J Am Oil Chem Soc (2010) 87:271–279
DOI 10.1007/s11746-009-1493-1

ORIGINAL PAPER

Computational Study of Macroscopic Properties of Macromolecules with Industrial Interest

Kepa K. Burusco · Carlos Jaime · Felicidad Franch-Lage · Lluís Beltran · Francesc Granero

Received: 28 May 2009 / Revised: 16 September 2009 / Accepted: 14 October 2009 / Published online: 12 November 2009

Abstract There is an increasing demand on the market for environmentally compatible lubricants. Refined vegetable oils have been used as biolubricants, but synthetic esters from renewable resources could also be considered as biolubricants, and can be prepared by combining different alcohols with linear or branched fatty acids. A tool has been developed to predict the properties of esters based on the chemical structure. A multilinear approach was used to correlate the experimental viscosity with theoretical parameters (diffusion coefficient, dipole moment and solvation energy) calculated from the expected chemical structure with a high degree of correlation.

Keywords Biolubricants · Fatty acid esters · Viscosity prediction

Introduction

Currently there is an increasing demand for environmentally compatible lubricants, particularly in areas where they can come into contact with water, food or people. From a functional point of view, lubricants decrease friction and wear, and also play other roles such as heat transfer, particle suspension in the engine, liquid sealing, antirust and water removal. To fulfil these requirements, lubricants need to have certain properties, such as cold stability, oxidation stability, hydrolytic stability, viscosity and a viscosity index, to name the most characteristic ones.

Biolubricants should have all the functionalities required by end users, but must also be biodegradable and have low environmental toxicity. Refined vegetable oils have been used as biolubricants, but synthetic esters, which may be partly derived from renewable resources, could also be considered as biolubricants. Moreover, synthetic esters offer an opportunity to modulate and design lubricants adapted to the functional requirements. They can be prepared by combining alcohols and polyols and linear or branched fatty acids.

In this study we have developed a tool for predicting the properties of esters based on the chemical structure. The chemical structure can be determined from theoretical computations before any chemical synthesis and certain parameters can be calculated using this chemical structure. Based on these calculated parameters we can estimate specific properties in order to select appropriate future functional molecules to be synthesized. For example, one of the critical parameters when designing a lubricant is the viscosity (μ) and, by combining the Stokes–Einstein equation (Eq. 1) with the Einstein–Smoluchowsky equation (Eq. 2) we obtain an expression (Eq. 3) that correlates viscosity (μ) with the temperature (T), time (t), molecular radius (r) and diffusion coefficient (autodiffusion if the solvent and solute are identical) (D) or average square displacement $\langle x^2 \rangle$. Consequently, it is possible to predict the viscosity of compounds by computing the parameters of chemical structures and using standard parameters (time, temperature, etc.).

$$D = \frac{k_B T}{6\pi\mu r} \quad (1)$$

Dedicated to Prof. Pelayo Camps on his 65th anniversary.

K. K. Burusco · C. Jaime (✉) · F. Franch-Lage
Department of Chemistry, Faculty of Sciences,
Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
e-mail: carlos.jaime@uab.cat

L. Beltran · F. Granero
Industrial Química Lasem, Av. de la Indústria no. 7,
Polígon Industrial Pla del Camí, 08297 Castellgalí, Spain

$$\langle x^2 \rangle = 6Dt \quad (2)$$

$$\mu = \frac{k_B T}{6\pi D r} = \frac{k_B T t}{\pi r \langle x^2 \rangle} \quad (3)$$

Experimental Details

Building Molecules

Five residues were defined (those for Gly, Tmp, Pen, Lin and Ole, Fig. 1). Six molecules were built by combining these five residues. This approach has two main advantages over the classical one: (1) each residue or building block is much smaller than the molecule itself (computational time will be shorter); and (2) if a new alcohol or fatty acid is used, it is only necessary to compute the corresponding residue (once again saving computational time).

The two residues derived from cutting an ester through the CO–O bond (one alkoxyl and one alkanoyl) are then assembled by adding an acetyl group to the alkoxy fragment and a methoxyl group to the alkanoyl fragment (Fig. 1).

The residues were then built through atomic charge computation by gausian98 and creation of units by Amber modules.

Molecular Dynamics Simulations

All computations were performed using the parm99 [1] force field implemented in the AMBER [2] v.7 program package [3, 4] because it allows molecular dynamics simulations to be computed and the trajectory files to be analyzed to obtain the geometrical properties of the molecule studied.

