Theor Chem Acc (2004) 112: 313–317
DOI 10.1007/s00214-004-0589-9

Theoretical
Chemistry Accounts

Regular article

Application of semiempirical quantum chemical methods as a scoring function in docking

V. Vasilyev, A. Bliznyuk

ANU Supercomputer Facility, The Australian National University, Canberra, ACT, Australia

Received: 15 July 2003 / Accepted: 18 December 2003 / Published online: 8 June 2004

Abstract. A crucial point in docking simulations is the scoring function used for estimation of the target-ligand interaction energy. The usual practice is to employ fast but simplified empirical scoring functions. Rigorous quantum chemical methods are too slow to screen virtual combinatorial libraries consisting of thousands of molecules, but they can be used in the final step of the simulations for assessing the results obtained. At this stage quantum chemical calculations can be performed only for the 10–100 top binders predicted by simplified scoring functions, and only using linear-scaling semiempirical quantum chemical methods such as MOZYME. The possibilities and potentialities of the quantum chemical methods for estimation of the binding affinities in docking simulations are a largely unexplored area, so the main goal of this study is a detailed evaluation of the potential and limitations of the MOZYME methodology for estimation of the target-ligand binding energies and its comparison with available experimental data.

Keywords: Semiempirical – MOZYME – Docking

Introduction

Predicting the binding affinity between a target and a ligand forming a noncovalent complex is a very challenging task that has been pursued for decades. Although plenty of computational approaches for estimation of the binding free energy of the target-ligand complexes have been developed [1, 2, 3, 4, 5, 6, 7, 8], no efficient and reliable method to evaluate the fitness between the target and the ligand in the general case is available [1, 2, 3, 4, 5, 6, 7, 8].

The development of an accurate method which will be able to predict real binding affinities and which will be, at the same time, computationally efficient enough to screen virtual combinatorial libraries consisting of many thousands of molecules is a very difficult task. One possible approach is to apply a two-stage ranking, i.e. to rapidly scan possible solutions using fast but simplified scoring functions to obtain initial "good" binders, followed by more sophisticated methods for assessing the results obtained. Up to now, the most "advanced" techniques for estimation of binding affinities, such as free-energy perturbation [9], the linear-response approximation [10], and a combination of the molecular mechanical energies with the continuum solvent approaches (Molecular Mechanics/Poisson–Boltzmann surface area) [11, 12], have been derived from empirical force field methods. As a result, they possess both the advantages and pitfalls of the force field approximation.

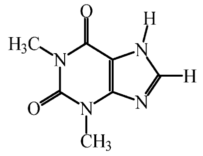
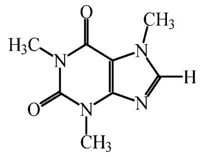
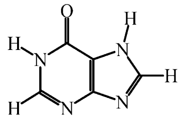
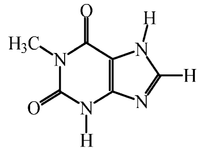
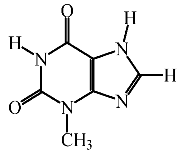
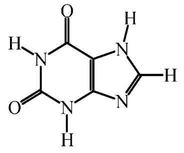
The increasing power of modern computers and the development of new computational methodologies have made it possible to perform quantum chemical calculations on several thousands of atoms [13, 14, 15], thus giving a more consistent and rigorous treatment of a molecular system. For example, quantum chemical methods describe both polarization and charge-transfer effects in a more realistic manner than by using point charges in force field approaches. Currently, quantum chemical treatment of several thousands of atoms with subsequent geometry optimization can be done using only so-called linear-scaling methods at the semiempirical level, for example, MOZYME [16].

The main goal of the present study is a detailed evaluation of the potential and limitations of the MOZYME methodology for estimation of the target-ligand binding energies and its comparison with available experimental data.

Proceedings of the 11th International Congress of Quantum Chemistry satellite meeting in honor of Jean-Louis Rivail

