Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties

Peter Ertl,* Bernhard Rohde, and Paul Selzer

Cheminformatics, Novartis Pharma AG, WKL-490.4.35, CH-4002 Basel, Switzerland

Received July 17, 2000

Molecular polar surface area (PSA), i.e., surface belonging to polar atoms, is a descriptor that was shown to correlate well with passive molecular transport through membranes and, therefore, allows prediction of transport properties of drugs. The calculation of PSA, however, is rather time-consuming because of the necessity to generate a reasonable 3D molecular geometry and the calculation of the surface itself. A new approach for the calculation of the PSA is presented here, based on the summation of tabulated surface contributions of polar fragments. The method, termed topological PSA (TPSA), provides results which are practically identical with the 3D PSA (the correlation coefficient between 3D PSA and fragment-based TPSA for 34 810 molecules from the World Drug Index is 0.99), while the computation speed is 2−3 orders of magnitude faster. The new methodology may, therefore, be used for fast bioavailability screening of virtual libraries having millions of molecules. This article describes the new methodology and shows the results of validation studies based on sets of published absorption data, including intestinal absorption, Caco-2 monolayer penetration, and blood−brain barrier penetration.

Introduction

Fast and reliable estimation of molecular transport properties, particularly intestinal absorption and blood−brain barrier penetration, is one of the key factors accelerating the process of drug discovery and development.[1−4] Traditionally, calculated values of octanol/water partition coefficient have been used for this purpose. In recent years, several new parameters have been introduced for absorption prediction, including molecular size and shape descriptors, hydrogen-bonding capabilities, and surface properties.[3−8] A set of rules imposing limitations on log *P*, molecular weight, and number of hydrogen bond donors and acceptors (known as the "rule of five") introduced by Lipinski[8] has become particularly popular. Another very helpful parameter for the prediction of absorption is the polar surface area (PSA) defined as the sum of surfaces of polar atoms in a molecule. This parameter is easy to understand and, most importantly, provides good correlation with experimental transport data. It has been successfully applied for the prediction of intestinal absorption,[9,11] Caco-2 monolayers penetration,[12−15] and blood−brain barrier crossing.[16−17] Various protocols have been reported to calculate the PSA, differing in the definition of "polar atoms" (some authors regard only N and O atoms to be polar, whereas other approaches include also other heteroatoms), different methodologies for generating the 3D structure (CORINA, CONCORD, geometry optimization, conformational sampling), or the surface itself (van der Waals, Connolly, or Lee−Richards). According to our in-house experience, however, the results of these various approaches are highly correlated, even when absolute values may differ due to differences in computational protocols and different sets of atomic radii used. Time requirements to calculate the PSA are generally high, up to tens of minutes per molecule when the geometry is optimized or a conformational search is performed, although recently a fast single-conformer method for PSA calculation was reported with a throughput of several molecules per second.[18] Furthermore calculations require specialized software to generate 3D molecular structures and to determine the surface itself.

In today's era of drug development, shaped by high-throughput screening and combinatorial chemistry, fast bioavailability screening of virtual libraries consisting of hundreds of thousands even millions of molecules is required. For this reason we developed a new, fast, and straightforward protocol to calculate the PSA based on the topological information only.

Methodology

The new methodology for the calculation of PSA termed TPSA (topological PSA) described here is based simply on the summation of tabulated surface contributions of polar fragments (i.e. atoms regarding also their bonding pattern). The workflow of the new methodology compared to the traditional way to calculate PSA is shown schematically in Figure 1.

The contributions of polar fragments were determined by least-squares fitting of the fragment-based TPSA to the single conformer 3D PSA for a large set of druglike structures. Molecules from the World Drug Index[19] were used for this procedure. This database was preprocessed by removing molecules with apparent valence errors, molecular weights outside the interval 100−800, and molecules not having at least one oxygen, nitrogen, sulfur, or phosphorus atom. This

* To whom correspondence should be addressed. Phone: +41 61 69 67413. Fax: +41 61 69 67416. E-mail: peter.ertl@pharma.novartis.com.

