especially, conceptual DFT based descriptors along with other descriptors and analyzing their applicability could lead to drastic improvement in QSPR models.

Based on this fact, structure–property relationship for the data set containing 49 fatty acids congeners on the lipophilic behaviour (logP*o/w*) is analyzed. Traditional regression procedures along with cross-validation are carried out to evaluate the predicting power of the developed model. It has been shown that using the entire data set, the number of single bond index NSB and ELUMO descriptors provides a reasonably good coefficient of determination (R^2 = 0.999) and cross-validated squared correlation coefficient R^2_{cv} = 0.997 value indicating the significance of the developed model.

5. References

* E-mail address: s.ali.bavafa@gmail.com

[1]. A. Leo, C. Hansch and D. Elkins, Chem Rev. **71**, 525–616 (1971).

[2]. J.Sangster, Fundamentals and Physical Chemistry, John Wiley & Sons. 1997, pp. 178.

[3]. I. Moriguchi, S. Hirono, I. Nakagome and H .Hirano, Chem. Pharm. Bull. **42**, 976 (1994)

[4]. W.M. Meylan and P.H. Howard, Pharm. J. Sci. **84**, 83 (1995).

[5]. V.K. Gombar and K. Enslein, J. Chem. Inf. Comput. Sci. **36**, 1127 (1996).

[6]. S.C. Basak, B.D. Gute and G.D. Grunwald, J. Chem. Inf. Comput. Sci. **36**, 1054 (1996).

[7]. J.J. Huuskonen, D.S. Livingstone and I.V. Tetko, J. Chem. Inf. Comput. Sci. **40**, 947 (2000).

[8]. I.V. Tetko, V.Y.Tanchuk and A.E. P. Villa, J. Chem. Inf. Comput. Sci. **41**, 1407 (2001).

[9]. L. Molnar, G.M.Keseru, A.Papp, Z. Gulyas and F. Daras, Bioorg. Med. Chem. Lett. **14**, 851 (2004).

[10]. A.F. Dupart, T.Huynh and G. Dreyfus, J. Chem. Inf. Comput. Sci. **38**, 586 (1998).

[11]. IUPAC Compendium of Chemical Terminology (2nd ed.). International Union of Pure and Applied Chemistry. Retrieved on 2007-10-31.

[12]. Joseph. R. Healthy intakes of n−3 and n−6 fatty acids: estimations considering worldwide diversity. American Journal of Clinical Nutrition (American Society for Nutrition)., 2006, 83, 1483S–1493S.http://www.ajcn.org/cgi/content/full/83/6/S1483.

[13]. O. Hirohmi, I. Yuko, S. Yueji, H. Tomohito and L. William, ω3 fatty acids effectively prevent coronary heart disease and other late-onset diseases: the excessive linoleic acid syndrome. World Review of Nutritional Dietetics. **96**, 83–103 (2007).

[14]. M.J. Frisch et al., Gaussian 03, Revision B03, Gaussian Inc., Pittsburgh PA, 2003.

[15]. D.E. Glendening, A.E. Reed, J.E. Carpenter and F . Weinhold, NBO, Version 3.1

[16]. SPSS is a statistical software of SPSS Inc., USA

[17]. http://chem.sis.nlm.nih.gov/chemidplus

[18]. http://hmp.biology.ualberta.ca/~knox/hmdb