J|A|C|S
ARTICLES
Published on Web 01/06/2006

Combination of a Modified Scoring Function with Two-Dimensional Descriptors for Calculation of Binding Affinities of Bulky, Flexible Ligands to Proteins

Csaba Hetényi,*,† Gábor Paragi,‡ Uko Maran,§ Zoltán Timár,|| Mati Karelson,§ and Botond Penke‡,||

Contribution from the Department of Biochemistry, Eötvös Loránd University, 1/C Pázmány P. sétány, H-1117 Budapest, Hungary, Protein Chemistry Research Group, Hungarian Academy of Sciences, 8 Dóm tér, H-6720 Szeged, Hungary, Department of Chemistry, Tartu University, 2 Jakobi Street, EE−51014 Tartu, Estonia, and Department of Medical Chemistry, University of Szeged, 8 Dóm tér, H-6720 Szeged, Hungary

Received September 1, 2005; E-mail: csabahete@yahoo.com

Abstract: Bulky, flexible molecules such as peptides and peptidomimetics are often used as lead compounds during the drug discovery process. Pathophysiological events, e.g., the formation of amyloid fibrils in Alzheimer's disease, the conformational changes of prion proteins, or β-secretase activity, may be successfully hindered by the use of rationally designed peptide sequences. A key step in the molecular engineering of such potent lead compounds is the prediction of the energetics of their binding to the macromolecular targets. Although sophisticated experimental and in silico methods are available to help this issue, the structure-based calculation of the binding free energies of large, flexible ligands to proteins is problematic. In this study, a fast and accurate calculation strategy is presented, following modification of the scoring function of the popular docking program package AutoDock and the involvement of ligand-based two-dimensional descriptors. Quantitative structure−activity relationships with good predictive power were developed. Thorough cross-validation tests and verifications were performed on the basis of experimental binding data of biologically important systems. The capabilities and limitations of the ligand-based descriptors were analyzed. Application of these results in the early phase of lead design will contribute to precise predictions, correct selections, and consequently a higher success rate of rational drug discovery.

Introduction

Flexible, peptidic molecules are often involved in rational drug design. These compounds find various applications for important biochemical problems such as the inhibition of β-secretase,[1] a key enzyme in the pathomechanism of Alzheimer's disease,[1] or the blocking of various types of trypsins.[2] Similarly, the beta sheet breaker peptides have proved useful in hindering self-aggregation of the β-amyloid peptide of Alzheimer's disease and conformational changes in prion proteins of transmissible spongiform encephalophaties.[3] The number of such relevant applications of peptides as potent bioactive partners or lead compounds is still increasing. In rational drug discovery, estimation of the free energies of binding (ΔG_b) of bioactive ligands to their macromolecular targets is an essential step in the molecular engineering process. Although sophisticated methods do exist for the experimental measurement of binding thermodynamics (e.g., isothermal titration calorimetry[4]), they are usually time-consuming and/or require special conditioning for problematic cases such as amyloid aggregation.[5]

Different in silico strategies for the structure-based calculation[6] of ΔG_b have become an alternative to the instrumental techniques. One branch of these computational methods works on a statistical ensemble of structures produced by a molecular dynamics (MD) simulation. The MD-based techniques, e.g., the linear interaction energy method[7] supported by perturbation theory,[8] have been successfully applied to modified peptides,[9]

† Eötvös Loránd University.
‡ Hungarian Academy of Sciences.
§ Tartu University.
|| University of Szeged.

(1) (a) Hong, L.; Koelsch, G.; Lin, X.; Wu, S.; Terzyan, S.; Ghosh, A. K.; Zhang, X. C.; Tang, J. *Science* **2000**, *290*, 150−153. (b) Ghosh, A. K.; Shin, D.; Downs, D.; Koelsch, G.; Lin, X.; Ermolieff, J.; Tang, J. *J. Am. Chem. Soc.* **2000**, *122*, 3522−3523. (c) John, V.; Beck, J. P.; Bienkowski, M. J.; Sinha, S.; Heinrikson, R. L. *J. Med. Chem.* **2003**, *46*, 4625−4630.
(2) Fodor, K.; Harmat, V.; Hetényi, C.; Kardos, J.; Antal, J.; Perczel, A.; Patthy, A.; Katona, G.; Gráf, L. *J. Mol. Biol.* **2005**, *350*, 156−169.
(3) (a) Soto, C.; Sigurdsson, E. M.; Morelli, L.; Kumar, R. A.; Castaño, E. M.; Frangione, B. *Nature Med.* **1998**, *4*, 822−826. (b) Soto, C.; Kascsak, R. J.; Saborio, G. P.; Aucouturier, P.; Wisniewski, T.; Prelli, F.; Kascsak, R.; Mendez, E.; Harris, D. A.; Ironside, J.; Tagliavini, F.; Carp, R. I.; Frangione, B. *Lancet* **2000**, *355*, 192−197. (c) Hetényi, C.; Körtvélyesi, T.; Penke, B. *Bioorg. Med. Chem.* **2002**, *10*, 1587−1593. (d) Hetényi, C.; Szabó, Z.; Klement, É.; Datki, Z.; Körtvélyesi, T.; Zarándi, M.; Penke, B. *Biochem. Biophys. Res. Commun.* **2002**, *292*, 931−936. (e) Dobson, C. M. *Nature* **2005**, *435*, 747−749.
(4) (a) Leavitt, S.; Freire, E. *Curr. Opin. Struct. Biol.* **2001**, *11*, 560−566. (b) Campoy, A. V.; Freire, E. *Biophys. Chem.* **2005**, *115*, 115−124.
(5) Kardos J.; Yamamoto, K.; Hasegawa, K.; Naiki, H.; Goto Y. *J. Biol. Chem.* **2004**, *279*, 55308−55314.
(6) (a) Murphy, K. P. *Med. Res. Rev.* **1999**, *19*, 333−339. (b) Lazaridis, T. *Curr. Org. Chem.* **2002**, *6*, 1319−1332.
(7) (a) Åqvist, J. *J. Comput. Chem.* **1996**, *17*, 1587−1597. (b) Marelius, J.; Hansson, T.; Åqvist, J. *Int. J. Quantum Chem.* **1998**, *69*, 77−88.