10.1021/jm000942e CCC: $19.00
Published on Web 09/19/2000

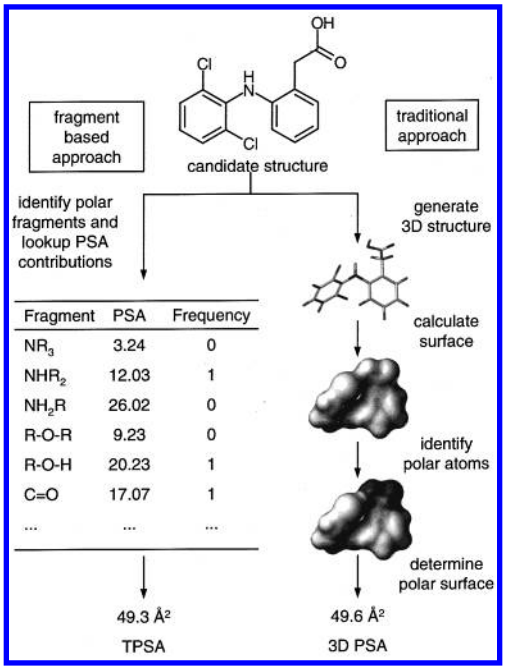

Figure 1. Comparison of the new methodology with the traditional way to calculate PSA.

Table 1. Atomic Contributions (Å^2) to PSA

atom type[a]	PSA contrib	atom type[a]	PSA contrib
[N](−*)(−*)−*	3.24	[nH](:*):*	15.79
[N](−*)=*	12.36	[n+](:*)(:*):*	4.10
[N]#*	23.79	[n+](−*)(:*):*	3.88
[N](−*)(=*)=* [b]	11.68	[nH+](:*):*	14.14
[N](=*)#* [c]	13.60	[O](−*)−*	9.23
[N]1(−*)−*−*−1[d]	3.01	[O]1−*−*−1[d]	12.53
[NH](−*)−*	12.03	[O]=*	17.07
[NH]1−*−*−1[d]	21.94	[OH]−*	20.23
[NH]=*	23.85	[O−]−*	23.06
[NH2]−*	26.02	[o](:*):*	13.14
[N+](−*)(−*)(−*)−*	0.00	[S](−*)−*	25.30
[N+](−*)(−*)=*	3.01	[S]=*	32.09
[N+](−*)#* [e]	4.36	[S](−*)(−*)=*	19.21
[NH+](−*)(−*)−*	4.44	[S](−*)(−*)(=*)=*	8.38
[NH+](−*)=*	13.97	[SH]−*	38.80
[NH2+](−*)−*	16.61	[s](:*):*	28.24
[NH2+]=*	25.59	[s](=*)(:*):*	21.70
[NH3+]−*	27.64	[P](−*)(−*)−*	13.59
[n](:*):*	12.89	[P](−*)=*	34.14
[n](:*)(:*):*	4.41	[P](−*)(−*)(−*)=*	9.81
[n](−*)(:*):*	4.93	[PH](−*)(−*)=*	23.47
[n](=*)(:*):* [f]	8.39		

[a] An asterisk (*) stands for any non-hydrogen atom, − for a single bond, = for a double bond, # for a triple bond, : for an aromatic bond; atomic symbol in lowercase means that the atom is part of an aromatic system. [b] As in nitro group. [c] Middle nitrogen in azide group. [d] Atom in a three-membered ring. [e] Nitrogen in isocyano group. [f] As in pyridine *N*-oxide.

yielded a set consisting of 34 810 reasonably druglike molecules. These molecules contain the 43 polar atom types listed in the Table 1. In addition to commonly used polar fragments with oxygen and nitrogen, we included also "slightly polar" fragments containing phosphorus and sulfur. It has been reported that PSA with inclusion of sulfur atoms provides better correlation with human jejunum permeability than just O- and N-based PSA.[7] The broader list of fragments provides the possibility to try various models and select the best one for a particular dataset. An even better solution would be to scale contributions of polar fragments according to the strength of the hydrogen bonds they form (as already suggested in ref 10).