Long molecular dynamics simulations (MD) were performed in two steps always under vacuum conditions: solvent box not included, molecule not centered, SHAKE [5, 6] protocol (bonds involving hydrogen atoms are constrained), and cut-off of 12 angstroms for electrostatic and non-bonded interactions. The first step is the *heating slope* and *equilibration* process performed under in vacuo conditions. This step lasts 300 ps, with a time step of 1 fs, 298 K in equilibrium, and the thermal coupling constant is modulated along the simulation to avoid *blow-up* terminations. The second step is the *sampling* process, which is also performed under in vacuo conditions. This step lasts for 5,000 ps, with a time step of 1 fs, 298 K, bath thermal coupling of 0.5 ps and sampling frequency of 1.0 snapshot per picosecond; the trajectory is stored in this step for further analysis. The thermal coupling constant remains unchanged throughout the simulation at a value of 0.5 ps.

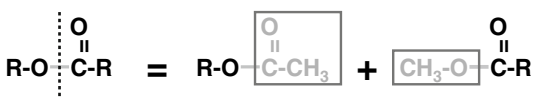

Fig. 1 The two residues derived from cutting an ester through the CO–O bond

The set of molecular parameters used in this work was obtained through analysis of MD trajectories employing AMBER 7 analysis modules: Radius of Gyration (CARNAL), Dipolar Moment (ptraj), Diffusion (ptraj) and GBSOL Solvation Free Energy (MM-GBSA).

The coefficient diffusion is directly related to the mean square deviation obtained by the ptraj module, although care must be taken with these computed values because the molecular dynamics simulations were performed under in vacuo conditions while the experimental coefficient diffusion is obtained in very different conditions (pure liquid).

Determining Viscosity

Viscosity was measured at the Industrial Química Lasem laboratories according to ASTM D-445/65, corresponding to the "Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids", using a viscosimeter Cannon–Fenske routine. Six measures were taken for the IsoOle, GlyOle and PenOle, while eight measures were taken for TmpOle.

Results and Discussion

The macromolecules studied were esters derived from the reaction between long fatty acids like linoleic (Lin), oleic (Ole), stearic (Ste) and isostearic (Ist) acids, and polyfunctional alcohols like glycerine (Gly), trimethylolpropane (Tmp), and pentaerythrite (Pen) (Fig. 2). Some isopropyl (Iso) esters were used as the standard. The esters were named

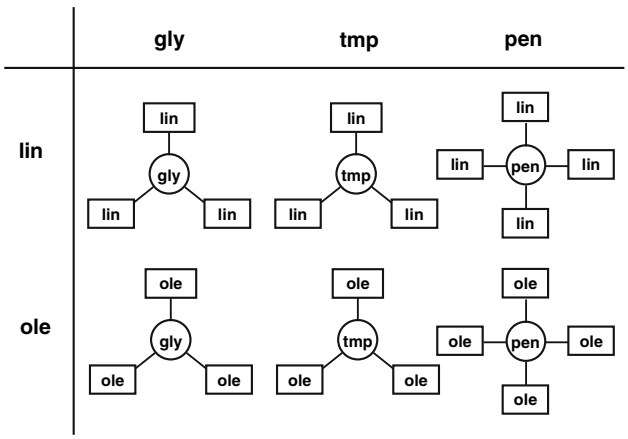

Fig. 2 Schematic structure of the esters studied in this work (for the sake of simplicity only those derived from the linoleic (Lin) and oleic (Ole) acids are represented)

Table 1 Mean square displacement (D), radius of gyration (RadGyr) and computed (as single units) and experimental viscosities (μ, in cStokes) for the studied molecules

	D (Å^2/ps) × 10^6	RadGyr (Å)	"μ"calc	μ exp
Centered				
GlyLin	433	6.01	383.8	NA
GlyOle	367	6.07	449.6	39.07 ± 0.57
TmpLin	1,133	6.12	144.1	NA
TmpOle	1,033	6.08	159.2	46.47 ± 0.05
PenLin	517	6.55	295.5	NA
PenOle	283	6.45	547.5	66.72 ± 0.42
Non-centered				
IsoOle	2,750	5.00	72.8	5.16 ± 0.13
GlyLin	2,033	6.02	81.7	NA
GlyOle	333	6.02	498.2	39.07 ± 0.57
TmpLin	2,200	6.10	74.5	NA
TmpOle	1,667	6.02	99.6	46.47 ± 0.05
PenLin	83	6.46	1,858.3	NA
PenOle	33	6.49	4,624.7	66.72 ± 0.42

Non available data is indicated by NA

by combining the alcohol and acid acronyms to form a six letter word (e.g., IsoOle for isopropyl oleate).

Viscosity Computed from Single Molecules

Only seven biolubricants were studied: IsoOle, GlyOle, TmpOle, PenOle, GlyLin, TmpLin and PenLin. Molecules were initially built using Macromodel [7] version 9.0 tools and were then fully minimized in the same program (MM3* [8] Force-Field, 5,000 steps, Polak-Ribiere Conjugate Gradient [9] with a derivative criterion of 0.001 kJ·mol^{-1} Å^{-1}). Atomic charges from the resulting structures were computed using gaussian-98 [10] with the following conditions: Restricted Hartree–Fock, set of basis functions 6-31G*, and Merz-Kollman [11] charges.