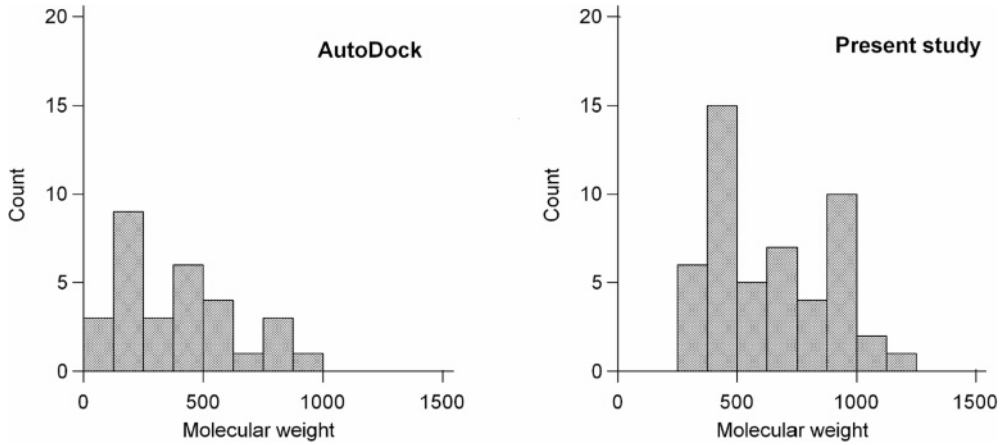

Figure 1. Distribution of molecular weights of the 30 ligands of the AutoDock calibration set[12] and the 50 compounds investigated in the present study. In the case of the present study, the number of compounds with higher molecular weights is significantly larger.

as well. Another strategy for the calculation of ΔG_b is the use of a single protein−ligand complex structure (preferably the crystallographic structure or an energy minimum). This approach requires a scoring function (SF), along with a parameter set appropriate for the type of ligand molecules investigated. The SFs developed for rapid calculation of ΔG_b are primarily implemented to drive the docking simulations.[10] In most of the cases they are parametrized for different types of small, druglike compounds to fit the requirements of the virtual high-throughput screening of compound libraries. It has been demonstrated in a number of studies that the crystallographic ligand positions in the protein−ligand complexes can be calculated precisely by using the appropriate SFs.[11] As SFs have been successfully used in calculations on various small compounds, it is a rational (but not trivial) wish to extend their applicability to larger, flexible ligands.

In the present study, the SF of the popular docking program package AutoDock 3.0[12] is tested and modified by using a set of flexible, peptidic ligands of biologically important complex systems. Predictive quantitative structure−activity relationships (QSARs) are developed for experimental ΔG_b values, using the modified SF of AutoDock and two-dimensional (2D) molecular descriptors of the ligand molecules. Our aim is to extend the capabilities of the SFs by means of easy-to-calculate ligand-based descriptors so as to develop a new, hybrid calculation strategy that combines advantages of the intermolecular terms of the SF and the ligand-based 2D descriptors for the rapid and accurate calculation of ΔG_b data for the problematic, bulky ligand molecules.

Methods

Protein−Ligand Systems. In the present study, 53 different protein−ligand complexes with known experimental values of ΔG_b ($\Delta G_{b(exp)}$) were involved. Complexes having large, peptidic ligands (MW > 350, Figure 1) and physiological importance (e.g., the "om"-series of β-secretase inhibitors; see Introduction for references on pathophysiological role of β-secretase) were prioritized for this study. Systems with di/tripeptide ligands were also selected to balance the structural data set. The atomic coordinates of 41 of the complexes, 1a30, 1abo, 1b05, 1b32, 1b3f, 1b3g,1b3l, 1b46, 1b51, 1b52, 1b58, 1b5i, 1b5j, 1b9j, 1bai, 1cka, 1fkn (om99-2), 1hhi, 1hhh, 1hhj, 1hhk, 1jet, 1jeu, 1jev, 1joj, 1k9r, 1m4h (om00-3), 1mcb, 1mcj, 1ody, 1qkb, 1str, 1vac, 1vwf, 2er9, 2rkm, 2vaa, 2vab, 4sga, 5sga, and 5er1 were obtained from the Protein Databank[13] (PDB). 12 β-secretase-inhibitor systems (om12, om13, om14, om15, om16, om17, om18, om19, om22, om23, om24, and om99-1)[14b] with no PDB structures available were modeled by modification of the 1fkn structure. $\Delta G_{b(exp)}$'s were compiled from previous studies.[14] Detailed data on the protein−ligand complexes and the corresponding codes are listed in the Supporting Information, Table A.