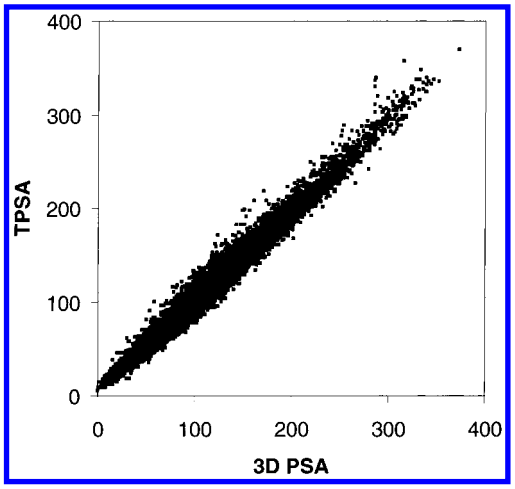

Figure 2. 3D PSA vs TPSA for 34 810 molecules from the World Drug Index.

The goal of the fitting procedure was to determine such values for surface fragment contributions (c(fragment$_i$) in eq 1) which provide maximum correlation with the 3D PSA (eq 1):

$$\text{3D PSA} = \sum_{i}^{n_{\text{types}}} n_i \cdot c(\text{fragment}_i) \qquad (1)$$

where 3D PSA is the traditionally calculated PSA (based on 3D molecular structure), n_{types} is the number of types of polar fragments, c(fragment$_i$) is the coefficient to optimize (i.e. surface contribution of fragment i), and n_i is the frequency of fragment i in the molecule.

The 3D PSA used as a target in the fitting was calculated from CORINA[20] geometries considering the van der Waals surfaces belonging to O, N, S, and P atoms, including also their attached hydrogens. All structure manipulation, processing of SMILES, identification of polar fragments, statistical analysis, etc., were done by using an in-house molecular development kit written in Java.

The statistical analysis provided very good correlation between 3D PSA and TPSA with the following statistical parameters: $r^2 = 0.982$, $r = 0.991$, $\sigma = 7.83$ Å^2, average absolute error = 5.62 Å^2. The fragment contributions obtained from eq 1 are listed in Table 1 and a graph of 3D PSA vs TPSA is shown in Figure 2.

The only class of molecules showing larger deviations between TPSA and 3D PSA (visible as outliers in Figure 2) are large macrocycles with many polar substituents. These substituents are usually buried in the center of the ring and are therefore not accessible to solvent. The calculated TPSA for such systems is too large compared with the 3D PSA. Another question is, however, how accurately the single conformer we used in the fitting procedure represents such complex molecules. In these cases, the polar surface obtained as an average of several representative ring conformations would probably provide a better fit to the TPSA.

Validation of the Methodology

The new TPSA methodology has been validated by using published data of various types of drug transport properties including intestinal absorption, blood−brain barrier penetration, and Caco-2 cell permeability. All these datasets have been already studied by using PSA descriptors calculated by different protocols. The results of validation studies are summarized in Table 2, where the squares of correlation coefficients (r^2) between observed molecular transport properties and 3D PSA used in the original publications, as well as the new TPSA, are given. The correlation between TPSA and 3D

