ELSEVIER

Carbohydrate Research 314 (1998) 141–155

CARBOHYDRATE
RESEARCH

A comparison and chemometric analysis of several molecular mechanics force fields and parameter sets applied to carbohydrates☆

Serge Pérez [a,*], Anne Imberty [a], Soren B. Engelsen [b], Jan Gruza [a], Karim Mazeau [a], Jesus Jimenez-Barbero [c], Ana Poveda [d], Juan-Felix Espinosa [c], Bouke P. van Eyck [e], Glenn Johnson [f], Alfred D. French [f], Marie Louise C.E. Kouwijzer [g], Peter D.J. Grootenuis [g], Anna Bernardi [h], Laura Raimondi [h], Hanoch Senderowitz [i], Viviane Durier [j], Gérard Vergoten [j], Kjeld Rasmussen [k]

[a] *Centre de Recherches sur les Macromolécules Végétales (associated with Université Joseph Fourier, Grenoble.), CNRS, BP53X, F-38041 Grenoble, France*

[b] *Food Technology, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

[c] *Instituto de Quimica Organica, CSIC, Juan de la Cierva 3, E-28006 Madrid, Spain*

[d] *SIdI-UAM, E-28047 Cantoblanco, Madrid, Spain*

[e] *Department of Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, NL-3584 CH Utrecht, Netherlands*

[f] *USDA, Southern Regional Research Center, 1100 Robert E. Lee Blvd., New Orleans, LA 70124, USA*

[g] *G.B.B. Institute, University of Groningen, Njenborgh 4, NL-9747 AG Groningen, Netherlands*

[h] *Dipartimento di Chimica Organica e Industriale, Universita di Milano, via Golgi, I-19-20133 Milan, Italy*

[i] *Department of Chemistry, Columbia University, New York, NY 10027, USA*

[j] *CRESIMM, Université des Sciences et Techniques de Lille-Flandres-Artois, F-59655 Villeneuve d'Ascq, France*

[k] *Department of Chemistry, The Technical University of Denmark, DK-2800 Lyngby, Denmark*

Received 3 July 1998; accepted 12 November 1998

Abstract

Carbohydrates are thought to be especially difficult to model because of their highly polar functionality, their flexibility, and their differences in electronic arrangements that occur during conformational and configurational changes, such as the anomeric, exo-anomeric and gauche effects. These issues have been addressed in recent years, yielding several contributions to set up some relevant parameterizations that would account for these specific features of carbohydrates. Within the framework of a workshop involving the participation of 11 research groups active in the field, several commonly used molecular mechanics force fields and special carbohydrate parameter sets have been considered. The application of 20 force fields and/or sets of parameters to a series of seven test cases provided a fairly general picture of the potentiality of these parameter sets for giving a consistent image of structure and energy of carbohydrate molecules. The results derived from a chemometric analysis (principal component analysis, PCA) give a global view of the performances of the force fields and parameter sets for carbohydrates. The present analysis (i) provides an identification of the parameter sets which differ from the bulk, (ii) helps to establish the relationship that

☆ The workshop was organized in the framework of the 9th European Carbohydrate Symposium, Utrecht, Netherlands, 6–11 July, 1997.

* Corresponding author. Tel.: +33-476-03760330; fax: +33-476-03760329.

E-mail address: perez@cermav.cnrs.fr (S. Pérez)

PII: S0008-6215(98)00305-X

exists between the different parameter sets, (iii) provides indications for selecting different parameter sets to explore the force field dependency (or the lack of thereof) of a given molecular modeling study. Through the PCA, we have created a force field landscape on which the different force fields are related to each other on a relative scale. New carbohydrate force fields can easily be inserted into this landscape (PCA model) and related to the performance of existing force fields.

Keywords: Molecular mechanics; Force field; Carbohydrate; PCA

1. Introduction

The molecular mechanics method is becoming a standard tool for chemists and biologists. This method is a foundation for the plethora of molecular modeling software that is now available for a wide range of computers, including inexpensive desktop computers. The application of molecular mechanics to biological molecules and macromolecules has been more focused on solving problems related to the field of nucleic acids and peptides than to carbohydrates. The need for special treatment of carbohydrates follows from their densely packed, highly polar functionality, the dependence of their conformational behavior on stereoelectronic effects (anomeric, exo-anomeric and gauche effects) along with an enhanced conformational flexibility. These issues have been addressed in recent years, and several contributions have been made to set up some relevant parameterizations that would account for these specific features of carbohydrates.

The actual question in routine work in glycoscience is not so much how to perform the calculation, as what force field and/or parameterization to use. If you cannot solve your modeling problem with a simple plastic model, then the choice of the force field is important. Within the framework of a workshop involving the participation of 11 research groups active in the field, several commonly used molecular mechanics force fields and special carbohydrate parameter sets have been considered. Purposely, this exercise has been conducted with a fairly naive approach, trying to provide answers to fairly basic questions that a dedicated carbohydrate chemist or biologist would like to address without being a professional molecular modeler. To this end, the molecules in the test were chosen to reflect several levels of increasing complexity usually found in conformational analysis of carbohydrates. The desired molecular properties were held to simple static energetic and geometrical features which all modeling programs should be able to perform. A particular protocol was set up in order to minimize the bias resulting from environment and conformational search methods.

From the selection of test cases, i.e., molecules and molecular properties, and in the context of conditions of applications of the workshop, it is practically impossible to give an objective appreciation of the suitability and quality of the force fields. Theoretically, the experiment would be the golden standard for the comparison with the computational results. However, the lack of trusted experimental data with errors less than the differences between the simulated data is a major obstacle. Instead we have decided to cluster the force field performances on a completely objective but relative ground using a chemometric approach. It provides a wide assessment of the self-consistency (or the lack thereof) for the structures and energetics of carbohydrates derived from the application of molecular mechanics calculations. The present paper describes a force field landscape on which the different force fields are related to each other on a relative scale. New carbohydrate force fields can easily be inserted into this landscape and related to the performance of existing force fields and carbohydrate parameterizations.