Molecular dynamics simulations of 10 ns length were performed to compute the diffusion coefficient [12, 13] and the radius of gyration for each studied molecule. The diffusion coefficient is directly proportional to the slope of the "mean square displacement" [14]. The slope is obtained directly from Eq. 4, where "d" is the dimension of the space.

$$\lim_{t\to\infty}\frac{\mathrm{d}}{\mathrm{dt}}\langle\Delta r_i(t)^2\rangle = 2dD \tag{4}$$

Molecules normally translate during the molecular dynamics simulation, and to stop molecules from escaping from the given margins the centering option in Amber was used to center the molecule after a certain number of picoseconds (values obtained using this option are shown under the "centered" caption in Table 1). The obtained slopes of the "mean square displacement" were always positive except for GlyOle, which was negative. This value is physically impossible, and we attributed the negative diffusion to the very large initial displacement of the molecule followed by much smaller diffusion. In cases like this, the absolute value was used. The apparent viscosity ("μ") given in Table 1 corresponds to $1/r\cdot\langle x^2\rangle$: a value which is proportional to the viscosity obtained from Eq. 3.

Another set of 5 ns molecular dynamics simulations (the simulation for IsoOle was only 2 ns long due to the larger movement of this molecule) was performed without using the centering option (data can be found under the "non-centered" caption in Table 1). Figure 3 shows the graphical results for two compounds, TmpLin and IsoOle, as examples.

A detailed analysis of the results presented in Table 1 indicates that although the values of the two methodologies are not totally coincident, the values from the non-centered simulations roughly agree with the experimental viscosities, and indicate that the viscosity increases with the molecular size. There are some disagreements for molecules of intermediate size, but they seem to come from the computed diffusion coefficient, which is very different in the two methods and does not follow a clear tendency, most probably because the molecular dynamics simulations were performed under in vacuo conditions while the experimental coefficient diffusion is obtained in very different conditions (pure liquid). In summary, a reasonable general

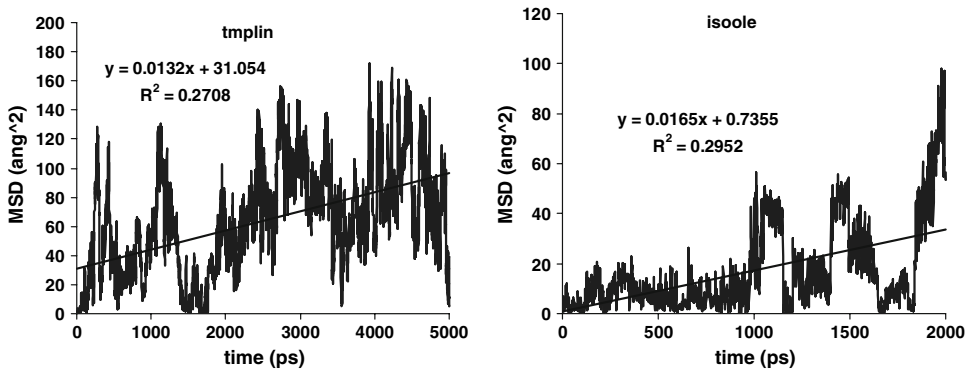

Fig. 3 Graphical representation of the "mean square displacement" obtained using the non-centered option and showing the average value (slope of the tendency line) for TmpLin and IsoOle as examples

agreement between the computed and experimental viscosities was obtained, although the methodology employed was not flexible enough to take advantage of the modular function of Amber.

Viscosity Computed from Modular Sets (Residues)

Building large molecules from a library of previously created independent fragments (called residues) is a methodology that takes advantage of the modular function of Amber. In the present case, in which we are studying branched esters, a library containing residues for the alcohols and fatty acids would be very useful, and would allow a large number of molecules to be built easily.

The computations of the corresponding residues for Gly, Tmp, Pen, Ole and Lin were carried out, as well as those for Iso (see computational details). These residues were used to build seven molecules (IsoOle, GlyOle, GlyLin, TmpOle, TmpLin, PenOle and PenLin). Their viscosity was computed from the mean square displacement and radius of gyration as indicated above in the computations from single molecules section.

The comparison of the computed viscosities shows that the two methodologies are absolutely equivalent and that they give similar results. Figure 4 contains the representation of the relationship between the theoretical viscosity obtained from molecules built as a unique set or as fragments.

Least-Square Multilinear Regression Approach

The results from the previous studies clearly indicate that the computed viscosities are in more than reasonable agreement with the experimentally obtained data, although it is also very clear that the mean square displacement (related to the diffusion coefficient) was not properly computed.

