Bioinformatics Advance Access published September 21, 2007

Structural Bioinformatics

Structure-based calculation of drug efficiency indices

Csaba Hetényi[1,2,*], Uko Maran[1], Alfonso T. García-Sosa[1] & Mati Karelson[1]

[1]Institute of Chemical Physics, University of Tartu, 2 Jakobi Street, 51014 Tartu, Estonia, [2]Department of Organic Chemistry, Faculty of Pharmacy, Semmelweis University, Budapest, Hungary.

Associate Editor: Prof. Anna Tramontano

[*]ABSTRACT

Motivation: The efficiency indices (EI's) have been derived from the experimental binding affinities of drug candidates to macromolecules. These "two-in-one" measures include information on both pharmacodynamics and pharmacokinetics of the candidate molecules. The time-consuming experimental measurement of binding affinities of extensive molecule libraries may become a bottle-neck of large scale generation and application of EI's.

Results: To overcome this limitation, structure-based calculation of new EI's is introduced using the modified free energy function of the popular program package AutoDock. The results are validated on experimental binding data of biochemical systems such as potent inhibitors bound to -secretase, a key enzyme of Alzheimer's disease and various drug-protein complexes. Application of new EI's is tested. Thermodynamics of EI's and their role in virtual high throughput screening of drugs and in the development of docking programs are discussed.

Contact: csabahete@yahoo.com

Supplementary Information: accompanies this manuscript on the publisher's web site.

1 INTRODUCTION

The mechanism of drug action generally involves a long chain of interactions with the molecules of the human body. There are numerous experimental and *in silico* drug design tools describing the terminal link of these chains, i.e. the estimation of equilibrium binding affinities (BA) of drug candidates (ligands) to the targeted macromolecules. Although BA is undoubtedly a key property, other pharmacokinetic and non-equilibrium links in the chain such as absorption, distribution, and excretion of the candidate molecules also affect drug-likeness (Swinney 2004, Swinney 2006).

Accordingly, most of the current *in silico* molecular design strategies (Lipinski and Hopkins 2004) include modeling steps for the equilibrium binding and also for the pharmacokinetics of drugs. Atomic level techniques have been introduced for structural calculation of binding in ligand-target complexes. Computational molecular docking (Fig. 1) is the most advanced among these techniques (Brooijmans and Kuntz 2003). The BA values of the ligands can be calculated directly from docked ligand-protein complex structures with free energy (scoring) functions. Another important step is the optimization of pharmacokinetics and drug-likeness of ligand databases using empirical rules of selection (Lipinski et al. 1997). These rules define limit values of simple, size-dependent molecular descriptors, e.g. the molecular weight (M_W) which can be used for filtering of compound databases.

[*]To whom correspondence should be addressed.

Recently, new measures, the efficiency indices (EI) were introduced (Abad-Zapatero and Metz 2005, Hopkins et al. 2004) linking the above mentioned different steps of drug design. EI's have promptly gained applications connecting structural diversity and biological activity of drugs (Schuffenhauer et al. 2006) and in optimization of synthetic receptors (Chen et al. 2006). The introduction of EI's was inspired by earlier studies (Kuntz et al. 1999) showing the usefulness of normalization of BA with the number of heavy atoms (N_{HAT}) for drug design purposes.

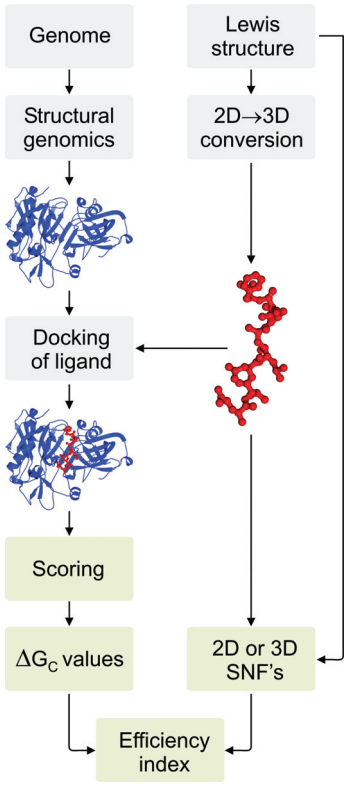

Fig. 1. The pathway of *in silico* drug design connecting genome and drug efficiency. Structural genomics projects generate new protein structures at an unprecedented rate (Yang and Tung 2006). To efficiently use this increasing amount of 3D information for drug design, high throughput methods are necessary, which can reduce the complexity of drug (ligand)-protein interactions to comparable measures (indices). The sequence of grey boxes show, that starting from the 2D Lewis structures of a ligand, 3D ligand-protein complexes can be obtained via conversion, modeling and docking. In the present study (beige boxes), a set of biologically relevant ligand-protein complexes were used for calculation of binding free energy

(ΔG_C). A representative complex of β-secretase (blue), a key enzyme of Alzheimer's disease and its potent peptidic inhibitor ligand, GluValAsnLeu(Ψ)AlaAlaGluPhe (red) is included in this figure. Further references on the role of β-secretase can be found in works of Hetényi et al. (2006) and Hong et al. (2000). Both 2D and 3D representation of the ligand molecules can be used for calculation of size-dependent normalization factors (SNF). The ratio of ΔG_C and SNF is the efficiency index, which is a practical 'two-in-one' measure of drug design. This figure was prepared using PyMol (DeLano 2006).