2. Methods

Strategy.—Molecular mechanics correspond to a set of empirical methods for calculation of molecular structures and properties using atoms as the elementary constituents. It involves the development of a force field in which molecular systems can be fully relaxed (energy minimization or geometry optimization) or in which they can move according to classical motional equations (molecular dynamics). Two choices have to be made when developing a molecular mechanics force field: (1) what should be the functional form of the potential energy functions (e.g., the use of either a Morse potential or a simple harmonic term for bonded interactions); (2) what should be the numerical values of functional parameters in the potential energy functions (e.g., the magnitude of the force constant of the harmonic term for bonded interactions). The choice of the functional form of the potential energy functions is a qualitative choice which may strongly affect both the precision and efficiency of the designed force field. The choice of numerical values of functional parameters is a quantitative one, which will mainly affect the precision of the force field, although convergence behavior during energy minimizations may change quite drastically. The latter quantitative choice can either be made by estimating the parameters from the literature or by optimizing them in a systematic manner, e.g., in the least-squares sense by comparison with experimental data and/or those derived from ab initio calculations on relevant model compounds.

Although we will mention pertinent experimental results along the way, we wish to retain our focus on the main question of this report: is the selection of force field an important choice in a modeling study? Therefore, we will mostly compare the results among themselves. The experimental evidence available for the atomic structures and energies of carbohydrates is primarily geometries and crystal packing in solid phase, anomeric equilibria in solution, translational and rotational diffusion in solution, and molecular vibrations in condensed phase. All of these experimental properties can be modeled, but they require consideration of a number of choices for modeling the environment and for sampling statistically representative ensembles of conformers contributing to the experimental evidence. Not all modeling software offers the same approaches for consideration of the environment or statistical sampling, and the influence of the choices will remain unknown. For example, what is the influence of using Ewald summation instead of simply constructing a mini-crystal when simulating solid phase?

A common confusion in the application of molecular mechanics methods regards the various types of energies. The vast majority of experimental determinations of energy differences yield Gibbs free energies. On the other hand, the potential energies calculated by a molecular mechanics program are most comparable with enthalpy. Some modeling software can calculate free energies, either by a normal mode analysis or from extended molecular dynamics simulations. All of these methods require a complete description of the potential energy hypersurface besides the reliable values of the steric energies for local minima that are reported herein.

Therefore, we have decided to investigate the performances of the force fields on an objective but relative basis. The only program effect we had to accept is the different implementations of energy minimization algorithms and convergence criteria. Although we are aware that not all software packages encourage the use of gradient-based convergence criterion, we judged these effects to be minor. Apart from this, the only bias in the comparison is in the selection of molecules and molecular properties.

Descriptions of selected molecules.—As the modeling of environment and conformational search methods differ from computer program to computer program, we wanted to avoid this bias and supplied each contributor with Cartesian coordinates of all the molecular test systems to which the contributors only had to apply their preferred carbohydrate force field for energy minimization and calculation of the desired molecular properties.

A set of seven molecules has been selected for this study. The choice was based on the conformational properties of each molecule. Fig. 1 gathers graphical representations of most of the molecules investigated here.

Methyl 5-deoxy-β-D-xylofuranoside. Two conformations corresponding to opposite pucker on the pseudorotational wheel, i.e., the envelope shape ^3E for the North pseudorotamer (Fig. 1A) and $_3E$ for the South one (Fig. 1B), were constructed.

Methyl β- and α-D-glucopyranosides. The two anomers of methyl glucopyranoside (Fig. 1C and D) were generated in the $^4C_1$ ring conformation.

Methyl α-D-glucopyranoside and α-D-galactopyranoside (Fig. 1E). The three staggered conformations of the hydroxymethyl group were generated for the two epimers in the $^4C_1$ ring conformation. These three orientations are referred as to gauche–gauche (gg), gauche–trans (gt) and trans–gauche (tg) values of − 60, + 60 and 180° for the O-5–C-5–C-6–O-6 torsion angles, respectively [1].

α-D-Glucopyranose·H_2O. The three-dimensional structure of glucose monohydrate (Fig. 1F) as found in the crystal structure [2] was provided together with the same molecule with the water molecule 50 Å away.

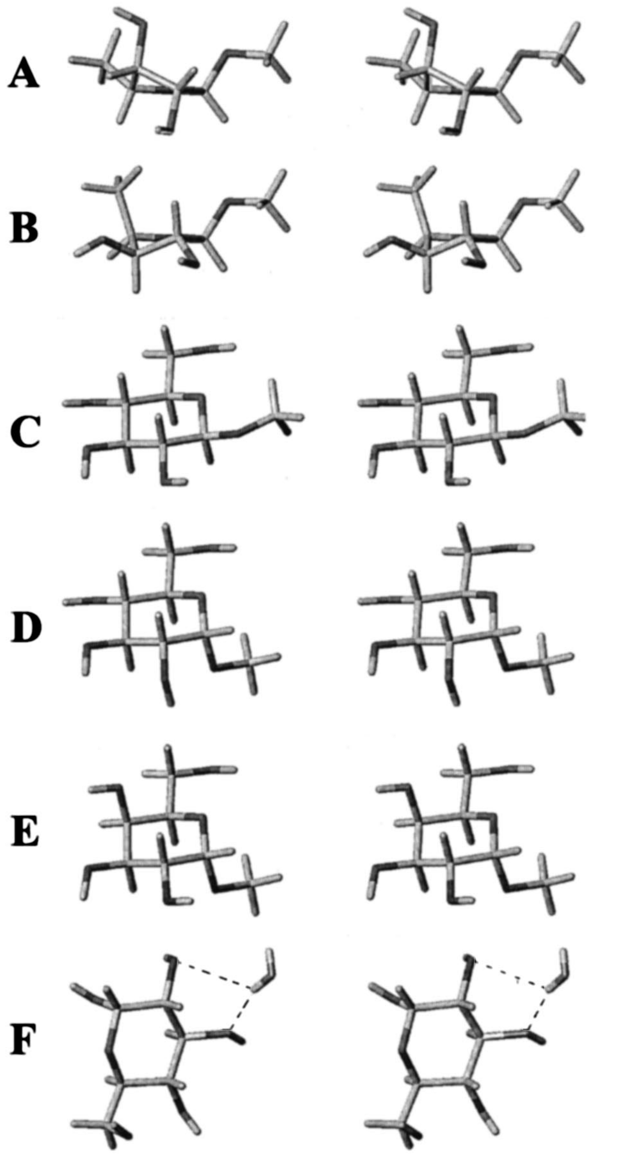

Fig. 1. 3D stereodrawings of some of the molecules used in the present study.