Therefore, a least-square multilinear regression approach was designed in which the parameters for computing viscosity were deduced from molecular parameters, such as molecular size and intermolecular interactions. It is well known that viscosity increases with molecular size and branching. It is also very reasonable to expect larger viscosity values for molecules with a large capacity to interact with each other. The molecular size is dependent on the number of carbon atoms in the molecule, the molecular weight and the molecular radius of gyration, while intermolecular interactions are related to molecular properties such as molecular dipole moment and solvation energy. The temperature used in the molecular dynamics simulations was also included as a component in the multilinear approach.

The first tests were carried out with a total of eight molecules: isopropyl, glycerine, trimethylolpropane and pentaerythrite oleates and isostearates (Ist). It is important to make a note on isostearic acid at this point. Commercial isostearic acid is a mixtures of different acids, but in this work we only considered what is generally its main component: 2-ethylhexadecanoic acid. All eight molecules were built from the previously created library of residues and were subjected to long molecular dynamics simulations of 5 ns in a vacuum. Table 2 shows all the previously mentioned properties of the computed molecular as well as the experimental viscosity.

These computed values were used to obtain a total of sixteen different sets of computed viscosities. These sets differ in the molecular property used in the multilinear approach, although all sets consider the dipole moment and the solvation energy as variables. Sets that gave rise to negative viscosity values were discarded. Table 3 shows the seven sets (marked with letters A–G) that gave positive viscosity values. The best correlation was considered to be the one with: slopes close to one, lines passing through the origin and a correlation coefficient (r^2) that is also close to one. Figure 5 shows the best correlation obtained (set "F" in Table 3) and indicates that the computed viscosity can be obtained from the radius of gyration, dipole moment, solvation energy, temperature, number of carbon atoms and molecular weight. No conclusions can be extracted from the coefficients of the different molecular properties because variables are not normalized; however, the best expression for computing viscosity is as follows:

$$\mu_{(\text{theo.})} = -51.6[\text{RGyr}] + 182.5[\text{Dip.Mom.}] + 1.2[\text{GBSOL}] + 0.5[\text{T}] - 65.6[\text{no.C}] + 4.4[\text{MW}]$$

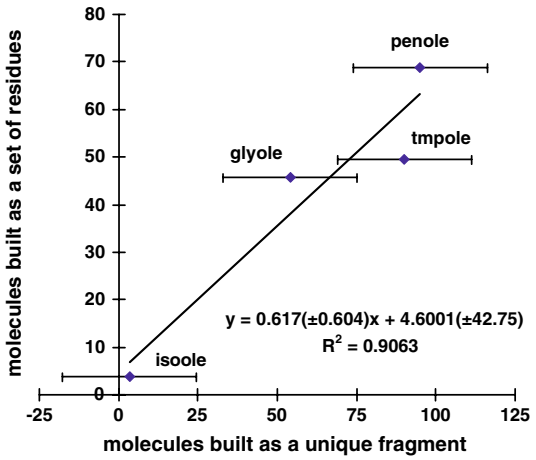

Fig. 4 Graphical representation of the relationship between the viscosities computed as molecules built as a unique fragment or as a set of residues. The slope and intercept are given with their confidence intervals at a confidence level of 95%

Table 2 Computed molecular properties

Molecule	D (Å^2/ps) × 10^8	RadGyr (Å)	Dip.Mom. (Debye)	GBSol (kcal/mol)	*T* (K)	No. C	MW (g/mol)	μ exp[a] (cSt.)
IsoOle	2.11	4.60	0.30	−0.11	298	21	324	5.1
IsoIst	2.82	5.06	0.31	1.34	298	21	326	7.1
GliOle	0.14	6.08	0.62	−3.14	298	57	884	39.0
GliIst	0.41	6.32	0.77	0.89	298	57	890	145.4
TmpOle	0.01	6.05	0.61	−2.16	298	60	926	46.0
TmpIst	0.17	6.36	0.57	1.59	298	60	932	91.6
PenOle	0.13	6.59	0.57	−3.03	298	77	1,192	68.0
PenIst	0.08	6.31	0.45	1.83	298	77	1,200	140.0

(*D* diffusion coefficient, *RadGyr* radius of gyration, *Dip.Mom.* dipole moment, *GBSol* solvation energy, *T* temperature, *no.C* number of carbon atoms, *MW* molecular weight) for the eight molecules used in the multilinear approach and experimental viscosity (μ, cStokes at 313 K). Molecules were built as single units

[a] Viscosity was measured at the Industrial Química Lasem laboratories according to ASTM D-445/65

Table 3 Coefficients of the different sets of multilinear approaches (from A to G) used to obtain the computed viscosities

(*D* diffusion coefficient, *RadGyr* radius of gyration, *DipMom.* dipole moment, *GBSol* solvation energy, *T* temperature, *no.C* number of carbon atoms, *MW* molecular weight, *Slope* slope of the linear regression, *Ord.* value at origin, *r2* correlation coefficient). Molecules were built from residues