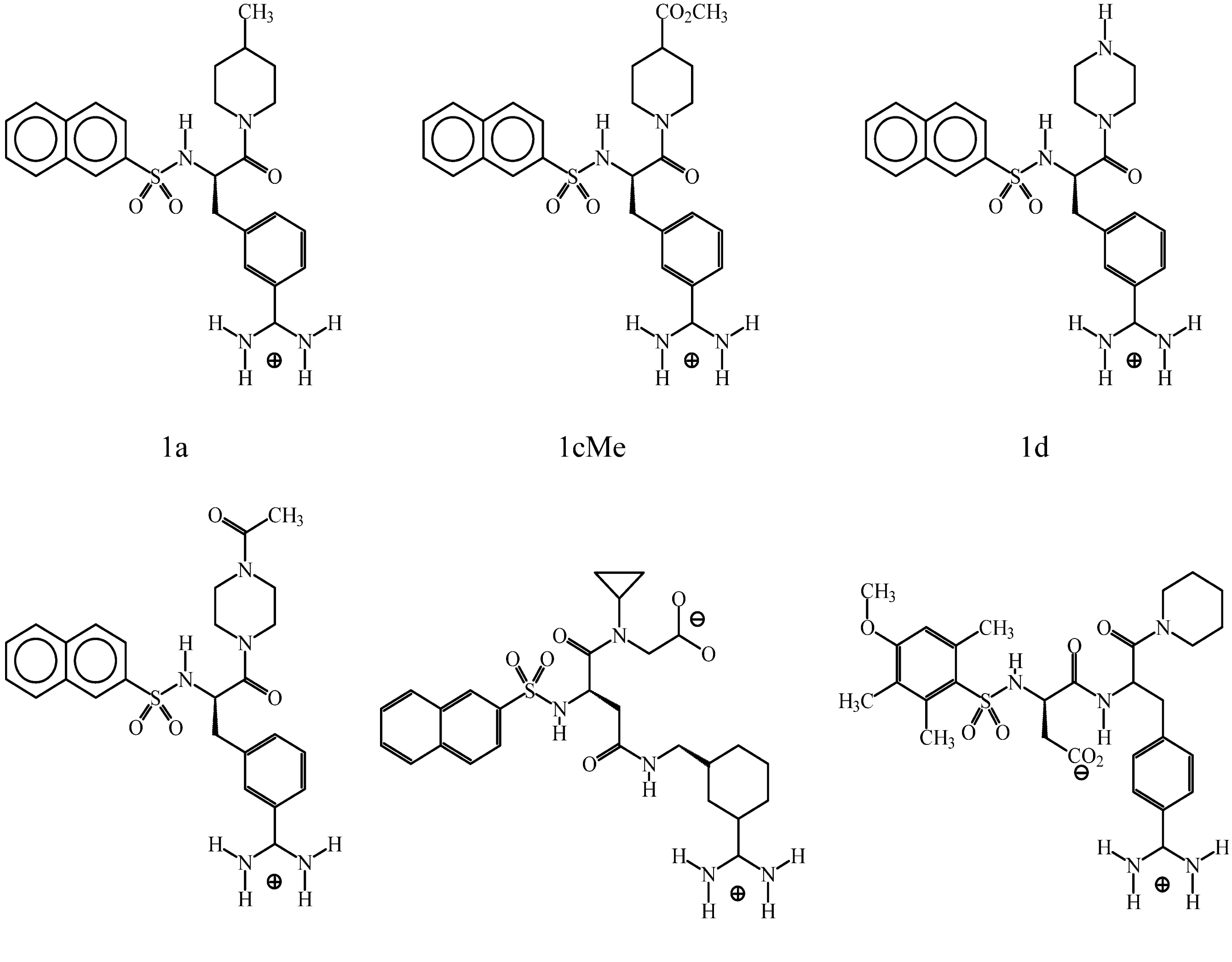
Correspondence to: V. Vasilyev
e-mail: vvv900@anusf.anu.edu.au

Table 1. Relative binding energies (in kilocalories per mole) for the RNA–theophylline analog complexes. Experimental binding free energies were calculated from dissociation constants and were taken from Ref. [18] [a]$\Delta G_{\mathrm{Exp}} = 8.9$ kcal/mol [b]AM1 binding energy is −3.62 kcal/mol [c]AM1 binding energy is 19.27 kcal/mol

Compound	Experiment	AM1 in gas-phase	AM1 with COSMO
 Theophylline	0[a]	1.17	1.40
 Caffeine	5.55	3.26	0.26
 Hypoxanthine	3.00	2.59	3.23
 1-Methylxanthine	1.99	1.93	2.69
 3-Methylxanthine	1.10	0.57	0.12
 Xanthine	1.96	0.0[b]	0.0[c]
r^2		0.557	0.004

Model systems

The theophylline-binding RNA aptamer (Protein Data Base reference code 1O15 [17]) provided the first set for testing the MOZYME methodology. This complex along with five other theophylline analogs (Table 1) was already a target of an extensive study using thermodynamic integration and the Molecular Mechanics Poisson–Boltzmann surface area (MM/PBSA) [18]. 1O15 consists of 33 RNA residues and theophylline [17]—1,086 heavy atoms. Theophylline analogs were built using the theophylline coordinates and then



1a 1cMe 1d

1dAc 2 3

Fig. 1. Trypsin ligands considered in this study. Compound designations are the same as in Ref. [20]

optimized in the binding site of thyophylline using the Amber94 force field [19]. The use of counterions for neutralizing the molecules would require sampling of many conformations as the positions of the counterions are not well-defined. Unfortunately, the number of possible conformations that can be computed using semiempirical methods is very limited owing to relatively high computational cost. Thus, we chose to add hydrogen atoms to neutralize the total charge of −32. These hydrogens are quite far from the position of the ligands and we are confident that such substitutions can be introduced with little or no effect on the binding energies. Thus, our quantum chemical system consisted of 1,118 atoms.