Disaccharides. Three low energy conformations of the α-D-Man-(1 → 3)-β-D-Man disaccharide arising from different orientations of the two torsion angles of the glycosidic linkage were constructed: ($\Phi = 75°$, $\Psi = 180°$), ($\Phi = 70°$, $\Psi = 100°$) and ($\Phi = 80°$, $\Psi =$ -50°) with Φ and Ψ defined as the dihedral angles O-5′–C-1′–O-1′–C-3 and C-1′–O-1′–C-3–C-4, respectively. The same procedure was applied for two conformations of the β-D-GlcNAc-(1 → 2)-β-D-Man disaccharide, ($\Phi = -70°$, $\Psi = 150°$) and ($\Phi = 90°$, $\Psi = 160°$), and three conformations of the Fuc(α1–2)Galβ disaccharide, ($\Phi = -80°$, $\Psi = 180°$), ($\Phi = -70°$, $\Psi = -120°$) and ($\Phi = -90°$, $\Psi = 50°$). For these two disaccharides, Φ and Ψ were defined as O-5′–C-1′–O-1′–C-2 and C-1′–O-1′–C-2–C-3, respectively.

Force fields included in the test.—Ten different software packages were considered in the present investigation. For some of them, different implementations (i.e., AMBER and AMBER*), different sets of parameters or different solvation models were used, leading to 20 different force fields. The list of these 20 combinations is given in Table 1.

AMBER. The 94s force field [3] of the AMBER molecular modeling program [4] was given the code number FF-12. This force field was augmented for carbohydrates by several authors. One parameter set was designed by Homans [5]. This parameterization adds the CHARMm-like molecular mechanics potential for carbohydrates developed by Ha et al. [6] to include the glycosidic linkage based on crystallographic data of pyranose systems and ab initio calculations on dimethoxymethane. We assigned this force field the code number FF-10 and ran it with a dielectric constant of 4.0. Glennon et al. [7] presented an AMBER-based force field especially for monosaccharides and (1 → 4)-linked polysaccharides (FF-8). Woods et al. [8] developed the GLYCAM parameter set for molecular dynamics simulations of glycoproteins and oligosaccharides (FF-1). FF-1 and FF-8 were run with a dielectric constant of 4. The AMBER* version of MacroModel 5.5 [9] was used together with the all-atom pyranose force field [10]. Calculations were run in vacuum (FF6), but also using the GB/SA water solvation model [11] (FF14).

Table 1
List of force fields and parameterization

	Force Field	Internet adress	Ref.
FF1	AMBER-GLYCAM 4.1	http://www.amber.ucsf.edu/amber/ff94_glycam.html	[8]
FF2	CHARMM-22	http://yuri.harvard.edu/charmm/charmm.html	[13]
FF3	TRIPOS-PIM	http://www.tripos.com/ and imberty@cermav.cnrs.fr	[17][a]
FF4	SPACIBA	vergoten@choia.univ-lille1.fr	[12]
FF5	MM3(94)/(96)	http://www.tripos.com/products/mm3.html	[a]
FF6	AMBER* (vacuum)	http://www.columbia.edu/cu/chemistry/mmod/mmod.html	[b]
FF7	CHEAT-95	http://www.msi.com/support/quanta/cheat95.html	[14]
FF8	AMBER-Glennon		[7]
FF9	MM3(92)		[19,20]
FF10	AMBER-Homans		[5]
FF11	GROMOS-87	http://igc.ethz.ch/gromos/ (for GROMOS 96 only)	[16]
FF12	AMBER-94	http://www.amber.ucsf.edu/amber/ff94.html	[3]
FF13	Biosym CVFF-92		[c]
FF14	AMBER* (GB/SA water)	http://www.columbia.edu/cu/chemistry/mmod/mmod.html	[b]
FF15	MM3* ($\varepsilon = 1$)	http://www.columbia.edu/cu/chemistry/mmod/mmod.html	[b]
FF16	MM3* ($\varepsilon = 80$)	http://www.columbia.edu/cu/chemistry/mmod/mmod.html	[b]
FF17	CFF/PEF95SAC	http://struktur.kemi.dtu.dk/cff/cffhome.html	[21]
FF18	Biosym CVFF_95_01.01		[b]
FF19	Biosym CFF	http://www.msi.com/solution/products/InsightII/index.html	[b]
FF20	Cerius-Dreiding	http://www.msi.com/solution/products/Cerius2/index.html	[24]

[a] Tripos Associates.
[b] Biosym Technologies/MSI software.
[c] The Macromodel package.

SPACIBA. This program has been developed as a vibrational molecular force field with a particular emphasis on monosaccharides and oligosaccharides [12]. The potential energy functions use a modified Urey–Bradley–Shimanouchi force field. In the present investigation a dielectric constant of 1 was used. This force field was assigned the code number FF-4.

CHARMm. The general-purpose force field [13] CHARMm-22 has been given the code number FF-2. CHEAT95 [14] is a CHARMm-based force field in which the carbohydrate hydroxyl groups are represented by extended atoms. This hydroxyl atom type is parameterized in such a way that a simulation of an isolated molecule mimicks the molecule in the aqueous solution. Therefore, in the present study, this force field, which has been given code number FF-7, was not used for modeling either for O-methyl groups, or for energy evaluation when water molecules are explicitly present. As for disaccharides, the methyl group attached to the anomeric oxygen in the selected molecules was removed.

GROMOS. The present investigation uses the GROMOS-87 [15] force field [16], which was assigned the code number FF-11. GROMOS was developed for MD simulations of solutions rather than for energy minimization of isolated molecules. Some parameters, especially charges, were not present in the original formulation and had to be estimated. It should be noted that a more recent version of the program (GROMOS-96) has not been tested in the present investigation.