In EI's, the normalized quantities (Eq. 1) are represented by commonly used measures of BA such as the experimental free energy of binding (ΔG_E), the negative logarithms of experimental dissociation constant (pK_d), inhibition constant (pK_i) or inhibitor concentration at 50 % inhibition (pIC_{50}). The above mentioned simple descriptors, i.e. M_W (Abad-Zapatero and Metz 2005) or N_{HAT} (Hopkins et al. 2004) are typical examples of the size-dependent normalizing factors (SNF).

$$EI = \frac{BA}{SNF} \tag{1}$$

The EI's were originally defined with experimental BA values (Abad-Zapatero and Metz 2005, Hopkins et al. 2004). However, the use of structure-based, calculated binding free energy (ΔG_C) values from scoring (Hetényi et al. 2006) of docked ligand-protein structures instead of ΔG_E may become successful alternative for obtaining EI's. Remarkably, computational docking has an advantage of producing atomic level protein-ligand complex structures within reasonable time. The calculation of ΔG_C (scoring) is either performed along with the docking calculations or independently in post-docking mode (Fig. 1). In both cases it requires negligible time and, therefore, allows reduction of time-consuming and expensive biochemical measurements of BA's. Picking up the speed of *in silico* docking and scoring, the calculation of EI's can become an essential part of high throughput, structure-based virtual compound screening and drug design. The aim of the present study is to introduce and investigate rapid calculation of various EI's on the basis of a set of biologically relevant structural and thermodynamic experimental data.

2 METHODS

Binding data and structure-based free energy calculation of protein-ligand systems. ΔG_E and ΔG_C values of 53 protein-ligand complexes were adopted from a previous study (Hetényi et al. 2006) and listed in Supplementary Information. Proteins having large, peptidic ligands (M_W>350) and physiological importance such as the β-secretase enzyme of Alzheimer's disease (Fig. 1), HIV-1 protease, streptavidin, and immunoglobulins were prioritized for the study. The atomic coordinates of 41 of the complexes, were obtained from the Protein Databank (PDB, Berman et al. 2000). 12 β-secretase- inhibitor systems (om12, om13, om14, om15, om16, om17, om18, om19, om22, om23, om24 and om99-1) with no PDB structures available were modeled by modification of the 1fkn structure. Details on the systems, modeling and minimization of the complexes can be found in the previous paper (Hetényi et al. 2006). Although the peptidic ligands of these systems may become excellent lead compounds, they cannot be considered as drugs (Rishton 2003). Thus, a set of an additional 20 drug-protein complexes (Table 1) having both PDB structures and ΔG_E values was collected and used in the external validation and application tests of the new EI's introduced in this study. The sources and the procedure of collection of these data are described in details in the Supplementary Information. Altogether the 53+20 ligands represent a wide range of compounds including larger, lead-like nondrugs and actual drugs.

The ΔG_C's were calculated using the minimized protein-ligand complexes, according to the modified AutoDock 3.0 (AD3, Morris et al. 1998) and AutoDock4 (AD4, Huey et al. 2007) scoring functions (Eq. 2).

Table 1. The 20 drug-protein complexes of the external validation set

PDB code	Protein	Drug
1aj6	gyrase	novobiocin
1cea	plasminogen	aminocaproic acid
1dhi	dihydrofolate reductase	methotrexate
1dwc	alpha-thrombin (small subunit)	argatroban
1f5l	urokinase-type plasminogen activator	amiloride
1fkf	FK 506 binding protein	tacrolimus
1h61	pentaerythitol tetranitrate reductase	hydrocortisone
1hvy	thymidylate synthase	raltitrexed
1hxw	HIV protease	ritonavir
1j3j	dihydrofolate reductase	pyrimethamine
1jt1	FEZ-1, class B3 metallo-beta-lactamase	captopril
1m2z	glucocorticoid receptor	dexamethasone
1odi	purine nucleoside phosphorylase	adenosine
1ohr	aspartylprotease	nelfinavir
1p62	deoxycytidine kinase	gemcitabine
1sqn	progesterone receptor	norethindrone
1t7j	drug resistant HIV protease	amprenavir
1uw6	acetylcholine-binding protein	nicotine
2aou	histamine N-methyltransferase	amodiaquine
2gss	glutathione S-transferase P1-1	ethacrynic acid

$$\Delta G_C(AD3) = \underbrace{f_{elec}\sum_{i,j}\frac{q_j q_i}{\varepsilon(r_{ij})r_{ij}} + f_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + f_{hbond}\sum_{i,j}\xi(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right)}_{\Delta H_C} + \underbrace{f_{sol}\sum_{i,j} S_i V_j e^{(r_{ij}^2/2\sigma^2)}}_{\Delta G_{s,C}} \tag{2}$$

The f coefficients were determined empirically from a multi-linear regression (MLR) to a set of 30 protein-ligand complexes (AutoDock calibration set) with known binding constants (Morris et al. 1998). 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 function (ε) (Morris et al. 1996). 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. ξ(t) is a directional weight depending on angle t at the H-bonds. S and V denote the solvation parameter and fragmental volume, respectively, in the

solvation function of Stouten et al. (1993). In the scoring function 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 of σ=3.5 Å. For simplicity, the sum of Coulombic and Lennard-Jones (enthalpic) terms is marked as ΔH_C and the last, desolvation term is marked as $\Delta G_{s,C}$. Remarkably, the AutoDock4 scoring function has different parametrization of the ΔG_C(AD4) part, especially for the desolvation term. Details on the new AD4 scoring function can be found in the original paper of Huey et al. (2007). In the present study, all systems were re-scored using the epdb command of AutoDock4. Besides the ΔG_C(AD4), i.e. the intermolecular enthalpic+desolvation terms, the full AD4 binding free energy (ΔG_{full}(AD4)) was also calculated and checked for applicability in EI calculations.