Variable/entry	A	B	C	D	E	F	G
D	2,100.5	−2,848.7					
1/*D* × 10^3			0.0	0.0	−1.4		
RadGyr	−37.8	8.0				−51.6	−8.6
1/RadGyr			−1,377.4	1,647.1	−332.7		
Dip.Mom.	216.4	90.7	109.7	198.0	249.4	182.5	239.3
GBSol	12.3	17.4	14.4	12.8	15.6	12.0	15.0
T			0.9	−1.5		0.5	
No. C	−49.1			−56.9		−65.6	
MW	3.4			3.8		4.4	
Slope	0.8	1.0	1.0	1.2	1.0	1.0	1.1
Ord.	−14.7	−0.9	−17.2	12.9	2.4	0.0	−10.7
r^2	0.96	0.86	0.84	0.95	0.77	0.98	0.64
Goodness	+	++	−	+	+	+++	−

Prediction of Viscosity Values for Novel Molecules

Once an excellent correlation was obtained, the next step of the work was to predict the viscosity for a total of 40 novel molecules. Only the best correlation was used for this prediction (correlation F in Table 3). The novel molecules were obtained by random combinations of the corresponding residues, and were named by combining the acronym of the alcohol with that of the fatty acid. Figure 6 contains the representation of all the residues used.

It is worth mentioning here once again that the final goal of this work is to design a new biolubricant prepared from esterification of polyols and fatty acids. With the aim of finding a novel molecule with large viscosity, the oleic acid was chosen as one of the best targets because it can be derived later by transforming the double bond in oxygenated functions. Therefore, two new residues were constructed: the oleic acid with one new ester group in one of

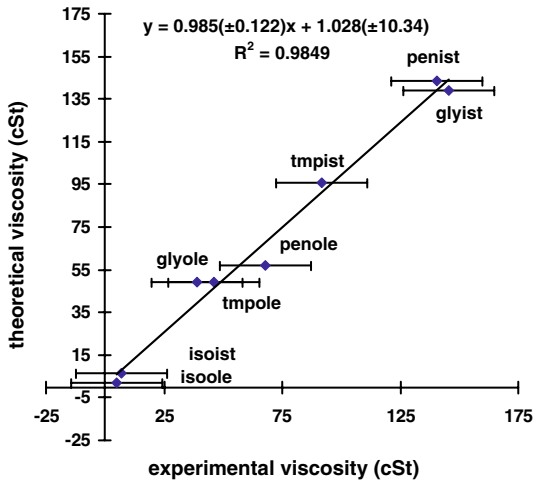

Fig. 5 Graphical representation of the best correlation obtained between the experimental and computed viscosity for the eight studied molecules. The slope and intercept are given with their confidence intervals at a confidence level of 95%

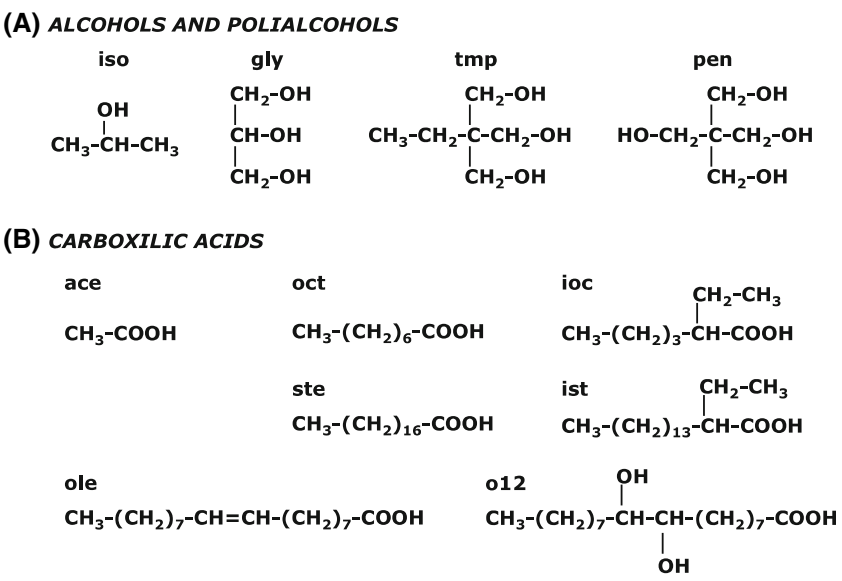

Fig. 6 Schematic representation of the residues used to build the 40 new molecules

the double bond carbons (Ol1) and the oleic acid with two ester groups, one in each double bond carbon (O12). These two new residues could be esterified with some of the following acids: acetic (Ace), octanoic (Oct), isooctanoic (Ioc), oleic (Ole), stearic (Ste) and isostearic (Ist). The corresponding names in Table 4 are formed by two or three sets of three letters, in which the first set indicates the alcohol (Iso, Gly, Tmp or Pen), the second the fatty acid that esterifies the alcohol (Ole, Ste, Ist, Ol1 and O12) and the third set (if there is one) the acid that esterifies the Ol1 or O12 acids (Ace, Oct, Ioc, Ole, Ste and Ist).