TRIPOS. The original Tripos 5.3 (Tripos Associates) force field [17] has been complemented with a set of parameters (Tripos-PIM) for modeling carbohydrates and protein–carbohydrate interactions [18]. This parameterization was assigned the code number FF-3.

MM3. The MM3(92) force field (code number FF-9) represents a highly detailed type of force field notably including anomeric and exo-anomeric effects [19,20]. Two updated versions of MM3 appeared in 1994 and 1996 (Tripos Associates). The most recent one was given code number FF-5 in the present study. Both MM3(92) and MM3(96) programs were run using a dielectric constant of 4. The

MacroModel package contains the MM3* force field, which is described as a modification of the original MM3 force field available in the public domain. This force field has been used in the present study together with a dielectric constant of 1 (FF-15) and 80 (FF-16).

CFF/PEF95SAC. This force field (code number FF-17) is a standard type of molecular mechanics force field which uses the Morse potential to model covalent bonds, and includes an anomeric carbon atom type. CFF/PEF95SAC [21] has been developed based on the consistent force field approach (CFF) [22,23] using least-squares fitting of potential energy parameters on experimental data like internal coordinates, nonbonded distances, dipole moments, lattice energies, unit cell dimensions, and vibrational frequencies. Environmental variables have been explicitly taken into account using Ewald summation during the development of this force field.

CVFF and CFF. These two force fields are available within the INSIGHT/DISCOVER software (Biosym Technologies). CVFF uses a Morse function for bond stretching, and cross terms are included in the potential functions. Two versions were used, respectively, CFFV-92 (code number FF-13) and CVFF-95_0_1.01 (code number FF-18). CFF has been developed as a highly non-diagonal force field (code number FF-19). These two force fields share the ability to use general parameters when no explicit parameters are available. Therefore, no particular attention is given to carbohydrate molecules.

DREIDING. Under the code FF-20, we used the DREIDING 2.21 force field [24] available in Cerius^2 (Molecular Simulations). DREIDING 2 is meant to be used for general organic and main group compounds, and no special parameterization is available for carbohydrates.

Comparison protocol.—The force field comparison protocol was designed as follows:

- The test cases should cover a wide range of structural and energetic features that molecular modeling should be able to simulate.
- The test cases were converted into test molecules. All calculations were run on the same files of coordinates. They were generated at CERMAV (Grenoble, France) and stored in an anonymous ftp directory with free reading access.
- The geometry of each structure was fully optimized by energy minimization with all 20 force fields. The optimizations were carried out without any explicit conformational searching.
- A spreadsheet file (Excel, Microsoft, USA) was also made available with the final results (energy, bond distances, valence angles, dihedral angles, etc.) to be inserted.

Principal component analysis.—The PCA method [25] offers a simple and efficient way of describing the systematic variations in a complex data structure. It uses a two-dimensional data evaluation strategy in which many variables are registered (molecular geometry and energetics, in the present case) on the same objects (force fields, in the present study). The multidimensional data set is resolved into orthogonal components whose linear combinations approximate the original data set in a least-squares sense (systematic part) (Fig. 2). PCA gives a representation, on a graphic interface, of the main variance of the complex data structure (including missing data) by projecting the data onto a few orthogonal directions (principal components). This provides a way to analyze the performances of the force fields as a whole, initially with the sacrifice of understanding on a more detailed level but with the capability of providing interesting global information which later feeds back into low-level knowledge.

In PCA, the original data matrix (**X**) ($ff \times v$) (number of force fields times number of variables/calculated observables) is decomposed into a score matrix (**T**) ($ff \times pc$) (number of force fields times number of principal components) and a loading matrix (**P**) ($pc \times v$), and the residuals are collected in a matrix (**E**) ($ff \times v$): $\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$. Only a limited number of principal components (PCs) are in most cases relevant in describing the systematic information in **X**. The loading vectors for the principal components can be considered as hidden data structures that are common to all the objects. What makes the original data