Table 2. Codes and definitions of size-dependent normalizing factors (SNF) of ligands used in the denominator of efficiency indices (Eq. 1)

Code	Name of SNF	Definition	References
1D SNF's			
N_{AT}	Number of atoms		Karelson 2000
N_{HAT}	Number of heavy atoms		Hopkins et al. 2004
N_B	Number of σ-bonds		Karelson 2000
M_W	Molecular weight		Abad-Zapatero & Metz 2005
2D SNF's			
W	Wiener index	$W = \frac{1}{2}\sum_{i,j}^{N_{SA}} d_{ij}$	Wiener 1947
${}^0\chi$	Randic index (n=0)		
${}^1\chi$	Randic index (n=1)		
${}^2\chi$	Randic index (n=2)	${}^n\chi = \sum_{n-\text{length paths}}^{N_{SB}} (\delta_{i1}\ldots\delta_{in+1})^{-1/2}$	Randić 1975
${}^3\chi$	Randic index (n=3)		
${}^0\chi^v$	Kier&Hall index (n=0)		
${}^1\chi^v$	Kier&Hall index (n=1)		
${}^2\chi^v$	Kier&Hall index (n=2)	${}^n\chi^v = \sum_{n-\text{length paths}}^{N_{SB}} (\xi_{i1}\ldots\xi_{in+1})^{-1/2}$; $\xi_i = \frac{Z_i^v - H_i}{Z_i - Z_i^v - 1}$	Kier & Hall 1976
${}^3\chi^v$	Kier&Hall index (n=3)		
${}^1\kappa$	Kier shape index (n=1)	${}^n\kappa = (N_{SA}+\alpha)(N_{SA}+\alpha-1)^2({}^nP+\alpha)^2$	
${}^2\kappa$	Kier shape index (n=2)	${}^n\kappa = (N_{SA}+\alpha-1)(N_{SA}+\alpha-2)^2({}^nP+\alpha)^2$	
${}^3\kappa$	Kier shape index (n=3)	${}^n\kappa = (N_{SA}+\alpha-1)(N_{SA}+\alpha-3)^2({}^nP+\alpha)^2$, if N_{SA} is odd ${}^n\kappa = (N_{SA}+\alpha-3)(N_{SA}+\alpha-2)^2({}^nP+\alpha)^2$, if N_{SA} is even	Kier 1990
Φ	Kier flexibility index	$\Phi = ({}^1\kappa^2\kappa)/N_{SA}$	Kier 1990
J	Balaban index	$J = \left(\frac{q}{\mu+1}\right)\sum_{i,j}^{q}(s_i s_j)^{-1/2}$; $\mu = q-n+1$	Balaban 1982
3D SNF's			
$I_AI_BI_C$	Product of principal moments of inertia	$I_AI_BI_C = \prod_x I_x$; $I_x = \sum_i A_{M,i}r_{x,i}^2$; (x = A, B or C)	Karelson 2000
GI_B	Gravitation index (all bonds)	$GI_B = \sum_{i<j}^{N_B} \frac{A_{M,i}A_{M,j}}{r_{ij}^2}$	Karelson 2000
GI_P	Gravitation index (all pairs)	$GI_P = \sum_{i<j}^{N_A} \frac{A_{M,i}A_{M,j}}{r_{ij}^2}$	Karelson 2000

N_{SA}: number of atoms in the molecular graph (hydrogens excluded); **N_{SB}**: number of bonds in the graph; **n**: length of bonding path (topological distance, order of descriptor); **d_{ij}**: entry of the distance matrix corresponding to the number of bonds in the shortest path connecting the pair of atoms i and j; **δ**: coordination number of atoms; **v**: valence of atom in a molecule; **ξ**: value of atomic connectivity; **Z_i**: total number of electrons in atom i; **Z_i^v**: number of valence electrons in atom i; **H_i**: number of hydrogens directly attached to atom i; **^nP**: number of paths of length n in the molecular graph; **α**: sum of all ratios of the i^{th} atomic radius and radius of sp^3 carbon atom for all atoms in the graph minus 1; **q**: number of edges in the molecular graph; **m**: number of vertices in the graph; **μ**: cyclometric number; **s_i** and **s_j**: distance sums obtained by summation of row i and column i (or row j and column j) of the distance matrix; **A_M**: atomic mass; **$r_{x,i}$**: distance of the i^{th} atom from principal axis x; **A, B, C**: principal axes; **N_A**: number of atoms; **N_B**: number of bonds; **r_{ij}**: interatomic distance.

Regression analyses. The LR's were statistically analyzed and the SNF values were obtained using the program package CODESSA (ver. 2.0) (Katritzky et al. 1995, Karelson et al. 1996). Results of the regression analyses, i.e. mean square errors and t-values of the regression coefficients, the F-values, and the squares of the correlation coefficients (r^2) of the regressions are tabulated in section

Results. The principal moments of inertia were calculated for the binding conformations of ligand molecules using the Analyze program of the TINKER software package (Ren and Ponder 2003). Numerical values used in the calculations and the correlations of SNF's are tabulated in Supplementary Information.