Molecular Modeling. The Babel,[15] Vega,[16] VMD,[17] and PyMol[18] packages were applied for file conversion, visualization, and modeling. Some of the GROMACS[19,20] topology files were generated with the program ProDrg.[21]

Molecular Mechanics Minimization. A standard routine was applied for all complexes to create a uniform set of coordinate files. The GROMACS program package and the force field[19,20] and explicit SPC[22] water model were involved in the calculations. The protein−ligand complexes and surrounding water molecules were placed in a cubic box together with the appropriate amount of neutralizing counterions. Dissociable protons were added by a built-in GROMACS algorithm, except for the β-secretase complexes, where the active site

(8) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420−1426.
(9) Hansson, T.; Åqvist, J. *Protein Eng.* **1995**, *8*, 1137−1144.
(10) (a) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins* **2002**, *47*, 409−443. (b) Brooijmans, N.; Kuntz, I. D. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335−373. (c) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. *J. Med. Chem.* **2004**, *47*, 3032−3047.
(11) (a) Hetényi, C.; van der Spoel, D. *Protein Sci.* **2002**, *11*, 1729−1737. (b) Hetényi, C.; Maran, U.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1576−1583.
(12) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(13) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(14) (a) Donnini, S.; Juffer, A. H. *J. Comput. Chem.* **2004**, *25*, 393−411. (b) Ghosh, A. K.; Bilcer, G.; Harwood, C.; Kawahama, R.; Shin, D.; Hussain, K. A.; Hong, L.; Loy, J. A.; Nguyen, C.; Koelsch, G.; Ermolieff, J.; Tang, J. *J. Med. Chem.* **2001**, *44*, 2865−2868. (c) Wang, R.; Fang, X.; Lu, Y.; Wang, S. *J. Med. Chem.* **2004**, *47*, 2977−2980. (d) Turner, R. T.; Koelsch, G.; Hong, L.; Castenheira, P.; Ghosh, A.; Tang, J. *Biochemistry* **2001**, *40*, 10001−10006.
(15) Walters, P.; Dolata, M. S. Babel − A Molecular Structure Information Interchange Hub. Department of Chemistry, University of Arizona, Tucson, AZ 85721.
(16) Pedretti, A.; Villa, L.; Vistoli, G. *J. Mol. Graph.* **2002**, *21*, 47−49.
(17) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33−38.
(18) DeLano, W. L. *PyMol Molecular Graphics System*; DeLano Scientific: San Carlos, CA, 2002.
(19) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model* **2001**, *7*, 306−317.
(20) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Comm.* **1995**, *91*, 43−56.
(21) Schuttelkopf, A. W.; van Aalten, D. M. F. *Acta Crystallogr. D* **2004**, *60*, 1355−1363.
(22) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*. Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, 1981; pp 331−342.

was protonated according to the results of a recent study.[23] The systems were optimized with steepest descent and conjugate gradient methods at tolerance levels of 1000 and 600 kJ mol^{-1} nm^{-1} and maximum step sizes of 0.05 and 0.001 nm, respectively. The optimum coordinates of the protein and ligand molecules were extracted for the subsequent calculations. Whenever necessary (e.g., 1ody) the crystallographic water molecule was also extracted as an essential part of the active site of the protein.

Scoring. Grid maps of 120 × 120 × 120 grid points at a spacing of 0.375 Å were generated around the center of the ligand binding site by the utility Autogrid of the program package AutoDock 3.0.[12] Heavy atoms and polar H atoms of the protein molecules were supplied with Kollman's partial charges. Atomic solvation parameters and fragmental volumes were inserted via the utility Addsol.[12] Gasteiger charges[24] were assigned to the ligand molecules. Charges of apolar H atoms were merged with charges of the connecting C atoms and aromatic atoms were selected by the utility Autotors.[12] The free energies of binding of the ligands to the proteins were calculated by using the SF implemented in the program package AutoDock[12] (eq 1):

$$\Delta G_{AD} = f_{elec}\sum_{i,j}\frac{q_j q_i}{\epsilon(r_{ij})r_{ij}} + f_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + f_{hbond}\sum_{i,j}\xi(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + f_{sol}\sum_{i,j}S_i V_j e^{(-r_{ij}^2/2\sigma^2)} + T_{HBD} + T_{TOR}$$

where

$$T_{HBD} = \sum_i P_{HBD,i};\ P_{HBD,i} = \begin{cases} 0.118\ \text{kcal/mol} & \text{if atom}_i = \text{polar H (H in a polar covalent bond)} \\ 0.236\ \text{kcal/mol} & \text{if atom}_i = \text{O} \\ 0.000\ \text{kcal/mol} & \text{if atom}_i \neq \text{polar H or O} \end{cases}$$

$$T_{TOR} = P_{TOR}N_{TOR} \quad (1)$$

ΔG_{AD} (the calculated AutoDock binding free energy) is the sum of three intermolecular interaction energy terms, one desolvational free energy term (these four terms are referred to as "bimolecular" in the next sections) and two "monomolecular" terms describing hydrogen-bonding (T_{HBD}) and torsional penalties (T_{TOR}) of the ligand molecule. It should be noted that the original formula of the AutoDock SF[12] is reorganized in eq 1 to make a distinction between the bimolecular and the ligand-based (monomolecular) terms.