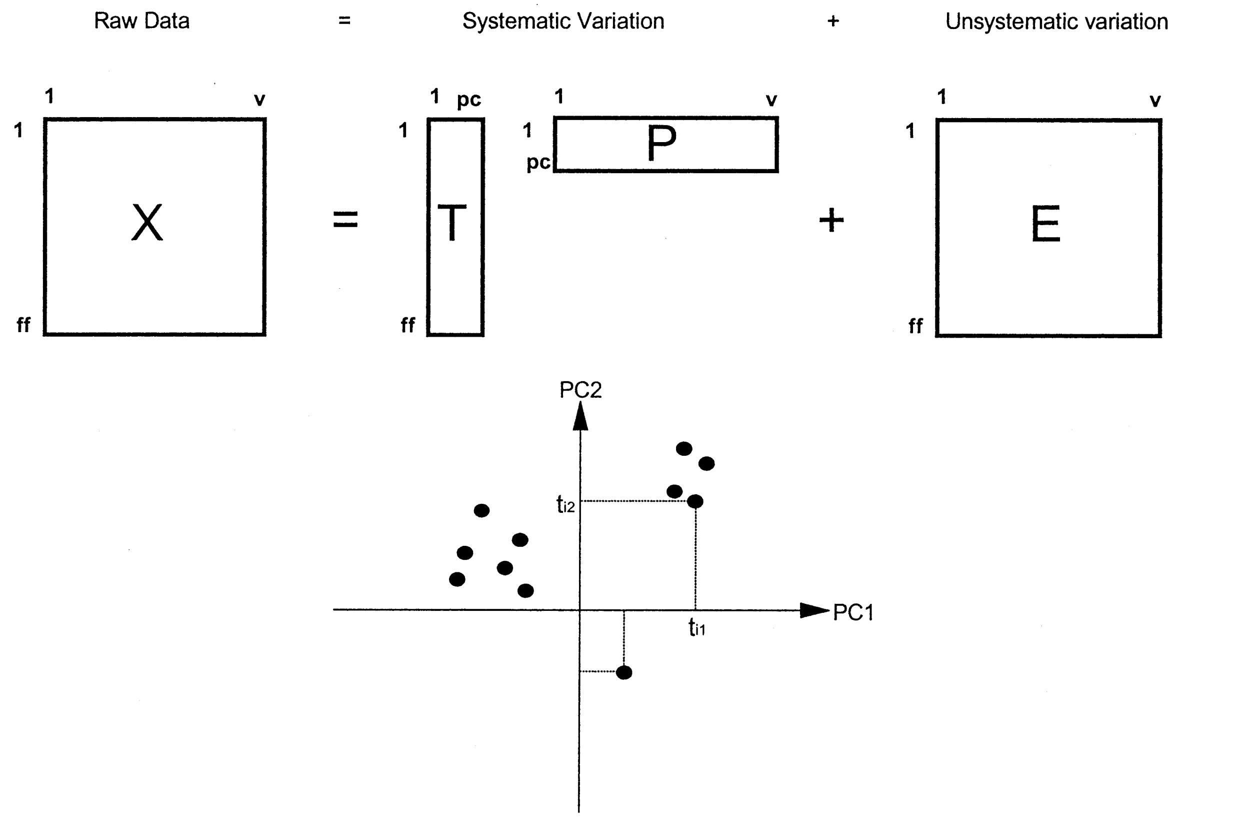

Fig. 2. Schematic PCA figure. (Top) The resolving of the original data set (**X**) into a systematic varying part described as score vectors (**T**) and loading vectors (**P**) and a residual noise matrix (**E**). (Bottom) The score vectors that describe the individual variation of the objects to be analysed can then be plotted against each other in scatter plots describing the main variation in the data set (**X**).

structures different is the amount (scores) of hidden data structures. The scores contain information about the objects and the loadings about the variables.

Prior to PCA, the data in the spreadsheet were pre-PCA transformed in two different steps: (i) all dihedral angles with mean value in the range − 120 to − 180° or in the range 120–180° were standardized; i.e., if a given dihedral variable in most force fields was calculated in the range 120–180°, then the few dihedral angles lying in the range − 120 to − 180° were converted to lie in the range 180–240° to avoid nonquantitative discontinuities in the data set; and (ii) all variables were divided by the standard deviation and the mean value subtracted in order to keep the variables and their variations in the same range. All PCA calculations were made by the program Unscrambler version 6.11b (CAMO ASA, Trondheim, Norway).

Cluster analysis.—Cluster analysis (see [26]) is generally used to classify a set of objects (e.g., force fields) into groups with similar properties (e.g., calculates molecular features); for example, if we desire the entire data set to be separated into the most possible homogeneous subsets (clusters). The basis for the cluster algorithm is a similarity/dissimilarity measure between the objects which can be considered as a generalized distance measure. In this work a hierarchical k-mean method based on Euclidian distances is used. In that method, the object pairs that are most similar are first linked together and in subsequent calculations replaced by their centroid. These linked objects are then successively extended to larger object groups by inclusion of new objects which have the highest similarity with the already selected objects. The end product is a tree structure (dendrogram) that reflects the similarity (heritage) between more and more homogeneous clusters. In contrast to PCA, the inclusion of a new object in the

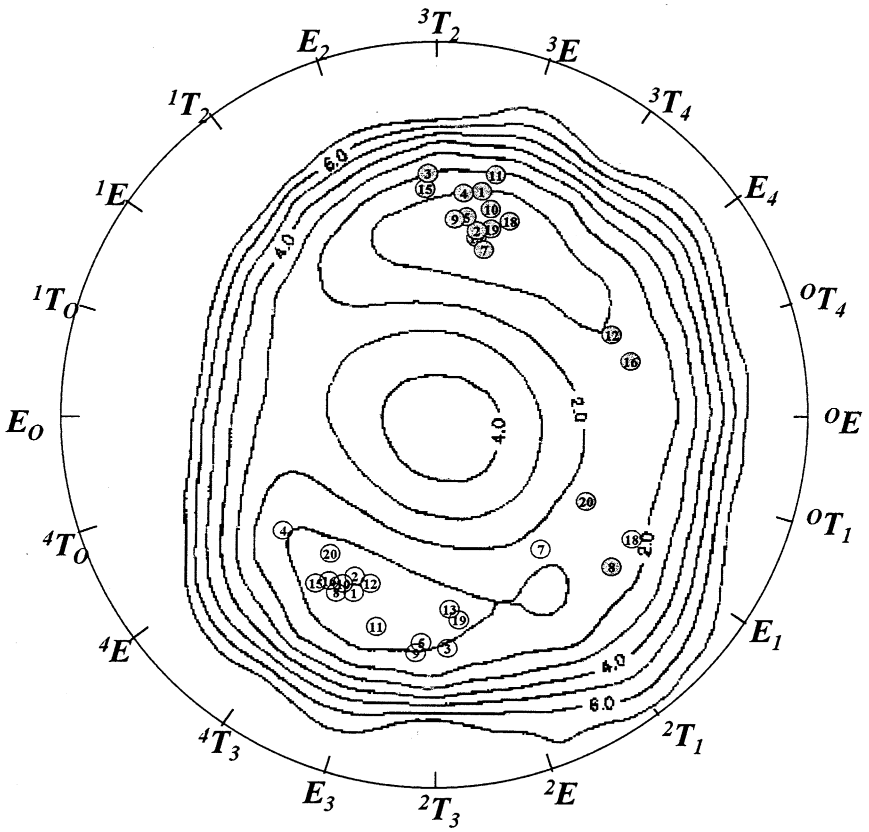

Fig. 3. Dispersion in the Cremer–Pople parameter space for methyl 5-deoxy-β-D-xylofuranoside. The puckering parameters observed after optimization by each force field (see label) are superimposed on the isoenergy map calculated using the MM3 software.

cluster analysis will require a complete recalculation.

Cluster analysis calculations were performed using Matlab ver. 5.1 (The Mathworks, Natick, MA) installed with the PLS_Toolbox ver. 1.53b (Eigenvector Research, Manson, WA).