3 RESULTS

Definition of new EI's. The list and definitions of the SNF's (Eq. 1) corresponding to new, and formerly published (Abad-Zapatero and Metz 2005, Hopkins et al. 2004) EI's can be found in Table 2. Some of these SNF's are commonly applied as two-dimensional (2D) descriptors in quantitative structure-activity relationship (QSAR) equations (Devillers and Balaban 1999) and show relatively large degree of correlation with each other (Supplementary Information). The more complicated SNF's contain information also on molecular complexity involving internal (topological) distances and branching of the ligands resulting in their more unique profile and, in some cases, moderate correlations with each other. Whereas 2D descriptors are derived solely from the Lewis formula, i.e. the empirical connectivity list or molecular graph of the ligands, calculation of ΔG_C requires the knowledge of spatial atomic positions in the protein-ligand complex. In a recent study (Hetényi et al. 2006), it was found that ΔG_E's of even large, flexible peptides (Fig. 1) can be predicted with a modified scoring function (ΔG_C) of the docking program package AutoDock 3.0 (Morris et al. 1998). As ΔG_C shows a significant correlation with the ΔG_E values (Hetényi et al. 2006) it was selected to represent BA in the structure-based, calculated EI values throughout the present investigations (BA = ΔG_C in Eq. 1).

Correlation of experimental and calculated EI's. To test the reliability and predictive value of calculated EI's, simple linear regression (LR) analyses were performed with EI's obtained from the measured ΔG_E values (Eq. 3).

$$\frac{\Delta G_{E,k}}{SNF_k} = \alpha \frac{\Delta G_{C,k}}{SNF_k} + \beta + \varepsilon_k \qquad (k = 1, 2, \ldots, N) \tag{3}$$

Where α, and β represent the regression coefficient and the intercept, respectively. The ε_k's are the residuals at each data point. The total number of data points (N), i.e. the number of protein-ligand systems adopted from the previous study (Hetényi et al. 2006) was 50. A systematic series of LR's were developed for EI's based on the SNF's of Table 2 and ΔG_C's calculated with the scoring schemes of AutoDock3.0 and AutoDock4, respectively. The results and statistical parameters of the LR's are summarized in Table 3 and in the Supplementary Information.

Table 3. Statistical parameters of linear regressions (Eq. 3) obtained for efficiency indices based on SNF's of different dimensionality

SNF	r^2	$r^2_{cv,LOO}$	$r^2_{cv,L50\%O}$	$r^2_{cv,EXT}$	F-value	r^2	$r^2_{cv,LOO}$	$r^2_{cv,L50\%O}$	$r^2_{cv,EXT}$	F-value
	AutoDock3.0					AutoDock4				
1D SNF's										
N_{AT}	0.857	0.845	0.852	0.493	286.90	0.839	0.826	0.835	0.718	250.28
N_{HAT}	0.896	0.886	0.891	0.593	413.66	0.887	0.877	0.884	0.758	378.11
N_B	0.865	0.854	0.863	0.522	308.23	0.848	0.835	0.844	0.731	266.98
M_W	0.889	0.879	0.889	0.607	386.55	0.881	0.870	0.879	0.767	354.22
2D SNF's										
W	**0.962**	**0.954**	**0.954**	**0.910**	**1216.46**	**0.960**	**0.953**	**0.952**	**0.931**	**1139.22**
$^0\chi$	0.891	0.882	0.890	0.589	394.40	0.884	0.873	0.879	0.757	364.67
$^1\chi$	0.893	0.884	0.889	0.582	402.59	0.884	0.873	0.883	0.752	364.56
$^2\chi$	0.918	0.910	0.911	0.686	540.18	0.913	0.905	0.908	0.803	503.64
$^3\chi$	0.916	0.908	0.912	0.856	524.80	0.906	0.897	0.904	0.909	461.11
$^0\chi^v$	0.892	0.882	0.830	0.548	397.04	0.881	0.870	0.878	0.749	355.47
$^1\chi^v$	0.886	0.876	0.874	0.571	372.51	0.871	0.860	0.856	0.764	325.15
$^2\chi^v$	0.914	0.906	0.910	0.692	509.66	0.904	0.895	0.903	0.832	449.73
$^3\chi^v$	0.905	0.897	0.896	0.803	458.55	0.889	0.880	0.884	0.896	384.51
$^1\kappa$	0.870	0.859	0.866	0.602	322.41	0.862	0.850	0.860	0.782	300.81
$^2\kappa$	0.791	0.774	0.788	0.603	181.45	0.777	0.759	0.776	0.801	167.30
$^3\kappa$	0.781	0.764	0.726	0.719	170.81	0.784	0.768	0.729	0.859	174.61
Φ	0.742	0.723	0.739	0.667	137.79	0.729	0.709	0.727	0.845	129.26
J	0.871	0.855	0.860	0.847	325.16	0.854	0.834	0.845	0.889	280.01
3D SNF's										
$I_AI_BI_C$	**0.966**	**0.929**	**0.938**	**0.961**	**1345.18**	**0.963**	**0.933**	**0.938**	**0.963**	**1246.91**
GI_B	0.900	0.891	0.890	0.660	432.20	0.892	0.882	0.887	0.787	396.71
GI_P	0.927	0.919	0.926	0.796	606.68	0.921	0.914	0.917	0.863	563.02

r^2: the squared correlation coefficient; **r^2_{cv}**: the square of the cross-validated correlation coefficient (LOO: leave-one-out method, L50%O: leave-50%-out method, EXT: external validation set of 20 drug-protein systems)

All LR's are statistically significant, and show higher r^2 values than the correlation (r^2=0.706) obtained between ΔG_E and ΔG_C (Heténi et al. 2006). Importantly, the high r^2 values in Table 3 are not trivial consequences of this correlation in the previous work, as the SNF values are different for the 50 different ligand molecules (Eq. 3).