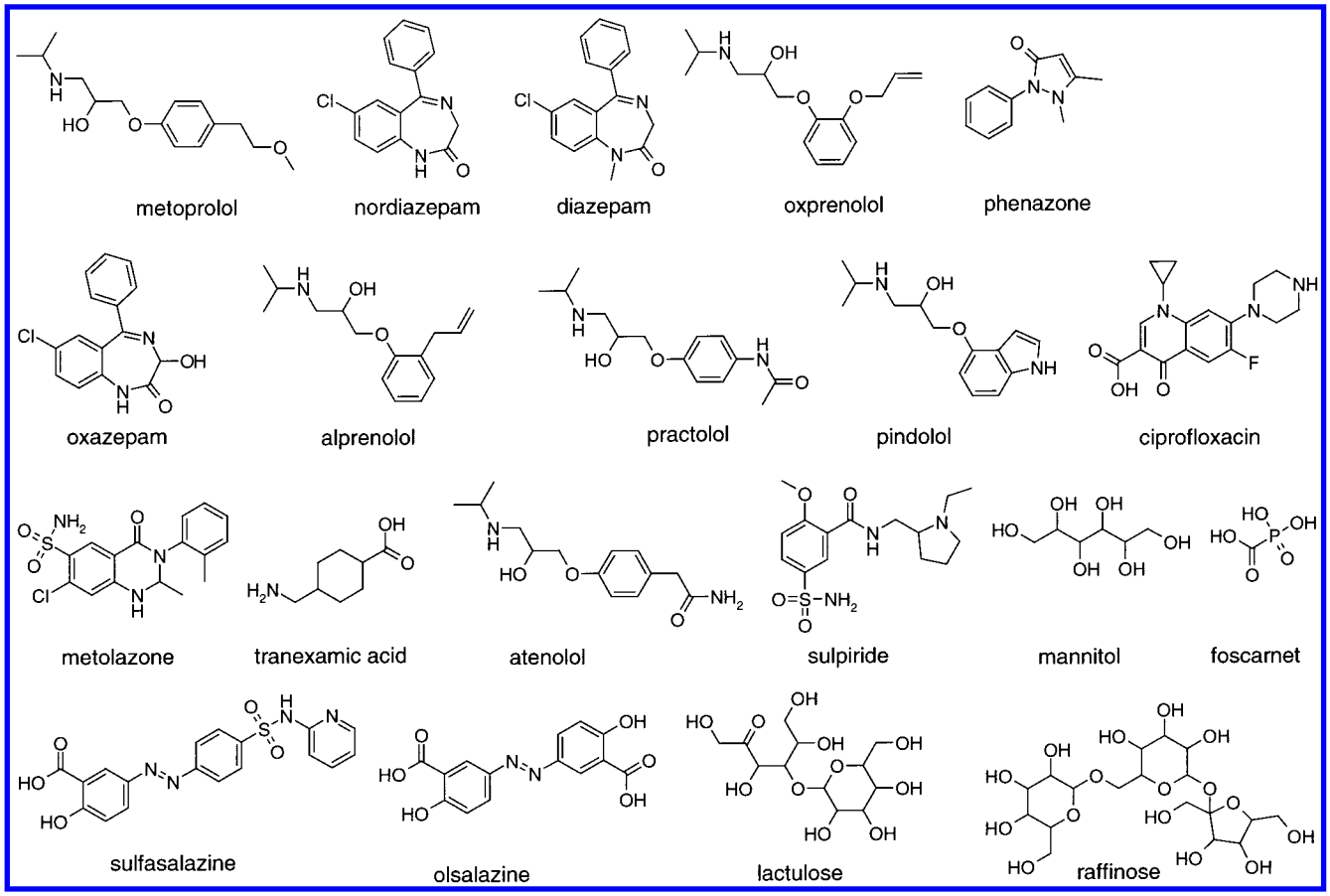

Figure 3. Molecules from the Palm dataset.[9]

Table 2. Summary of Validation Studies

bioavailability type	n^a	r^2 TPSA	r^2 3D PSA	r^2 TPSA/ 3D PSA
oral drug absorption	20	0.91	0.94[b]	0.98
Caco-2 permeability	9	0.96	0.98[c]	0.99
blood−brain barrier penetration	45	0.78	0.84[d]	0.95
human jejunum permeability	13	0.75	0.76[e]	0.99
Caco-2 permeability	17	0.56	0.45[f]	0.94
blood−brain barrier penetration	57	0.66	0.67[g]	0.99

[a] Number of molecules in the dataset. [b] Dynamic PSA, sigmoidal model.[9] [c] Dynamic PSA in water environment, linear model[13] r^2 for sigmoidal model = 0.99. [d] Value for dynamic PSA; for single-conformer PSA $r^2 = 0.78$.[17] [e] Also sulfur fragments were included in the PSA calculation.[7] [f] Ref 5. [g] Ref 16.