The *f* coefficients were determined empirically from a multilinear regression (MLR) to a set of 30 protein−ligand complexes (AutoDock calibration set) with known binding constants.[12] The indices *i* and *j* correspond to ligand and protein atoms, respectively. The Coulombic term includes the partial charges (*q*) and a distance-dependent dielectric permittivity value (ϵ).[25] A, B, C, and D are the Lennard−Jones parameters in the dispersion/repulsion (12−6) and H-bonding (12−10) formulas, and *r* denotes the distance between the atomic pairs. $\xi(t)$ is a directional weight depending on angle *t* at the H-bonds.[12] T_{HBD} accounts for the broken H-bonds between the ligand and solvent molecules, and it is calculated by summation of the P_{HBD} penalty constants for the polar H or O atoms in the ligand molecule. In practice, these constants are added to the appropriate atomic affinity grid maps during calculation. The value of P_{HBD} for polar H atoms was derived[12] as $P_{HBD} = 0.0656 \times 0.36 \times 5$ kcal/mol, where 0.0656 is f_{hbond}, the MLR coefficient, 0.36 is the proportion of H-bonding sites utilized on average, and 5 kcal/mol is the maximal well depth of the H-bonding interaction.[12] The constant P_{HBD} (for O atoms) is equal to 2 × P_{HBD} (for polar H's) counting for two possible H-bonds at O atoms. P_{TOR} has a constant (0.3113 kcal/mol) value per torsion. N_{TOR} is the number of free torsions in the ligand. The product (T_{TOR}) of P_{TOR} and N_{TOR} gives an estimate of the unfavorable torsional entropy loss upon ligand binding. S and V denote the solvation parameter and fragmental volume, respectively, in the solvation function of Stouten et al.[26] In the SF of AutoDock 3.0, only the C atoms of the ligand molecules are involved in the solvation model. The exponential term is an envelope function with a constant value[26] of $\sigma = 3.5$ Å. By elimination of T_{HBD}, T_{TOR}, or both terms, new, modified SFs (ΔG_H, ΔG_T, or ΔG_{TH}) are defined and applied in the present study.

Quantum Mechanics (QM) Calculations. At the ab initio level, the density functional method was used for calculation of the partial charges on the atoms of the ligand molecules.[27] The B3LYP functional and 6-311 basis set augmented with polarization functions were employed in the Gaussian98[28] calculations.

Development of Quantitative Structure−Activity Relationships (QSARs). The development and statistical analysis of the MLRs and the selection of 2D descriptors were achieved with the program package CODESSA (ver. 2.0).[29] The MLRs have the following general formula (eq 2):

$$\Delta G_{b(exp),j} = \sum_{i=1}^{n}\alpha_i D_{ji} + \text{constant};\quad (j = 1, 2, \ldots, N) \quad (2)$$

where *i* and *j* are the serial numbers of the descriptors and ligands, respectively, *N* is the total number of ligands (complex systems), *n* is the total number of descriptors involved in the model, D_{ji} denote the descriptors, and α_i's are the regression coefficients. The mean square errors and *t*-values of the regression coefficients, the *F*-values, the standard deviations (s^2), and the squares of the correlation coefficients (R^2) of the regressions were also calculated. The descriptor pool created with CODESSA formed the basis for the selection of ligand-based 2D descriptors (Supporting Information, Table B). The "best multilinear regression (BMLR)" procedure was applied for the development of QSAR models A and B (see Results and Discussion for the naming of QSARs). During the BMLR procedure the pool of descriptors is cleaned from insignificant descriptors ($R^2 < 0.1$) and the descriptors with missing values. In the following steps of BMLR, construction of the best two-parameter regression, the best three-parameter regression, etc. are done based on the statistical significance and noncollinearity criteria ($R^2 < 0.6$) of the descriptors. In BMLR, the descriptor scales are normalized, centered automatically, and the final result is given in natural scales. The final model has the best representation of the property in the given descriptor pool with the given number of parameters. Numerical values of the selected descriptors are tabulated in the Supporting Information, Table C. Having residuals ≥ 2.00 kcal/mol (QSAR B), three (codes 1hhj, om22, and om24) of the 53 systems were outliers and excluded from the final models. Two of them (om24 and 1hhj) were found to be outliers from models in other studies,[30,31] as well. Thus, QSARs with $N = 50$ systems and up to 3 descriptors were developed.

Results and Discussion

Test and Modification of the Scoring Function. For the 50 complexes of the present study the $\Delta G_{b(exp)}$'s had poor

(23) Park, H.; Lee, S. *J. Am. Chem. Soc.* **2003**, *125*, 16416−16422.
(24) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219−3228.
(25) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J. Comput.-Aided. Mol. Des.* **1996**, *10*, 293−304.
(26) Stouten, P. F. W.; Frömmel, C.; Nakamura, H.; Sander, C. *Mol. Simul.* **1993**, *10*, 97−120.
(27) Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964**, *136*, 864−871.
(28) Frisch, M. J. et al. *GAUSSIAN.98*, revision A.7; Gaussian,Inc.: Pittsburgh, PA, 1998.
(29) (a) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24*, 279−287. (b) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Reference Manual (ver. 2)*; Gainesville, Florida, 1994. (c) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027−1043.
(30) Tounge, B. A.; Reynolds, C. H. *J. Med. Chem.* **2003**, *46*, 2074−2082.
(31) Liu, Z.; Dominy, B. N.; Shakhnovich, E. I. *J. Am. Chem. Soc.* **2004**, *126*, 8515−8528.