An advantage of 2D descriptors such as the Wiener index (W) involved in the best correlation (Fig. 2, Table 3) is that they can be unambiguously and rapidly calculated from the internal connectivity information coded in the molecular graph (Table 2). For example, W involves a simple summation of shortest topological distances in a molecule. Comparably good correlations could be achieved at all other SNF's including Balaban index (J) which is also defined by internal topological distances (Table 2) and was found to be useful as a QSAR descriptor in prediction of the entropic parts of ΔG_E (Heténi et al. 2006). In addition, even the three outlier protein-ligand systems (1hhj, om22, om24) of the previous study (Heténi et al. 2006) could be involved in the models (N=53 in Eq. 3). In case of W the level of correlation (r^2=0.962) did not decrease when the three former outliers were included.

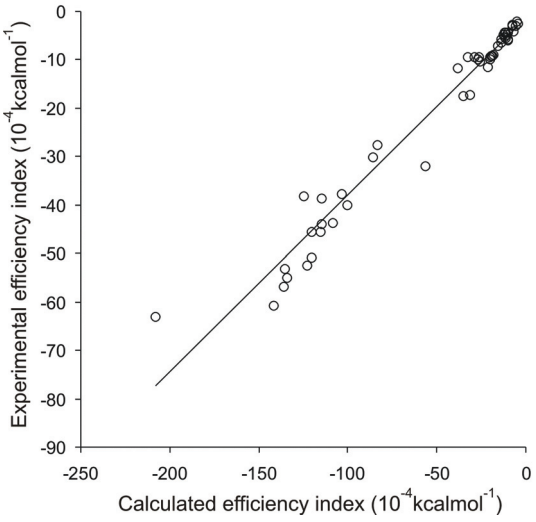

Fig. 2 The correlation of experimental and calculated efficiency indices (EI) using the Wiener index as a size-dependent normalizing factor (AD3 scoring).

Cross-validation of the correlations. There were different methods applied for cross-validation of the correlations presented in Table 3. The cross-validated correlation coefficients (r^2_{cv}) of the leave-one-out (LOO) and leave-50%-out (L50%O) methods (Table 3) shows that exclusion of one or more data points from the models does not decrease the level of correlation dramatically. A set of 20 drug-protein complexes (Table 1) was used as an external validation set (EXT). Most of the corresponding r^2 values are above 0.5 showing that the models can predict the EI values for smaller, drug ligands not included in the training set (50 systems). Notably, ΔG_C(AD4) produced higher r^2_{cv}(EXT) values for the external validation than ΔG_C(AD3), probably due to the more advanced solvation terms and the larger compound database included in its parametrization. The ΔG_{full}(AD4) function did not result better EI-correlations (data not shown) than ΔG_C(AD4), and, therefore ΔG_C(AD4) was selected for the final evaluations (Table 3).

The results of the cross-validated correlations in Table 3 let us conclude that structure-based calculation of EI's works for both the 'traditional' (Abad-Zapatero and Metz 2005, Hopkins et al. 2004) and the newly introduced 2D SNF's (W, χ's, J, etc.). The formulas in Eq. 3 and Table 2 and the validated models can be coded and applied as EI-calculators during the *in silico* drug design process (Fig. 1). Direct implementation of EI-calculator algorithms in docking/scoring program packages such as AutoDock is also possible.

Applications. (1) To check the applicability of two new EI's with the best correlations (Table 3) the distributions of ΔG_E and EI values were compared for the sets including the 50 peptidic compounds (nondrugs) and the 20 drugs, respectively. It was found (Fig. 3 and Supplementary Information) that overlapping distributions of ΔG_E's (Fig. 3A) of drugs and nondrugs are separated for the EI's (Fig. 3B). There is one or two order of magnitude difference (Table 4) in the median/average values of EI's for both W and $I_A I_B I_C$ and there are considerably large gaps between the minimum values of drugs and nondrugs, as well. These results emphasize the applicability of the new EI's in separation of drugs from nondrugs.

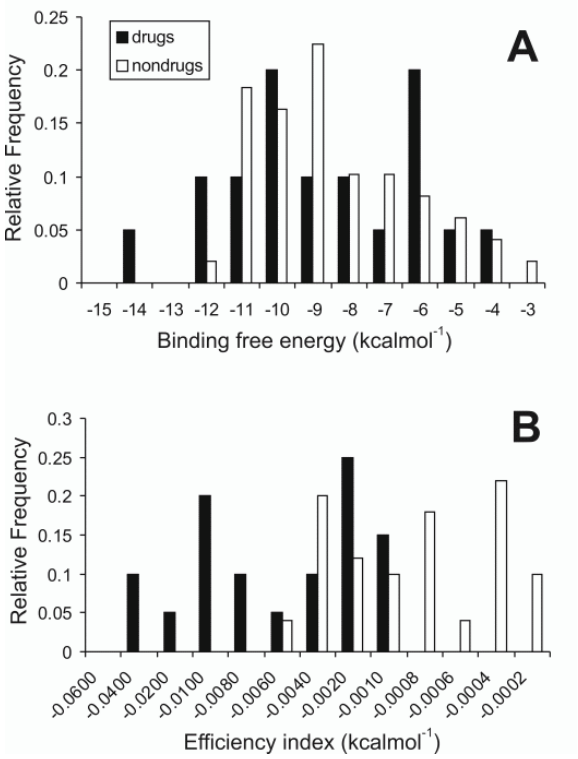

Fig. 3 Histograms showing the distribution of experimental binding free energy (A) and Wiener index-based efficiency index (B) values for drugs and nondrugs. (The scales cover the full range of values and the same number of bins were applied for both histograms.)