PSA is also included. In all cases, the TPSA methodology performs quite well, providing results of the same quality as the computationally much more demanding 3D PSA (including so-called dynamic PSA,[12] based on the Boltzmann-weighted average values computed from an ensemble of low-energy conformations obtained by a detailed conformational search). All molecular datasets described in Table 2 including molecular structures, experimental absorption data, and calculated TPSA are available as Supporting Information.

In Table 3 the PSA values calculated by three different methods, namely dynamic PSA,[9] single-conformer PSA,[10] and TPSA, are compared for 20 representative drugs from the dataset[9] (Figure 3). Both 3D methods use the same van der Waals atomic radii reported in ref 10. The methods provide highly correlated results, namely: dynamic PSA vs TPSA $r^2 = 0.982$; single-conformer PSA vs TPSA $r^2 = 0.991$; dynamic PSA vs single-conformer PSA $r^2 = 0.993$.

Table 3. Comparison of Calculated PSA Values for Compounds from Palm et al.[9]

name	% FA[a]	TPSA[b]	dyn 3D PSA[c]	3D PSA[d]
metoprolol	102	50.7	53.1	57.2
nordiazepam	99	41.5	45.1	47.5
diazepam	97	32.7	33.0	34.5
oxprenolol	97	50.7	46.8	53.2
phenazone	97	26.9	27.1	28.0
oxazepam	97	61.7	66.9	55.6
alprenolol	96	41.9	37.1	41.8
practolol	95	70.6	73.4	77.2
pindolol	92	57.3	56.5	60.9
ciprofloxacin	69	74.6	78.7	80.1
metolazone	64	92.5	94.5	95.9
tranexamic acid	55	63.3	69.2	71.5
atenolol	54	84.6	90.9	93.3
sulpiride	36	101.7	100.2	101.4
mannitol	26	121.4	116.6	129.6
foscarnet	17	94.8	115.3	117.3
sulfasalazine	12	141.3	141.9	148.6
olsalazine	2.3	139.8	141.0	147.0
lactulose	0.6	197.4	177.2	197.8
raffinose	0.3	268.7	242.1	266.8

[a] Percent (%) of drug absorbed after oral administration. [b] Fragment-based PSA. [c] Dynamic PSA.[9] [d] Single-conformer PSA.[10]

The main advantage of our fragment-based contribution method is the very high throughput, since the only processing step required is the identification of polar fragments in the molecules under study. By using topological information only (atomic connectivity) contained in the molecule's SMILES code,[21] the Java program is able to process more than 8000 molecules/min on a standard 450-MHz PC. This allows calculation

of PSA descriptors for very large databases or combinatorial libraries. Another advantage of the topological approach is the fact that the fragment contributions have been obtained by averaging surface values for geometries of tens of thousands of representative drug-like molecules. In this way, the effect of "conformational sampling" has been taken into account. This is otherwise only available at the expenses of a computationally demanding dynamic PSA approach.

Conclusions

A new methodology to calculate molecular polar surface area as a sum of fragment contributions has been described. The method is straightforward and does not require any computationally demanding steps such as 3D structure generation and surface calculation. The only input needed is the molecular topology (i.e. SMILES string). The method is therefore extremely fast and thus allows virtual bioavailability screening of very large collections of molecules. Despite this, the results are of a quality comparable with those obtained by using computationally much more expensive approaches, as documented by validation studies with published transport data including intestinal absorption, oral bioavailability, and blood−brain barrier penetration.

Acknowledgment. We thank our colleagues Sigmar Dressler and David Wadsworth for critical reading of the manuscript and for helpful comments.