3. Results and discussion

The present analysis is based on the assumptions that all participants have used the programs and concomitant force fields correctly and subsequently inserted the calculated number correctly in the spreadsheet. The analysis is furthermore based on the assumption that the 20 different force fields can be evaluated on the basis of 78 molecular quantities (12 energies, 4 bond lengths, 12 valence angles, 4 puckering parameters and 46 dihedral angles) contributing with equal weight.

Description of conformations and energies.— Many energetic and geometrical data have been collected after energy minimization of all tested molecules. All the data have been gathered in a table, which is available as supplementary material from the Corresponding Author. A short description of the data together with an evaluation of the differences as a function of the force field is given here.

Puckering parameters and energy difference between the North and South conformers of the methyl 5-deoxy-β-D-xylofuranoside. The first objective was to address the possibility of assessing the location and the relative energy of the two most stable conformers of a simple furanose. General-purpose force fields such as MM3, AMBER or CHARMM should, in theory, handle five-membered rings. On the other hand, none of the force field add-ons for carbohydrates were designed to handle furanose rings. In Fig. 3, the Cremer and Pople puckering parameters [27] of the optimized conformers have been reported on the pseu-

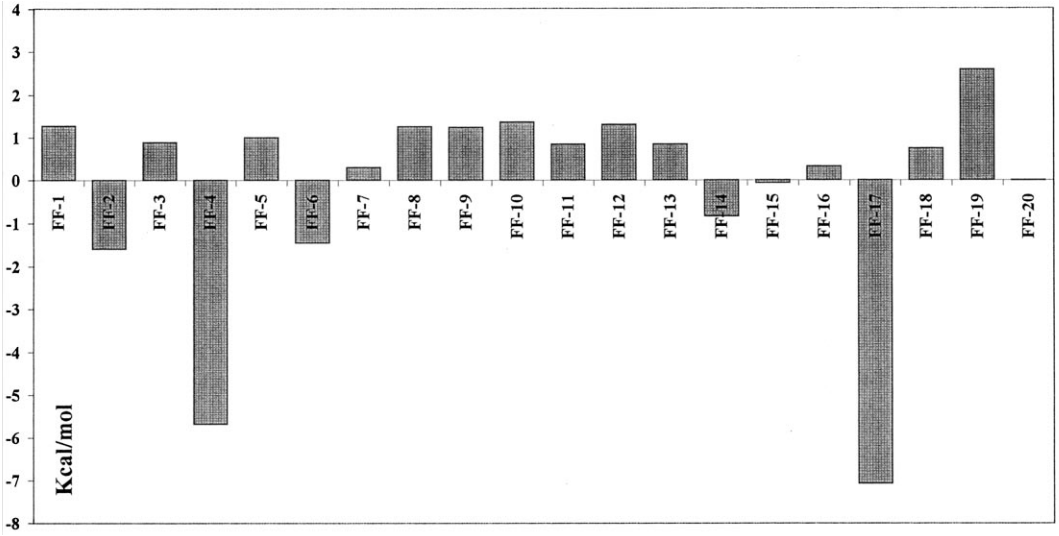

Fig. 4. Energy differences between methyl α- and β-D-glucopyranoside.

dorotational wheel, and these are superimposed on an MM3 energy map (Gruza and Imberty, unpublished results) using a procedure previously developed for conformational analysis of furanose rings [28]. In general, the refined conformers remained in the low-energy regions corresponding to the North and South minima, with the exception of the North conformer that converged to the South region when minimized using Dreiding (FF-20) or AMBER-Glennon (FF-8). It has been demonstrated recently that AMBER-native gives very low-energy barriers along the pseudorotational wheel of ribofuranoside [29], whereas the carbohydrate-modified parameters gave deeper minima. From NMR experiments (Gruza et al., personal communication) it is indicated that the Gibbs free energy difference should be small between the two populations, with a slight preference for the North conformer. Very scattered values are obtained in the present study, ranging from 4.0 kcal/mol preference for the North conformer by FF-2 (CHARM22) to 1.4 kcal/mol preference for the South one by FF-3 (Tripos-PIM).

Energy differences between the two anomers of methyl D*-glucopyranoside as well as distances and valence and torsion angles close to the anomeric centers*. The calculation of the geometry, conformation, and energy occurring as a function of the anomeric configuration of methyl glucopyranoside is obviously a key question of interest to any carbohydrate chemist. The experimental anomeric ratio of methyl glucopyranoside was determined by equilibrating the system in methanol using catalytic acid at 35 °C; it corresponds to a preference for the α anomer by 0.42 kcal/mol [30]. Fig. 4 depicts the energy preference between the two anomers of methyl glucoside as a function of the force field used for the calculations. In general, the force fields considered here fail to predict this anomeric effect, since most of them give a small energy preference for the β anomer. Two force fields gave unreasonably high preference for the α anomer, and only CHARM-22 (FF-2) and both AMBER versions for MacroModel (FF-6 and FF-14) yield energy differences close to the experimental value. AMBER*, together with the GB/SA solvation model (FF-14), gives the best fit with the experimental result. This pyranose parameterization has already been shown to reproduce anomeric free energies [10]. The structural aspect of the anomeric effect, i.e., the shorter C-1–O-1 bond length and smaller O-5–C-1–O-1 valence angle, observed for the β anomer (mean experimental values 1.385 Å and 107.6° for an equatorial anomeric group) when compared with the α anomer (1.398 Å and 111.9° for an axial anomeric group) [31], is better reproduced by force fields specially parameterized for carbohydrates such as FF-1 and FF-3, whereas the general DREIDING force field (FF-20) gives quite high values for the glycosidic bond