(2) The introduction of EI's in a virtual screening process improves the selectivity of screening. As a test case, the binding pocket of progesterone receptor was used as a target in the docking of 1760 compounds including an abridged version of the NCI Diversity Set (NCI/NIH, Lindstrom et al. 2003) and the native drug ligand norethindrone (1sqn, Table 1). ΔG_C's were collected and W- and $I_A I_B I_C$-based EI's were calculated. Details on the methods of these

procedures are described in the Supplementary Information. It was found, that the use of ΔG_C's alone ranked norethindrone to the best 10 % of the 1760 compounds. Re-ranking of the best 10 % according to W- and $I_AI_BI_C$-based EI's resulted norethindrone in the 2nd and 6th best position (< top 0.5 %) on the list of the 1760 compounds, respectively. This test showed that in a second ranking step these new EI's can improve the quality of selection of a real drug.

Table 4. Statistics of the distribution of experimental binding free energy values and efficiency indices based on Wiener index (EI_W) and $I_AI_BI_C$ (EI_{IAIBIC}) for drugs and nondrugs. ΔG_E and EI_W values are in kcalmol^{-1} units. EI_{IAIBIC} has a dimension of kcalmol^{-1}amu^{-3}Å^{-6}.

		Median	Average	Minimum	Maximum
ΔG_E	Drugs	-9.87	-9.30	-14.75	-4.63
	Nondrugs	-9.50	-8.86	-12.94	-3.89
EI_W	Drugs	-6.281×10^{-3}	-1.226×10^{-2}	-5.930×10^{-2}	-1.014×10^{-3}
	Nondrugs	-9.728×10^{-4}	-2.101×10^{-3}	-2.137×10^{-4}	-2.137×10^{-4}
EI_{IAIBIC}	Drugs	-3.151×10^{-10}	-6.564×10^{-9}	-6.783×10^{-8}	-4.644×10^{-12}
	Nondrugs	-4.190×10^{-12}	-3.579×10^{-11}	-3.283×10^{-10}	-1.677×10^{-13}

4 DISCUSSION

The background of the thermodynamics of EI's. The binding free energy (ΔG) can be written as the sum of experimental enthalpic (ΔH) and entropic (ΔS) binding contributions (Eq. 4), where T is the thermodynamic temperature.

$$\Delta G = \Delta H - T\Delta S \quad (4)$$

As an additive quantity, ΔS can be further split into translational (ΔS_t), rotational (ΔS_r), and vibrational (ΔS_v) entropy changes (Eq. 5) at the ligand molecule. In some articles (Noskov 2001), further contributions are also considered such as solvation/desolvation free energy (ΔG_s) of the ligand and/or the protein molecules, etc. As the SNF's depend solely on the ligands, involvement of protein effects is not necessary in the forthcoming discussion.

$$\Delta G = \Delta H - T(\Delta S_t + \Delta S_r + \Delta S_v) + \Delta G_s \quad (5)$$

The use of statistical thermodynamics expressions (Carlsson and Åqvist 2005, Murray and Verdonk 2002) for estimation of components S_t, S_r (Table 5) and S_v of molecular entropy is quite common. S_v depends on the frequencies of normal modes of the ligand molecule, which cannot be connected with the simple SNF's of this study. ΔG_s includes both enthalpic and entropic contributions (Zou et al. 1999) and partly depends on the molecular size and shape of the ligand via the solvent accessible surface area. Accordingly (Eq. 5), the division of ΔG_E (left side of Eq. 3) with SNF's results in normalized ΔH_E's and ΔS_E's.

On the right side of Eq. 3 there is ΔG_C (Eq. 6), including three terms (Eq. 2, Methods), which can be assigned (Brooijmans and Kuntz 2003, Calderone and Williams 2001) to the enthalpic (ΔH_C) contributions of binding. The fourth term of ΔG_C ($\Delta G_{s,C}$) is an estimate of ΔG_s which represents only a minor portion of ΔG_C (Eq. 6).

$$\Delta G_C = \Delta H_C + \Delta G_{s,C}. \quad (6)$$

Thus, the SNF-normalized ΔG_C (Eqs. 3 and 6) contains mostly normalized ΔH_C (and neglibible $\Delta G_{s,C}$). Most importantly, there are no terms estimating ΔS_t, ΔS_r, and ΔS_v on the right side.

If assuming that experimental entropy (S_E), i.e. S_t and S_r becomes zero after ligand binding, then ΔS_t and ΔS_r will include size-dependent factors, such as M_W or the product of principal moments of inertia ($I_AI_BI_C$), respectively (Table 5). However, it was correctly discussed (Carlsson and Åqvist 2005), that the assumption of zero final entropy is rather hypothetical as the ligand does fluctuate around its binding position. Whereas the formulas of Table 5 can hardly be applied for calculation of binding entropy of ligands in their present forms, they obviously show the dependence of molecular entropy on M_W and $I_AI_BI_C$ of ligands. Thus, normalization of ΔS_E (left side of Eq.3) with SNF's such as M_W or $I_AI_BI_C$ can be expected to decrease the ligand-dependency of the ΔS_E terms resulting in a constant part of the normalized ΔS_E.