Supporting Information Available: All molecular datasets described in Table 2 (data include molecular structures in computer-readable form,[22] experimental bioavailability, and calculated TPSA). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Blake, J. F. Chemoinformatics − Predicting the Physicochemical Properties of 'Drug-like' Molecules. *Curr. Opin. Biotechnol.* **2000**, *11*, 104−107.
(2) George, A. The Design and Molecular Modeling of CNS Drugs. *Curr. Opin. Drug Discovery Dev.* **1999**, *2*, 286−292.
(3) Clark, D. E.; Pickett, S. D. Computational Methods for the Prediction of Drug-likeness. *Drug Discovery Today* **2000**, *5*, 49−58.
(4) Stenberg, P.; Luthman, K.; Artursson P. Virtual Screening of Intestinal Drug Permeability. *J. Control. Relat.* **2000**, *65*, 231−243.
(5) van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Estimation of Caco-2 Cell Permeability using Calculated Molecular Descriptors. *Quant. Struct.-Act. Relat.* **1996**, *15*, 480−490.
(6) van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Chretien, J. R.; Raevsky, O. A. Estimation of Blood-Brain Barrier Crossing of Drugs Using Molecular Size and Shape, and H−Bonding Descriptors. *J. Drug Target* **1998**, *15*, 480−490.
(7) Winiwarter, S.; Bonham, N. M.; Ax. F.; Hallberg, A.; Lennernäs, H.; Karlén A. Correlation of Human Jejunal Permeability (in Vivo) of Drugs with Experimentally and Theoretically Derived Parameters. A Multivariate Data Analysis Approach. *J. Med. Chem.* **1998**, *41*, 4939−4949.
(8) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(9) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharm. Res.* **1997**, *14*, 568−571.
(10) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 1. Prediction of Intestinal Absorption. *J. Pharm. Sci.* **1999**, *88*, 807−814.
(11) Stenberg, P.; Luthman, K.; Ellens, H.; Pin Lee, Ch.; Smith, P. L.; Lago, A.; Elliot, J. D.; Artursson, P. Prediction of the Intestinal Absorption of Endothelin Receptor Antagonists Using Three Theoretical Methods of Increasing Complexity. *Pharm. Res.* **1999**, *16*, 1520−1526.
(12) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Artursson P. Correlation of Drug Absorption with Molecular Surface Properties. *J. Pharm. Sci.* **1996**, *85*, 32−39.
(13) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson P. Evaluation of Dynamic Polar Molecular Surface Area as Predictor of Drug Absorption: Comparison with Other Computational and Experimental Predictors. *J. Med. Chem.* **1998**, *41*, 5382−5392.
(14) Krarup, L. H.; Christensen, I. T.; Hovgaard, L.; Frokjaer, S. Predicting Drug Absorption from Molecular Surface Properties Based on Molecular Dynamics Simulations. *Pharm. Res.* **1998**, *15*, 972−978.
(15) Stenberg, P.; Luthman, K.; Artursson, P. Prediction of Membrane Permeability to Peptides from Calculated Dynamic Molecular Surface Properties. *Pharm. Res.* **1999**, *16*, 205−212.
(16) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 2. Prediction of Blood-Brain Barrier Penetration. *J. Pharm. Sci.* **1999**, *88*, 815−821.
(17) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J.-P. Polar Molecular Surface as a Dominating Determinant for Oral Absorption and Brain Penetration of Drugs. *Pharm. Res.* **1999**, *16*, 1514−1519.
(18) Pickett, S. D.; McLay, I. M.; Clark, D. E. Enhancing the Hit-to-Lead Properties of Lead Optimization Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 263−272.
(19) *World Drug Index Database WDI97;* Derwent Publications Ltd., distributed by Daylight Chemical Information Systems, Inc.
(20) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(21) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−26.
(22) Molecular structures available in the Supporting Information are encoded as SMILES strings. A free SMILES visualizer is available at http://www2.chemie.uni-erlangen.de/services/gifcreator/index.html.

JM000942E