The next test system consisted of trypsin and a series of closely related N^{α}-(2-naphthylsulphonyl)-L-3-amidino-phenylalanine derivatives (Fig. 1) for which thermodynamic and structural data are available [20, 21] (Protein Data Base codes 1K1I, 1K1J, 1K1L, 1K1M, 1K1N). Structure 1K1I was chosen as a reference and other ligands were placed into the active site after superimposition and were minimized. A typical quantum chemical system consisted of about 3,720 atoms with a total charge of +8 on the free enzyme.

Computational method

Quantum chemical calculations at the semiempirical Austin model 1 (AM1) [22] level were conducted using a linear-scaling method (MOZYME) [16] and an implicit solvation model (conductor-like screening model, COSMO) [23] as implemented in MOPAC2002 [24].

Since in general the lowest energy structures predicted by a molecular mechanics force field do not correspond to the low-energy structures at the AM1 quantum chemical level, we performed a partial optimization for each structure using the AM1 method in the gas phase before calculating binding energies with an implicit solvation model (COSMO).

Table 2. Calculation times for oneself-consistent-field calculation (in seconds)

	Number of atoms			
	1,118[a]	3,720[b]	6,194[c]	46,452[d]
Alpha ev68 1 MHz[e]	131.0	1,232.2	2,515.2	35–40 h
Pentium 4 2.66 GHz[f]	111.2	835.5	1,766.5	

[a] Theophylline–RNA complex (PDB reference code 1O15)
[b] Bovine trypsin–inhibitor complex (PDB reference code 1K1I)
[c] β-Secretase–inhibitor complex (PDB reference code 1FKN)
[d] Citrate synthase (PDB reference code 2CTS) in a water droplet
[e] HP Alpha Server SC system with 127 nodes. Each node contains 4×1GHz ev68 (Alpha 21264C) CPUs and between 4 and 16 GB of RAM
[f] Linux cluster with 150 Dell Precision 350 nodes, each containing a 2.66 GHz Pentium4 CPU with 1 GB (533 MHz dual channel) PC1066 RAMBUS

In the theophylline analog–RNA complexes, the optimized region included a ligand and all atoms of the RNA residues lying within 3 Å (about 180 atoms, i.e. about 540 variable parameters). The optimized region of the trypsin complexes consisted of the ligand and crystallographic water molecules lying within 5 Å from the ligand (about 120 atoms, i.e. about 360 variable parameters) with the structure of protein being kept fixed. Prior to calculating the final binding energies, all water molecules were deleted.

Calculation times

CPU times for one self-consistent-field calculation for several model systems on two hardware platforms are shown in Table 2. One can see that large biomolecular systems up to several tens of thousand atoms can now be treated at the semiempirical quantum chemical level of theory in reasonable time, with a full or partial optimization being feasible for systems up to 6,000–7,000 atoms. A typical calculation with partial optimization of the RNA–ligand complex (about 1,120 atoms and 540 variable parameters) took 4–5 h. Partial optimization of the trypsin–inhibitor complexes (about 3,700 atoms and 360 variable parameters) required at least 24 h.

Results and discussion

The relative AM1 interaction energies between the theophylline analogs and RNA versus the experimental data are summarized in Table 1. One can see that the relative AM1 interaction energies including solvation (COSMO) show no correlation with experimental data (regression coefficient $r^2 = 0.004$), with the absolute AM1 interaction energies being positive. By contrast, the relative AM1 gas-phase interaction energies show better agreement with experiment ($r^2 = 0.557$) and their absolute energies are negative.

In the case of trypsin–inhibitor complexes (Table 3), the relative AM1 interaction energies with COSMO more or less correlate with experimental ones ($r^2 = 0.380$) although their absolute interaction energies are again positive. Gas-phase AM1 interaction energies are strongly dependent on the total charge of the ligand, such that the interaction energies are either positive or negative for positively or neutrally charged ligands, respectively.