Table 4 contains the computed molecular properties and the predicted viscosity for the 40 designed molecules using the best correlation obtained (F in Table 3). It is worth noting here that only the computed viscosity for the isopropyl stearate is negative, and consequently should be discarded.

Principal Component Analysis

An exploratory analysis of the variables used for the multilinear regression was performed with a PCA. The model obtained considering three principal components explains 97.4% of the variance. A model with four components explains up to 99.5% of the variance; however, this is an overfitted model.

A careful analysis of the loading graph (including the experimental viscosity) for the two-first components (Fig. 7) clearly demonstrates that variables D and GBSOL are significantly different from the rest in the first component axis, and also shows that the radius of gyration, number of carbon atoms, molecular weight and dipole moment are quite correlated and are directly proportional to viscosity.

Stepwise Regression

A stepwise regression was performed with the aim of evaluating the most significant variables [15]. The computed data were standardized to ensure that the same variance was given to all the variables and to minimize the problems caused by using different units and value ranges. The analysis was performed with two different procedures: direct and indirect [16]. The two procedures converged to the same result:

$$\mu_{(\text{theo.})} = 96.2[\text{RGyr}] - 212.2[\text{Dip.Mom.}] + 12.5[\text{GBSOL}] - 385.6$$

The final result is thus a function that depends on only three variables: radius of gyration, dipole moment and solvation energy. The variable of radius of gyration is slightly out of the limit of significance allowed in this study (0.01). However, the straight line obtained when this variable is not considered is worse than when it is. Moreover, the confidence [17] on the prediction of viscosity when the previously described equation is used (see above) is $Y \pm 54.5$.

Partial Least Square

This is an alternative method for solving data in which the variables are highly correlated. The advantage of this type of regression is its high predictive capacity when the first components of the decomposition carried out are used.

A PLS model for eight observations was built with the most significant variables (radius of gyration, dipole moment and solvation energy). Two PLS-components were necessary to explain 99% of the variance. The loading coefficients of each component (PC1 and PC2) were as

Table 4 Computed molecular parameters (diffusion coefficient, radius of gyration, solvation energy, and dipole moment) used in the multilinear approach to calculate the macroscopic viscosity (cStokes)