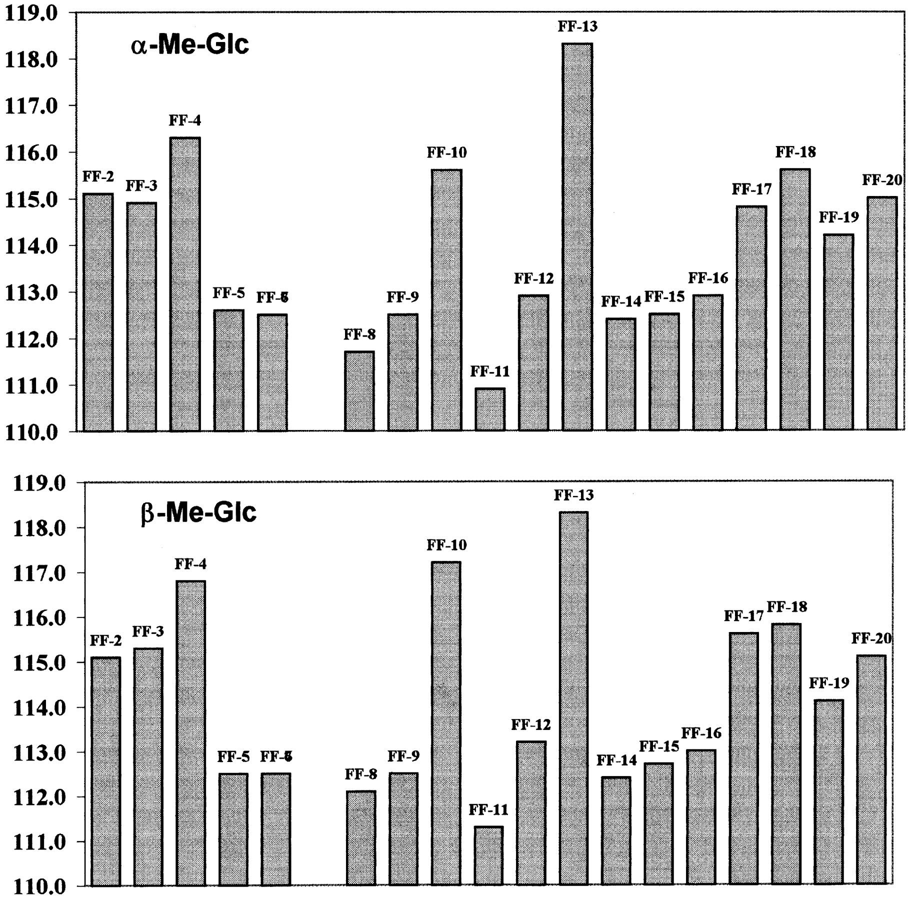

Fig. 5. Variations of the valence angle at the anomeric oxygen for methyl α- and β-D-glucopyranoside.

lengths in both anomers (1.45 Å). Variations of the glycosidic valence angles are displayed in Fig. 5. When compared with the mean experimental values of 113.8° for an axial linkage and 113.1° for an equatorial linkage [31], most of the force fields are in close agreement. However, some strong deviations are observed. In particular, GROMOS (FF-11) yields valence angles that are too small and CFF/PEF95SAC gives valence angles that are too large.

Energy difference and torsion angle for the three rotamers for the hydroxymethyl group of methyl α-D-glucopyranoside and methyl α-D-galactopyranoside. The population ratio of the rotational conformers around the C-5–C-6 torsion is an important issue, as it may govern some intermolecular interactions. From NMR experiments in water, the rotamer population was estimated to present a gg:gt:tg ratio of 57:38:5 for methyl α-D-glucopyranoside and 14:47:39 for methyl α-D-galactopyranoside [32]; however, there are a number of approximations in the analysis of the NMR data. The conformational preference of the hydroxymethyl group was evaluated for both the methyl α-D-glucopyranoside and methyl α-D-galactopyranoside. The results are in very poor agreement with the experimental data: for Me α-D-Glc, none of the calculations could reproduce the observed range of population, namely, gg > gt > tg. In fact, the order is generally reversed, with the gg rotamer displaying the highest energy. Only the CFF-based force fields (FF-17 and FF-19) are able to predict the gg rotamer as the preferred one. The agreement is not better when looking at Me α-D-Gal. In this latter case, only TRIPOS-PIM and CHEAT-95 predicted the correct gt > tg > gg population ranking. Such a poor agreement is not surprising and should not be considered as a drawback of the parameterization, since it has been demonstrated that the hydroxymethyl conformational behavior is very dependent on the solvation model [33].

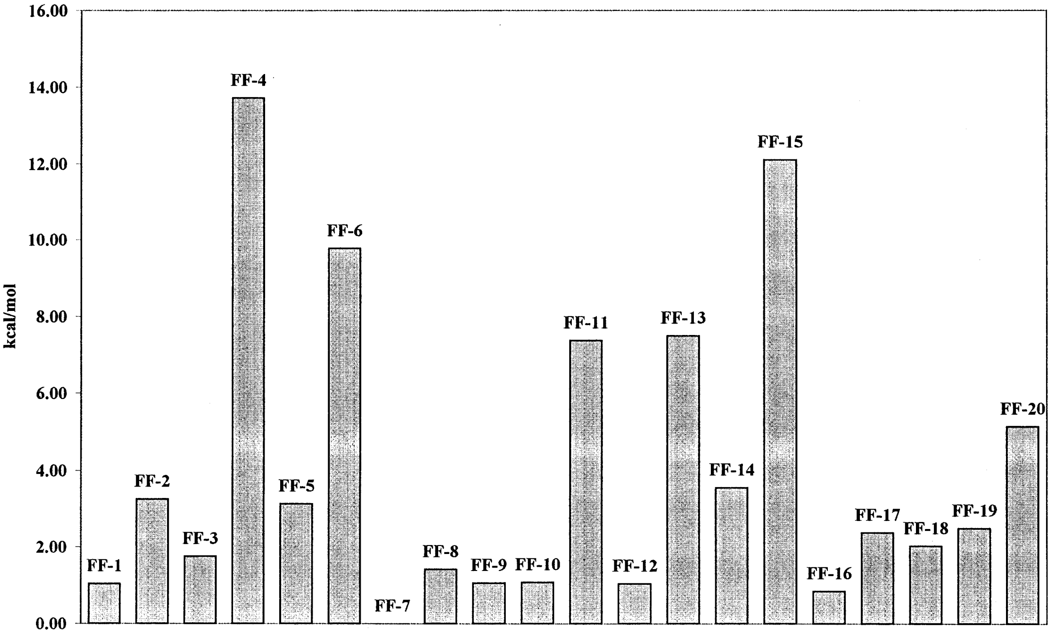

Fig. 6. Interaction energy of α-D-glucopyranose with one water molecule.