Table 3. Relative binding energies (in kilocalories per mole) for the trypsin–inhibitor complexes

Compound	Experiment	AM1 in the gas phase	AM1 with COSMO
1a	0.2	87.48	0.0[c]
1k1j(1cMe)	0.0[a]	87.84	3.90
1k1l(1d)	0.7	152.96	7.19
1k1m(1dAc)	0.2	89.20	6.59
2	2.8	0.30	8.67
1k1n(3)	1.3	0.0[b]	11.05

[a] $\Delta G_{\mathrm{Exp}} = -10.4$ kcal/mol
[b] AM1 binding energy is −54.11 kcal/mol
[c] AM1 binding energy is 4.47 kcal/mol

Since binding occurs in an aqueous environment, AM1/COSMO binding energies are more important than gas-phase ones. One can see that the AM1/COSMO approach systematically predicts positive absolute binding energies for diverse types of complexes, proteins and RNA. There are two major sources of errors in the binding energies. First of all, the AM1 method has known drawbacks in describing intermolecular interactions [25, 26, 27, 28]. Secondly, the COSMO method implemented in MOPAC2002 was not parameterized specifically for evaluation of solvation energies of biomolecules, and because the absolute solvation energies for the biomolecule–ligand complexes are approximately 100–1,000 times larger than the AM1 binding energies, even small relative errors in solvation calculations may have a large contribution to the binding energies. So, we think that the positive absolute binding energies obtained in AM1/COSMO calculations reflect the unsatisfactory COSMO parameterization, in particular, the atomic radii.

On the other hand, the relative AM1/COSMO binding energies are not all that unrealistic considering that the present work faces a number of limitations. For example, the conformational space available to the ligand in the target-binding pocket was not sampled to estimate the entropy contribution. This gives hope that a proper calibration of the COSMO solvation method as well as adequate sampling of the target-ligand conformational space will have a major impact on the quality of the semiempirical quantum chemical estimation of binding energies.

Conclusions

In this paper, we presented the first application of a semiempirical quantum chemical method to estimate binding affinities of druglike ligands to large biological molecules. Owing to the increase in computer power and algorithm performance, it is now possible to treat quantum chemically large molecular systems up to several tens of thousand atoms, with the full or partial optimization being feasible for systems of several thousands of atoms. However, to unleash the full potential of semiempirical quantum chemical methods in docking simulations further study has to be done to improve the solvation model.

References

1. Wang R, Lu Y, Wang S (2003) J Med Chem 46:2287
2. Grzybowski BA, Ishchenko AV, Shimada J, Shakhnovich EI (2002) Acc Chem Res 35:261
3. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Proteins Struct Funct Genet 47:409
4. Taylor RD, Jewsbury PJ, Essex JW (2002) J Comput-Aided Mol Des 16:151
5. Gohlke H, Klebe G (2001) Curr Opin Struct Biol 11:231
6. Pérez C, Ortiz AR (2001) J Med Chem 44:3768
7. Muegge I, Rarey M (2001) In: Lipkowitz KB, Boyd DB (eds) Reviews in computational chemistry, vol 17. Wiley-VCH, New York, pp 1–60
8. Ehrlich LP, Wade RC (2001) In: Lipkowitz KB, Boyd DB (eds) Reviews in computational chemistry, vol 17. Wiley-VCH, New York, pp 61–97
9. Jorgensen WL (1989) Acc Chem Res 22:184
10. Aqvist J, Medina C, Samuelsson J-E (1994) Protein Eng 7:385
11. Srinivasan J, Cheatham TE III, Kollman P, Case DA (1998) J Am Chem Soc 120:9401
12. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE III, (2000) Acc Chem Res 33:889
13. Greatbanks SP, Gready JE, Limaye AC, Rendell AP (2000) J Comput Chem 21:788
14. Titmuss SJ, Cummins PL, Rendell AP, Bliznyuk AA, Gready JE (2002) J Comput Chem 23:1314
15. Bliznyuk AA, Rendell AP, Allen TW, Chung S-H (2001) J Phys Chem B 105:12674
16. Stewart JJP (1996) Int J Quantum Chem 58:133
17. Clore GM, Kuszewski J (2003) J Am Chem Soc 125:1518
18. Gouda H, Kuntz ID, Case DA, Kollman PA (2003) Biopolymers 68:16
19. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) J Am Chem Soc 117:5179
20. Dullweber F, Stubbs MT, Musil Đ, Stürzebecher J, Klebe G (2001) J Mol Biol 313:593
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235
22. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902
23. Klamt A, Schüümann G (1993) J Chem Soc Perkin Trans 2 799
24. Stewart JJP (1999) MOPAC 2002, version 1.01. Fujitsu, Tokyo, Japan
25. Voityuk AA, Bliznyuk AA (1988) J Mol Struct (THEOCHEM) 164:343
26. Turi L, Dannenberg JJ (1993) J Phys Chem 97:7899
27. Hobza F, Hubalek F, Kabelac P. Majzlik M, Sponer J, Vondrasek J (1996) Chem Phys Lett 257:31
28. Dannenberg JJ (1997) J Mol Struct (THEOCHEM) 401:279