	D ($Å^2$/ps) × 10^8	RadGyr (Å)	Dip.Mom. (Debye)	GBSOL (kcal/mol)	*T* (K)	No. C	MW (g/mol)	μ calc	
								Mean (cSt.)	S.Dev. (cSt.)
IsoOle	2.11	4.60	0.30	−0.11	298	21	324	1.8	0.2
IsoSte	1.65	6.19	0.31	0.24	298	21	326	−64.1	8.1
IsoIst	2.82	5.06	0.31	1.34	298	21	326	6.7	0.8
GlyOle	0.14	6.08	0.62	−3.14	298	57	884	49.4	6.2
GlySet	0.29	6.40	0.58	−1.10	298	57	890	76.7	9.7
GlyIst	0.41	6.32	0.78	0.89	298	57	890	139.0	17.5
GlyOl1Ace	0.15	6.27	1.13	−16.37	298	63	1,137	692.2	87.4
GlyOl1Oct	0.24	6.79	1.27	−10.80	298	81	1,390	688.9	87.0
GlyOl1Ioc	0.12	6.59	0.96	−10.91	298	81	1,390	642.3	81.1
GlyOl1Ole	0.27	7.45	1.28	−5.18	298	111	1,811	607.0	76.6
GlyOl1Ste	0.07	7.29	0.87	−5.57	298	111	1,811	535.7	67.6
GlyOl1Ist	0.15	7.45	1.18	−4.58	298	111	1,811	596.7	75.3
GlyO12Ace	0.10	6.61	0.93	−14.71	298	70	1,274	802.2	101.3
GlyO12Oct	0.13	7.40	1.71	−1.61	298	105	1,781	993.5	125.4
GlyO12Ioc	0.02	7.15	1.15	−11.73	298	105	1,781	783.3	98.9
GlyO12Ole	0.44	8.31	1.35	5.89	298	165	2,623	736.5	93.0
GlyO12Ste	0.32	8.77	1.26	7.37	298	165	2,623	715.2	90.3
GlyO12Ist	0.11	8.50	0.94	4.39	298	165	2,623	634.6	80.1
TmpOle	0.01	6.05	0.61	−2.16	298	60	926	49.2	6.2
TmpSte	0.18	6.37	0.55	0.00	298	60	932	72.7	9.2
TmpIst	0.17	6.36	0.57	1.59	298	60	932	96.0	12.1
TmpOl1Ace	0.13	6.41	1.13	−14.23	298	66	1,178	693.4	87.5
TmpOl1Oct	0.26	6.89	1.40	−8.60	298	84	1,432	722.2	91.2
TmpOl1Oc2	0.18	6.80	1.82	−11.55	298	84	1,432	768.3	97.0
TmpOl1Ole	0.58	7.53	1.04	−4.63	298	114	1,853	553.8	69.9
TmpOl1Ste	0.19	7.54	1.21	−3.82	298	114	1,853	595.1	75.1
TmpOl1Ist	0.07	7.59	1.28	−3.53	298	114	1,853	608.8	76.9
TmpO12Ace	0.11	6.56	0.86	−14.27	298	72	1,318	858.2	108.3
TmpO12Oct	0.18	7.52	1.03	−3.66	298	108	1,823	826.5	104.3
TmpO12Ioc	0.11	7.08	1.08	−6.98	298	108	1,823	818.0	103.3
TmpO12Ole	0.08	8.46	1.17	2.55	298	168	2,665	644.5	81.4
TmpO12Ste	0.11	8.40	1.48	7.18	298	168	2,665	759.5	95.9
TmpO12Ist	0.15	8.49	1.50	2.68	298	168	2,665	705.1	89.0
PenOle	0.13	6.59	0.57	−3.03	298	77	1,192	57.2	7.2
PenSte	0.29	6.83	0.62	−0.51	298	77	1,200	118.4	14.9
PenIst	0.08	6.31	0.45	1.83	298	77	1,200	143.4	18.1
PenOl1Ace	0.28	6.73	1.29	−19.05	298	85	1,530	949.4	119.9
PenOl1Oct	0.11	7.28	0.98	−12.17	298	109	1,867	854.5	107.9
PenOl1Ioc	0.06	7.20	0.92	−15.09	298	109	1,867	813.6	102.7
PenOl1Ole	0.08	7.96	1.14	−6.59	298	149	2,428	759.6	95.9
PenOl1Ste	0.12	8.03	1.44	−3.08	298	149	2,428	852.8	107.7
PenOl1Ist	0.02	8.26	1.43	−5.62	298	149	2,428	808.7	102.1
PenO12Ace	0.07	7.07	1.83	−20.45	298	93	1,715	1,303.8	164.6
PenO12Oct	0.12	7.81	1.46	−0.60	298	141	2,388	1,246.4	157.4
PenO12Ioc	0.17	8.12	1.38	−7.30	298	141	2,388	1,136.2	143.4
PenO12Ole	0.22	9.16	1.55	2.39	298	221	3,510	913.8	115.4
PenO12Ste	0.21	9.22	1.14	12.02	298	221	3,510	951.7	120.1
PenO12Ist	0.09	9.25	1.49	6.73	298	221	3,510	952.0	120.2

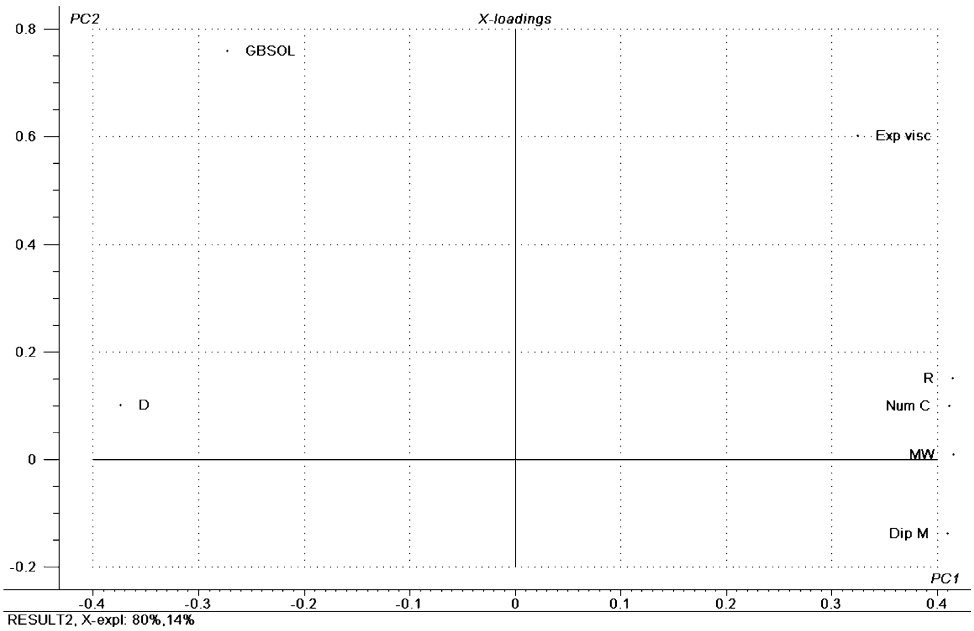

Fig. 7 Loading graph for the two main components

follows: radius of gyration = 0.686 and 0.233, dipole moment = 0.690 and −0.170 and solvation energy = −0.462 and 0.959.