Table 1. Correlation of Experimental and Calculated Binding Free Energy Values of the 50 Complexes[a,b]

terms excluded	scoring function (D_1) code	coefficient (α_1)	error of coeff.	*t*-value	R^2	R^2_{cv}	s^2	*F*-value
- - -	ΔG_{AD}	2.5622×10^{-1}	4.8939×10^{-2}	5.2355	0.364	0.323	3.27	27.41
	constant	−5.8868	6.2266×10^{-1}	−9.4542				
T_{HBD}	ΔG_H	2.9877×10^{-1}	4.3668×10^{-2}	6.8419	0.494	0.458	2.60	46.81
	constant	−4.5114	6.7514×10^{-1}	−6.6821				
T_{TOR}	ΔG_T	3.1491×10^{-1}	3.4981×10^{-2}	9.0021	0.628	0.601	1.91	81.04
	constant	−3.1140	6.6747×10^{-1}	−4.6653				
T_{HBD} and T_{TOR}	ΔG_{TH}	3.1686×10^{-1}	2.9505×10^{-2}	10.7392	0.706	0.684	1.51	115.33
	constant	−2.1434	6.4904×10^{-1}	−3.3023				

[a] Linear regressions (eq 2, $n = 1$) were performed using free energies calculated with the (modified) AutoDock SFs as descriptors (D_1) [b] ΔG_{AD} denotes the default AutoDock SF. ΔG_H, ΔG_T, and ΔG_{TH} denote the modified SFs with T_{HBD}, T_{TOR}, and both terms eliminated, respectively. Standard deviations (s^2), squares of the correlation coefficients (R^2), and leave-one-out cross-validated correlation coefficients (R^2_{cv}) of the regressions are tabulated.

correlation with the ΔG_{AD} values calculated with the original SF of eq 1 (squared correlation coefficient, $R^2 = 0.364$; Table 1). However, good correlation ($R^2 = 0.956$) was obtained[12] for the original calibration set of AutoDock 3.0. This apparent contradiction can readily be explained: ΔG_{AD} was originally calibrated on the basis of a diverse set of 30 druglike compounds, and the molecular weight distribution of the 30 ligands of the AutoDock calibration set[12] and that of the 50 ligands in the present study (Figure 1) are significantly different and shifted to larger molecular weights in the latter case. A plausible reason for the low R^2 value for the set of 50 ligands in the present study is the different compound composition from that for the calibration set. Thus, it is reasonable to re-examine the components of the original AutoDock SF using a set of bulky and flexible peptides in order to yield a better fit to the experimental binding free energies for ligands of this problematic type.

In accordance with this finding, eq 1 was inspected to select out terms that depend on the ligand and influence the efficiency of the scoring. One of the two ligand-based terms is T_{HBD}, which represents a penalty, i.e., the loss of free energy due to broken H-bonds between the ligand and water molecules during complex formation with the protein. The exclusion of T_{HBD} alone increases R^2 to 0.494 (ΔG_H). The other simple, ligand-based term in eq 1 is T_{TOR}, which accounts for the change in free energy upon freezing of the torsional degrees of freedom of the ligand. Elimination of this term results in a much better correlation (ΔG_T in Table 1; $R^2 = 0.628$) between the experimental and calculated ΔG_b's in comparison with ΔG_{AD}. Elimination of both terms yields $R^2 = 0.706$ (ΔG_{TH} in Table 1; Figure 2) and an s^2 of 1.51. This model is fairly promising in comparison with other ΔG_b calculators,[32] and therefore, ΔG_{TH} forms a good basis for further, predictive QSARs.

Similarly to the present results, the terms T_{HBD} and T_{TOR} were modified by other authors[33] in order to obtain a good binding free energy model for carbohydrate ligands. In a recent work,[2] the difference between the binding affinities of SGTI (*Schistocerca gregaria* trypsin inhibitor, a 35-amino-acid-long peptide) to two different trypsins was estimated correctly by elimination of these two ligand-based terms. It should be noted that the accumulation of constant penalties from T_{HBD} results in an erroneous positive sum of the free energy of binding for unusually large ligands such as SGTI.

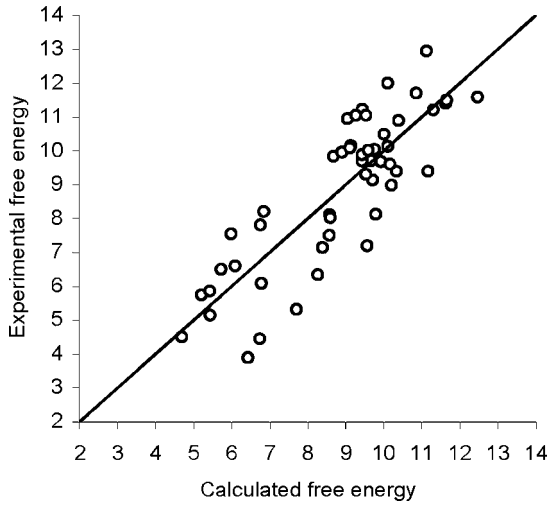

Figure 2. Correlation plot of experimental[14] and calculated binding free energy values (−kcal/mol) of the 50 complexes in the present study. Linear regression (eq 2, $n = 1$) was performed using free energies calculated with the modified SF ΔG_{TH} as a descriptor (D_1).

Development of QSARs Using ΔG_{TH} and Ligand-Based 2D Descriptors. The final correlation ($R^2 = 0.706$) obtained in the previous section is remarkably good showing the usefulness and good predictive power of the remaining bimolecular terms (ΔG_{TH}) having the original AutoDock parameters. Thus, instead of reparametrization of the whole SF, another strategy was followed in the present study. Keeping ΔG_{TH} as a descriptor, which can be reproducibly calculated for any protein−ligand complex structures, new, simple ligand-based descriptors were searched for in order to improve the correlation. Since both T_{HBD} and T_{TOR} can be derived from the 2D molecular graph without inclusion of any 3D information (eq 1),[12,33] the present search for ligand-based descriptors was restricted to 2D ones. A noteworthy advantage of 2D descriptors is that they are easy to calculate and require negligible computational time. Use of the CODESSA descriptor pool complemented with ΔG_{TH} furnishes the QSAR models in Table 2.