Table 5. Statistical thermodynamics formulas of molecular entropy

Molecular entropy	Formula
Translational (S_t) (Sackur-Tetrode)	$S_t = Nk\ln\left[\frac{Ve^{5/2}}{N}\left(\frac{2\pi kTM_W}{h^2}\right)^{3/2}\right]$
Rotational (S_r)	$S_r = Nk\ln\left[\frac{8\pi^2}{\sigma}\left(\frac{2\pi ekT}{h^2}\right)^{3/2}(I_AI_BI_C)^{1/2}\right]$

Note, that the Sackur-Tetrode equation used for discussion was originally derived for gas phase. **V**: volume available for the molecule; **N**: number of molecules; $\mathbf{M_W}$: molecular weight; **k**: Boltzmann's constant; **T**: thermodynamic temperature; **h**: Planck's constant; $\boldsymbol{\sigma}$: symmetry number; $\mathbf{I_AI_BI_C}$: product of principle moments of inertia (see also Table 1).

The constant part of SNF-normalized ΔS_E does not affect the level of correlation and the remaining SNF-normalized enthalpic terms in Eq. 3 correlate well with each-other (Table 3).

New 3D SNF's. To test the prediction of the previous section, i.e. the usefulness of $I_AI_BI_C$ as a 3D SNF, it was employed in Eq. 3. The statistical parameters of the corresponding LR (Table 3, Supplementary Information) show an excellent correlation (r^2=0.966) verifying the expectation. Remarkably, both the 3D $I_AI_BI_C$ and the 2D W involve the calculation of real or topological internal lengths of the ligand molecules, and, therefore their connection is trivial. Their correlation for the 50 ligands is r^2=0.864. Interestingly, the 2D W performed as well (Table 3) as the obviously more elaborate 3D $I_AI_BI_C$ in case of the 50 systems. It was also found, that $I_AI_BI_C$ works even for smaller sub-sets of the 50 investigated systems resulting in e.g. an r^2 of 0.973 for the 10 modeled β-secretase complexes alone (AD3 scoring).

Other internal distance-based 3D SNF's such as the gravitation index (GI), a descriptor successful in prediction of boiling points (Katritzky et al. 1996) also provided good LR results in calculation of EI's (Table 3).

Methodological aspects of the results. Scoring functions of docking programs are generally based on correlations of ΔG_E with ΔG_C. However, during the development of scoring functions, separate fit of experimental ΔH_E and ΔS_E to the corresponding enthalpic and entropic terms (Brooijmans and Kuntz 2003) of the scoring functions would be an ideal way (Murphy 1999) to decrease errors coming from overlapping and/or coupled terms. However, most of the experimental thermodynamic BA data available are ΔG_E values or pK's from which ΔG_E's can be calculated (Wang et al. 2004). The amount of enthalpic data is limited as experimental binding enthalpy (ΔH_E) can be obtained only by additional measurements with special techniques, e.g. isothermal titration calorimetry (Cam-

poy and Freire 2005). The LR's of the previous sections showed, that the SNF-normalization of ΔG_E provides excellent correlation with the normalized ΔH_C without additional measurements of ΔH_E, due to the high enthalpic content of both sides of Eq. 3 (see previous sections for details).

It can also be recognized (Eq. 3), that the reciprocals of the SNF's are actual weights in the weighted least squares fit of the calculated enthalpic terms to the experimental ΔG_E's. By using these weights during development of scoring functions, the degree of correlation and the accuracy of computational docking-scoring methods can be increased.

Practical applications. The EI's are simple indicators developed to aid rational drug design and hit-to-lead approaches (Keserű and Makara 2006). In the present study, new EI's involving 2D and 3D SNF's were introduced. It was shown, that precise, structure-based calculation of EI's is a real alternative of time-consuming measurements and that the new EI's can be used in separation of drugs from nondrugs. The calculation of EI's of a large set of available drugs will allow the determination of reference EI-limits for selection of drug-like candidates in the future. The building of an EI database for the precise determination of EI-limits has already been started in our laboratory. As the proposed EI-calculators are fast and cost-effective, they will help to reduce the number of experimental measurements and can easily be combined with available methods in high throughput computational docking and scoring (Fig. 1).

ACKNOWLEDGEMENTS

The authors are thankful to Foundation Innove (www.innove.ee) project No. 1.0101-0310 for the financial support. The referees of the manuscript are acknowledged for their constructive suggestions.

REFERENCES

Abad-Zapatero, C. & Metz, J. T. (2005) Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today,* 10, 464-469.

Balaban, A.T. (1982) Higly discriminating distance based topological index. *Chem. Phys. Lett.,* 89, 399-404.

Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, 28, 235-242.

Brooijmans, N. & Kuntz, I. D. (2003) Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* 32, 335-373.

Calderone, C. T. & Williams, D. H. (2001) An enthalpic component in cooperativity: the relationship between enthalpy, entropy, and noncovalent structure in weak associations. *J. Am. Chem. Soc.,* 123, 6262-6267.