There are two statistical parameters for evaluating the goodness of fit of a PLS model: R^2 (which is the variation explained by the model, and therefore is a measure of how well the model fits the data), and Q^2 (which is the variation of the training set predicted by the model according to cross validation, and therefore shows how well the model predicts new data). A large Q^2 ($Q^2 > 0.5$) indicates a good predictive capacity.

Our model has both a high R^2 and Q^2. The cumulative R^2 is 0.56, 0.86 and 0.88 for the first, second and third components respectively, while the cumulative Q^2 values are 0.51, 0.75 and 0.75 respectively. The tendency of Q^2 indicates that the PLS model with one more component would probably reduce the Q^2 parameter, which would make the model overfitted.

For easy comparison and interpretation of the models obtained in each of the procedures (Stepwise regression and Partial Least Square), the root mean square error of calibration (RMSEC) and the root mean square error of prediction (RMSEP) were calculated. The first one gives an idea of the error in the description of the data structure and the second one indicates the error in the predictive capacity. RMSEC was of 17.76 and 19.23 while RMSEP was of 31.86 and 31.02 for the stepwise regression and PLS, respectively. Hence similar results are obtained for the two models; the stepwise regression seems to explain the intrinsic structure of the data better, whereas the PLS model has a higher predictive capacity.

Conclusions

The predictive model, based exclusively on the direct use of the Einstein–Smoluchowsky equation, gave unsatisfactory results, most probably due to an inadequate computation of the diffusion coefficient derived from the conditions of the molecular dynamics simulations (in vacuo) while the experimental coefficient diffusion is obtained in very different conditions (pure liquid). In contrast, the least-square multilinear regression approach successfully reproduced the viscosity of a set of eight molecules, and gave correlation coefficients very close to the unity. Therefore, we used this approach to compute and predict the viscosity of a set of forty new and previously unknown molecules.

In spite of using a rather small number of experimental data (eight values), the principal component analysis performed on the computed results demonstrated that with only three variables, all of which were obtained from computations (radius of gyration, dipole moment and solvation energy), it is possible to explain more than 99% of the difference observed; using more variables allows the percentage to be increased up to 99.9%. Using a larger number of experimental results would have produced sounder results; however, there are many difficulties in obtaining good and reliable experimental results due to the lack of available compounds.

Acknowledgments Financial support from the "Ministerio de Industria, Turismo y Comercio" (Spain) through one of the components of the program CENIT (project PiIBE) is gratefully acknowledged.

References

1. Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J Comput Chem 21:1049–1074
2. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE III, De Bolt S, Ferguson D, Seibel G, Kollman PA (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput Phys Commun 91:1–41
3. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668–1688
4. Case DA, Pearlman DA, Caldwell JW, Cheatham TE III, Wang J, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV, Cheng AL, Vincent JJ, Crowley M, Tsui V, Gohlke H, Radmer RJ, Duan Y, Pitera J, Massova I, Seibel GL, Singh UC, Weiner PK, Kollman PA (2002) AMBER 7. University of California, San Francisco
5. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. J Comput Phys 23:327–341
6. Ryckaert JP (1985) Special geometrical constraints in the molecular dynamics of chain molecules. Mol Phys 55:549–556
7. MacroModel, version 9.0 (2005) Schrödinger, LLC, New York
8. Allinger NL (1989) Molecular mechanics: the MM3 force field for hydrocarbons. 1. J Am Chem Soc 111:8551–8566
9. Polak E, Ribiere G (1969) Note sur la convergence de directions conjugées. Revue Francaise Inf Rech Oper 3e Année 16:35–43
10. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA Jr, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Salvador P, Dannenberg JJ, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Andres JL, Gonzalez C, Head-Gordon M, Replogle ES, Pople JA (2001) Gaussian 98, revision A.11. Gaussian Inc., Pittsburgh
11. Singh UC, Kollman PA (1984) An approach to computing electrostatic charges for molecules. J Comput Chem 5:129–145
12. van Gunsteren WF, Berendsen HJC, Rullmann JAC (1981) Stochastic dynamics for molecules with constraints: Brownian dynamics of *n*-alkanes. Mol Phys 44:69–95
13. van Gunsteren WF, Berendsen HJC (1982) Algorithms for Brownian dynamics. Mol Phys 45:637–647
14. van Gunsteren WF, Berendsen HJC (1990) Gas chromatographic separation of enantiomers on cyclodextrin derivatives. Angew Chem Int Ed Engl 29:939–957
15. Hocking RR (1976) The analysis and selection of variables in linear regression. Biometrics 32:1–49
16. Rencher AC (2002) Methods of multivariate analysis, 2nd edn. Wiley, Wiley series in Probability and Statistics, New York, p 233
17. Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kauffman L (1988) Chemometrics: a textbook, 5th edn. Elsevier, Amsterdam, pp 185–188