The best three-descriptor model (B) in Table 2 includes the bimolecular ΔG_{TH} as a major descriptor and two monomolecular, 2D descriptors, the $RPCG_{EN}$ (relative positive charge based on electronegativity), and the Balaban index (J) (Figure 3).

(32) (a) Böhm, H.-J. *J. Comput.-Aided. Mol. Des.* **1998**, *12*, 309−323. (b) Venkatarangan, P.; Hopfinger, A. J. *J. Med. Chem.* **1999**, *42*, 2169−2179. (c) Marder, M.; Estiú, G.; Blanch, L. B.; Viola, H.; Wasowski, C.; Medina, J. H.; Paladini, A. C. *Bioorg. Med. Chem.* **2001**, *9*, 323−335. (d) Wang, R.; Lai, L.; Wang, S. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 11−26. (e) Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. *J. Med. Chem.* **2002**, *45*, 2469−2483.
(33) Laederach, A.; Reilly, P. J. *J. Comput. Chem.* **2003**, *24*, 1748−1757.

Table 2. Correlation of Experimental and Calculated Binding Free Energy Values of the 50 Complexes[a,b]

	descriptor (D_i)								
QSAR	*i*	abbreviation	coefficient (α_i)	error of coeff.	*t*-value	R^2	R^2_{cv}	s^2	*F*-value
A	1	ΔG_{TH}	3.1216×10^{-1}	2.4686×10^{-2}	12.6456	0.799	0.774	1.05	93.36
	2	$RPCG_{EN}$	3.2582×10^{1}	6.9963	4.6571				
		constant	−4.1980	6.9930×10^{-1}	−6.0031				
B	1	ΔG_{TH}	2.7077×10^{-1}	2.2926×10^{-2}	11.8105	0.859	0.838	0.76	93.17
	2	$RPCG_{EN}$	5.7129×10^{1}	8.1307	7.0263				
	3	J	-6.2410×10^{-1}	1.4148×10^{-1}	−4.4113				
		constant	−4.6864	6.0281×10^{-1}	−7.7743				

[a] Multilinear regressions (eq 2, $n = 2$ or 3) were performed with ΔG_{TH} and ligand-based 2D descriptors. [b] $RPCG_{EN}$: electronegativity-based relative positive charge (Sanderson's electronegativity scheme). J: Balaban index. For other notes, refer to Table 1.

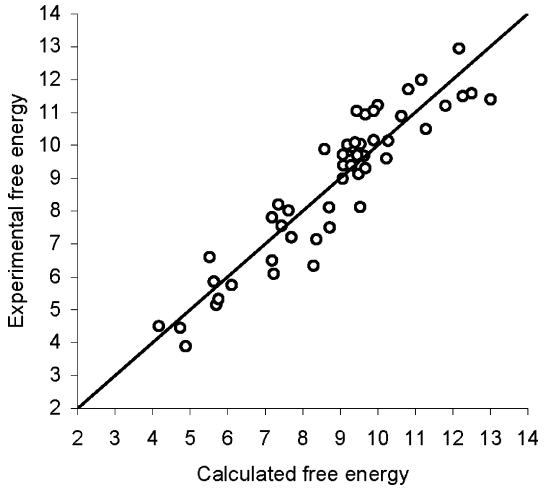

Figure 3. Correlation plot of experimental and calculated binding free energy values (−kcal/mol) of the 50 complexes in the present study in the case of QSAR B. The involvement of $RPCG_{EN}$ and *J* descriptors significantly improved the correlation as compared with Figure 2.

The $RPCG_{EN}$ values describe the distribution of positive partial charges in a molecule (eq 3):

$$RPCG_{EN} = \frac{\delta_{max}}{\sum_a \delta_a}; \quad a \in \{\delta_a > 0\} \tag{3}$$

where δ_{max} is the maximum value of the positive partial charges (charge excesses, δ_a) on the atoms (a) of the ligand molecule. In the CODESSA program, the δ values are assigned by a simple method,[34] which uses Sanderson's electronegativities of the atoms. Inspection of the δ_a values in our ligands reveals that most of them are located on H atoms connected to N or O atoms and on the C atoms of the amide bonds.

These H atoms with $\delta > 0$ are the possible H-bonding donor sites on the ligand molecules. Importantly, the regression coefficient of this descriptor is positive (Table 2), which means that it decreases the absolute value of the calculated binding free energy (the $RPCG_{EN}$ values are always positive, eq 3). Similarly, the eliminated T_{HBD} term contributed to ΔG_b with positive penalties, due to vanishing interactions between the ligand and water molecules. The RPCG descriptor was developed and used to account for the effects of polar intermolecular interactions.[35] These results let us conclude that $RPCG_{EN}$ describes (part of) the energy changes due to the altered H-binding system of the ligand during the attachment to a protein. To illustrate the molecular background of the $RPCG_{EN}$ descriptor, the systems 2rkm and 1vwf with ligands having maximum and minimum $RPCG_{EN}$ values (Supporting Information, Table C), respectively, are represented in Figure 4.