Campoy, A. V. & Freire, E. (2005) ITC in the post-genomic era...? Priceless. *Biophys. Chem.,* 115, 115-124.

Carlsson, J. & Åqvist, J. (2005) Absolute and relative entropies from computer simulation with applications to ligand binding. *J. Phys. Chem. B,* 109, 6448-6456.

Chen, W., Chang, C.-E. & Gilson, M. K. (2006) Concepts in receptor optimization: targeting the RGD peptide. *J. Am. Chem. Soc.,* 128, 4675-4684.

DeLano, W.L. (2006) *PyMol Molecular Graphics System*, DeLano Scientific, San Carlos CA, USA.

Devillers, J. & Balaban, A. T. eds. (1999) *Topological indices and related descriptors in QSAR and QSPR*. Gordon and Beach Science Publishers, Amsterdam.

Hetényi, C., Paragi, G., Maran, U., Timár, Z., Karelson, M., Penke, B. (2006) Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. *J. Am. Chem. Soc.,* 128, 1233-1239.

Hong, L., Koelsch, G., Lin, X., Wu, S., Terzyan, S., Ghosh, A.K., Zhang, X.C. & Tang, J. (2000) Structure of the protease domain of memapsin 2 (beta-secretase) complexed with inhibitor. *Science,* 290, 150-153.

Hopkins, A. L., Groom, C. R. & Alex, A. (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today,* 9, 430-431.

Huey, R., Morris, G.M., Olson, A.J. & Goodsell, D.S. (2007) A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.,* 28, 1145-52.

Karelson, M. (2000) *Molecular descriptors in QSAR/QSPR*, J. Wiley & Sons, New York.

Karelson, M., Lobanov, V. S. & Katritzky, A. R. (1996) Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.,* 96, 1027-1043.

Katritzky, A. R., Lobanov, V. S. & Karelson, M. (1995) QSPR:The Correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.,* 24, 279-287.

Katritzky, A.R., Mu, L., Lobanov, V.S. & Karelson, M. (1996) Correlation of Boiling Points With Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.,* 100, 10400-10407.

Keserű, G. M. & Makara, G. M. (2006) Hit discovery and hit-to-lead approaches. *Drug Discov. Today,* 11, 741-748.

Kier, L.B. & Hall, L.H. (1976) *Molecular Connectivity in Chemistry and Drug Reasearch*, Academic Press, New York.

Kier, L.B. (1990) in *Computational Chemical Graph Theory, ed.* Rouvray, D.H., Nova Science Publishers, New York, pp. 151-174.

Kuntz, I. D., Chen, K., Sharp, K. A. & Kollman, P. A. (1999) The maximal affinity of ligands. *Proc. Natl. Acad. Sci. USA,* 96, 9997-10002.

Lindstrom, W.H., Morris, G.M., Huey, R.H., Sanner, M.F. & Olson, A.J. (2003) The NCI Diversity Set for AutoDock. http://autodock.scripps.edu/resources/databases

Lipinski, C. & Hopkins, A. (2004) Navigating chemical space for biology and medicine. *Nature,* 432, 855-861.

Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.,* 23, 3-25.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. & Olson, A.J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.,* 19, 1639-1662.

Morris, G.M., Goodsell, D.S., Huey, R. and Olson, A.J. (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.,* 10, 293-304.

Murphy, K. P. (1999) Predicting binding energetics from structure: looking beyond DeltaG degrees. *Med. Res. Rev.,* 19, 333-339.

Murray, C. W. & Verdonk, M. L. (2002) The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput. Aided Mol. Des.,* 16, 741-753.

NCI/NIH, National Cancer Institute, Developmental Therapeutics Program. http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html

Noskov, S. Y. & Lim, C. (2001) Free energy decomposition of protein-protein interactions. *Biophys. J.,* 81, 737-750.

Randić , M. (1975) On characterization. of. molecular branching. *J. Am. Chem. Soc.,* 97, 6609-15.

Ren, P. & Ponder, J. W. (2003) Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B,* 107, 5933-5947.

Rishton, G.M. (2003) Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today*, 8, 86-96.

Schuffenhauer, A., Brown, N., Selzer, P., Ertl, P. & Jacoby, E. (2006) Relationships between Molecular Complexity, Biological Activity, and Structural Diversity. *J. Chem. Inf. Model.,* 46, 525-535.

Stouten, P.F.W., Frömmel, C, Nakamura, H. & Sander, C. (1993) An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. *Molec. Simulat.,* 10, 97-120.

Swinney, D. C. (2004) Biochemical mechanisms of drug action: what does it take for success? *Nat. Rev. Drug. Discov.,* 3, 801-808.

Swinney, D. C. (2006) Biochemical mechanisms of New Molecular Entities (NMEs) approved by United States FDA during 2001-2004: mechanisms leading to optimal efficacy and safety. *Curr. Top. Med. Chem.,* 6, 461-478.

Wang, R., Fang, X., Lu, Y. & Wang, S. (2004) The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.,* 47, 2977-2980.

Wiener, H. (1947) Structural determination of paraffin boiling points. *J. Am. Chem. Soc.,* 69, 17-20.

Yang, J.-M., Tung, C.-H. (2006) Protein structure database search and evolutionary classification. *Nucleic Acids Res.,* 34, 3646-3659.

Zou, X., Sun, Y. & Kuntz, I. D. (1999) Inclusion of solvation in ligand binding free energy calculations using the generalized-Born model. *J. Am. Chem. Soc.,* 121, 8033-8043.