It can be seen that the dipeptide ligand (KK) in 2rkm is completely buried inside the protein, while in 1vwf a considerable interaction interface remains between the octapeptide ligand and the surrounding solvent. In 2rkm, the energy contribution of the $RPCG_{EN}$ term to ΔG_b in QSAR B is 6.24, whereas in 1vwf it is only 1.41 kcal/mol. Although the complete burial of a ligand can be considered as an extreme case, the probability of the use of a higher percentage of available H-bonding atoms in the new interactions with the protein is higher for smaller (dipeptide) rather than for larger (octapeptide) ligands. Consequently, the energy penalty corresponding to the loss of ligand-surrounding water interactions should be higher for 2rkm than for 1vwf. The $RPCG_{EN}$ descriptor correctly reflects this observation, as the fewer positively charged H atoms the molecule has, the smaller the denominator and the larger the $RPCG_{EN}$ value, i.e., the penalty (eq 3). If a ligand contains an atom with high δ_{max} (possibly buried into protein), this further increases the penalty. The similar argumentation is also valid for the vanished dipole−dipole interactions between the ligand and the surrounding water, pointing to the generality of $RPCG_{EN}$ descriptor. Besides, $RPCG_{EN}$ contains also indirect information on the size of the molecule via the sum of the partial positive charges (eq 3).

The size of the molecule is directly described by the Balaban index[36] (eq 4) that occurs as the third descriptor in QSAR B:

$$J = \left(\frac{q}{\mu + 1}\right) \sum_{i,j}^{q} (s_i s_j)^{-1/2}; \quad (\mu = q - n + 1) \tag{4}$$

where *q* is the number of edges in the molecular graph, *n* is the number of vertexes in the graph, μ is the cyclometric number, and s_i and s_j are the distance sums obtained by summation of row i and column i or row j and column j, respectively, of the distance matrix between the atoms in the molecule. In J, only the heavy atoms are considered in the molecular graph.

Thus, the J describes not only the size of the molecule but also its internal branching and distances. Interestingly, the number of free torsions (N_{tor}) is a part of the excluded term T_{TOR}, whereas the torsional tree of a ligand is also a type of branching. Considering this and the fact that the change in rotational entropy depends on the moments of inertia, i.e., the

(34) Zefirov, N. S.; Kirpichenok, M. A.; Ismailov, F. F.; Trofimov, M. I. *Dokl. Akad. Nauk.* **1987**, *296*, 883−887.
(35) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2323−2329.
(36) Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

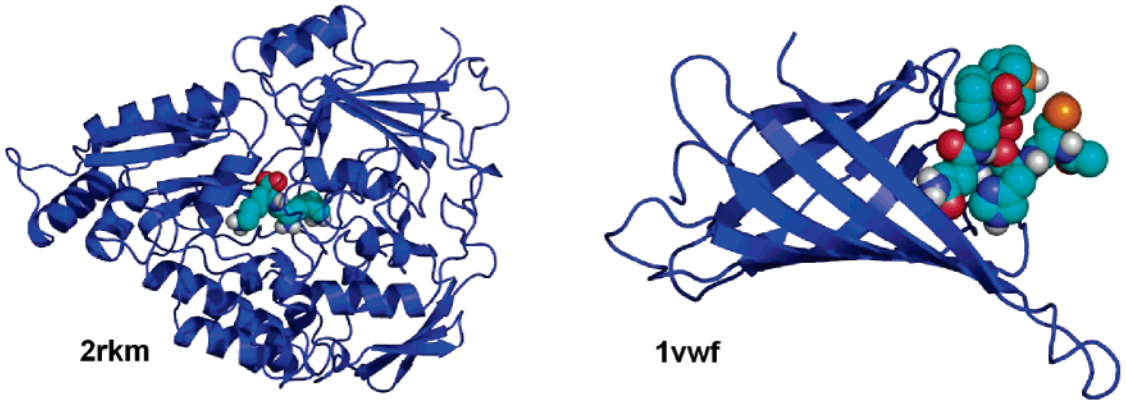

Figure 4. Small dipeptide ligand of the system 2rkm is buried deeply inside the protein, while the octapeptide ligand of 1vwf is sitting on the surface of the protein and its relatively large part can be involved in the ligand−solvent interaction; i.e., a small energy penalty occurs due to deceased H-bonds with the bulk solvent in the case of 1vwf. (Protein molecules and ligands are represented with cartoon and van der Waals surfaces, respectively.)

internal distances of the molecule, J may be descriptive of the change in free energy of binding upon the decrease of rotational and torsional degrees of freedom. In general, the J is based on the molecular structure according to graph theory and the distance matrix and reflects the relative connectivity and effective size of the flexible peptidic molecules. The magnitude of this descriptor increases with (i) an increase in branching and (ii) an increase in the number of atoms in the molecule. However, it would be a much more difficult task to give an analytical explanation for the role of the complex J descriptor than it was for $RPCG_{EN}$. QSAR B (Table 2) is comparable with other published ΔG_b calculators,[31−33,37] and as it concerns s^2, it is one of the best available calculators for the ΔG_b of large, flexible peptides.

Cross-Validation of the QSARs. The squared correlation coefficients of the leave-one-out cross correlation test (jackknife method) of QSARs are given in Table 2. These coefficients are fairly close to the original R^2's, emphasizing the statistical reliability of the models. The leave-20%-out test provides similarly good R^2 values: 0.780 and 0.848 for QSAR A and B, respectively. As a further test, it can be informative to separate a homogeneous subset of the 50 complexes and use the remaining systems as a training set to check the dependency of the results on this homogeneous part of the data. In our case, there is such a subset of 12 complexes (24%) among the 50, i.e., 12 of the 50 ligands investigated in this study have the same target protein (β-secretase) and are analogous in their structure, and the corresponding experimental inhibition constants used for calculation of the $\Delta G_{b(exp)}$'s were measured in the same laboratory.[14b,d] The results of this test for QSARs A and B are summarized in Table 3. It can be seen that, on the basis of the training set, good correlations are developed for the whole set of 50 points, and therefore, selection of the descriptors for the predictive QSARs is independent of the inclusion of the complexes of the homogeneous subset. Similar R^2 values (0.803 and 0.841 for QSAR-s A and B, respectively) can be calculated if correlating the predicted ΔG_b's of the subset of 12 systems (validation set) with the corresponding $\Delta G_{b(exp)}$'s, using the 38 systems as a training set.

Table 3. Cross-Validation Tests of Descriptor Sets of QSARs A and B Excluding a Homogeneous Subset of 24% of the Data Points[a,b]

	N = 38 (training set)				N = 50 (training set + 24% left out)	
QSAR	R^2	R^2_{cv}	s^2	F-value	R^2	s^2
A	0.841	0.813	0.85	92.36	0.797	1.10
B	0.893	0.868	0.59	94.26	0.857	0.79

[a] Multilinear regressions were trained for 38 of the 50 systems and tested on all 50 systems. [b] N corresponds to the number of systems (data points) used for the correlation. For other notes, refer to Table 1.

Robustness of the Second and Third Descriptors of the Models Obtained. The descriptor J is calculated directly from the molecular graph and is therefore robust, i.e., unambiguously defined by a single chemical formula. The RPCG values are calculated in two steps, as they are derived from the precalculated partial charges (charge excesses, δ, in eq 3) of the atoms of the molecules. It is known that there are several approaches for the assignment of partial charges to the atoms in a molecule. In the case of QSARs A and B, the RPCGs were calculated by using the electronegativity-based charge distribution of the molecules ($RPCG_{EN}$). However, it may be worthwhile to check whether RPCG remains descriptive on the basis of a different partial charge system. For this reason, QM-based RPCG values ($RPCG_{QM}$) were calculated and put in the QSARs instead of $RPCG_{EN}$'s as second descriptors. From among the numerous ways to calculate QM-based partial charges according to different principles (e.g., Mulliken,[38] Hirshfeld[39] charges, etc.), the Breneman and Wiberg approach[40] was selected for the present calculations. This approach reconstitutes the electrostatic potential of a molecule by atomic charges, which is appropriate for this study. It was found that the statistical parameters of the new correlation A_{QM} ($R^2 = 0.770$; $R^2_{cv} = 0.739$; $s^2 = 1.21$; details of the model are listed in the Supporting Information, Table D) are similar to those of A, with a slight decrease in the R^2 values and that J does not improve the model so effectively in this case (B_{QM}). However, the application of a completely different QM-based partial charge system on the ligand molecules, i.e., a 3D descriptor ($RPCG_{QM}$) instead of the 2D $RPCG_{EN}$, does not spoil the descriptive power of RPCG, which

(37) (a) Takamatsu, Y.; Itai, A. *Proteins* **1998**, *33*, 62−73. (b) Huo, S.; Wang, J.; Cieplak, P.; Kollman, P. A.; Kuntz, I. D. *J. Med. Chem.* **2002**, *45*, 1412−1419. (c) Vedani, A.; Dobler, M. *J. Med. Chem.* **2002**, *45*, 2139−2149. (d) Ma, X. H.; Wang, C. X.; Li, C. H.; Chen, W. Z. *Protein Eng.* **2002**, *15*, 677−681. (e) Hong, X.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 324−336.

(38) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833−1840, 1841−1846.
(39) Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129−138.
(40) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361−373.

can therefore be regarded as a robust quantity for the second descriptor in the present QSARs.

Conclusions

The results of the present study indicate that the use of different ΔG_b calculators for ligands of radically different sizes may be considered in future applications and development of docking/scoring methods. A semiempirical SF of a widely used docking method was modified and extended to achieve a precise fit of the structure-based, calculated binding free energy values to the experimental ΔG_b's for bulky, flexible peptidic ligands. The combination of the bimolecular descriptor ΔG_{TH} with additional ligand-based 2D descriptors yielded new, hybrid ΔG_b calculators with good predictive power. The results highlight the possibility of development of such hybrid calculators involving other SF-s in the future. Thorough tests and cross-validations of the QSARs were performed to verify the statistical relevance of the calculators and the descriptors. It was found, that the inclusion of bimolecular terms of the SF is obligatory for a diverse set of protein−ligand systems (ΔG_{TH} is the major descriptor in the QSARs). Both the scoring and the calculation of ligand-based 2D descriptors are rapid processes, even for the large ligands in this study. The precision of their present combination is at least comparable with that of other available calculators of binding thermodynamics. Thus, the proposed strategy is a real alternative for calculation of the binding affinities in the problematic cases of bulky, flexible lead compounds in the early phases of rational drug design. In practice, the docked lead compound−protein complexes can be supplied by AutoDock or other, appropriate automated docking methods and used with the hybrid calculators of the present study to obtain ΔG_b values.

Acknowledgment. This article is dedicated to Docent Dr. István Horváth, Department of Inorganic and Analytical Chemistry, University of Szeged in acknowledgment of his lectures in regression analysis. C.H. is a Békésy Fellow of the Hungarian Ministry of Education. U.M. is grateful to the Estonian Science Foundation for support (Grant No. 5805). This work was supported by OTKA TS 049817 and the Hungarian National Bureau of Research and Development.

Supporting Information Available: Details of the protein−ligand complexes (Table A). The Codessa descriptor pool used for selection of the appropriate 2D descriptors (Table B). Numerical values of the descriptors (Table C). The correlation of experimental and calculated binding free energy values of the 50 complexes. Multilinear regressions (eq 2, $n = 2$ or 3) were performed, using ΔG_{TH}, $RPCG_{QM}$, and ligand-based 2D descriptor J (Table D). Complete ref 28. This material is available free of charge via the Internet at http://pubs.acs.org.

JA055804Z