Interaction energy of glucose with one water molecule. Assessing the energy of hydration of a simple monosaccharide such as α-D-glucopyranose may appear to be a trivial question. Therefore, the monohydrate adduct of α-D-glucopyranose was extracted from the X-ray crystal structure [2]. As shown in Fig. 1, the presence of the water molecule allows the establishment of a three-centered hydrogen-bonding scheme. The energies calculated for this interaction are displayed in Fig. 6. They range from 0.85 (MM3* with $\varepsilon = 80$) to 13.7 kcal/mol (SPACIBA). It may be noted that, as expected, inclusion of the CB/SA water solvation model in the AMBER* calculations results in a decrease of the calculated interaction energy, even though its value is still above the one calculated by all the other AMBER-based force fields (FF-1, FF-8, FF-10 and FF-12). The same effect, but a lot more dramatic, is obtained in the MM3* calculations by increasing the dielectric constant from 1 (FF-15) to 80 (FF-16).

Relative energies and glycosidic torsion angles for several conformations of the α-D-Man-(1 → 3)-D-Man, β-D-GlcNAc-(1 → 2)-D-Man, and α-L-Fuc-(1 → 2)-D-Glc disaccharides. One would also like to estimate the low-energy conformers of disaccharides for which the global shape depends mainly on rotations about the glycosidic linkages, because the flexibility of the pyranose ring is rather limited and the different orientations of the pendent groups usually have a limited influence on the conformational space of the disaccharide. The α-D-Man-(1 → 3)-β-D-Man and β-D-Glc-(1 → 2)-β-D-Man disaccharides are part of the N-glycan moiety of glycoproteins. They have been the subject of many theoretical and experimental conformational studies. N-glycan oligosaccharides have been successfully co-crystallized with lectins, so several conformations of these oligosaccharides have been observed in the solid state [34]. It has therefore been demonstrated that the α-D-Man-(1 → 3)-β-D-Man disaccharide can adopt three different conformations, displaying large differences in the Ψ angle, whereas the β-D-Glc-(1 → 2)-β-D-Man disaccharide can adopt two different conformations, with 120° difference in Φ. All these conformations have been considered in the present investigation. The α-L-Fuc-(1 → 2)-D-Gal disaccharide was also proposed in order to evaluate the capacity of the force fields for handling L-sugars. The results of the energy minimizations are presented in Fig. 7, superimposed on the MM3 energy maps that have been demonstrated to be able to predict the conformations observed in the crystal structures for these two disac-

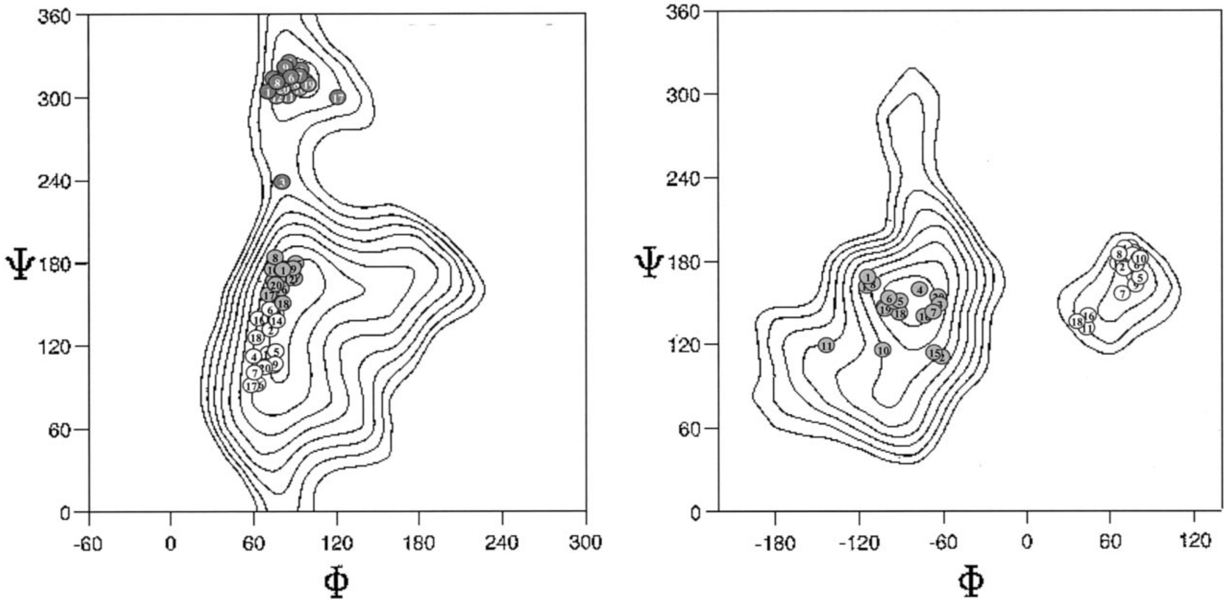

Fig. 7. Repartititon in the Φ and Ψ space for α-D-Man(1 → 3)-β-D-Man (left) and β-D-Glc-(1 → 2)-β-D-Man (right) disaccharides. The conformations optimized by each force field (see labels) are superimposed on the energy map calculated with the MM3 software.

charides [34]. In general, the conformers remain in their energy wells, except for the α-D-Man-(1 → 3)-β-D-Man disaccharide where a large conformational plateau joins two of the considered conformations. In several cases one conformer converged to the other one. In general, the main energy well was predicted to have the lowest energy. For the β-D-Glc-(1 → 2)-β-D-Man disaccharide, all the conformers stayed in their energy wells and the minimum governed by exo-anomeric effect (Φ about $-80°$) contained in all cases the conformers with lowest energy. The discrepancies appear in the energy difference between the two conformers, which ranges from 1.2 kcal/mol in AMBER-Glennon to 14 kcal/mol in AMBER*.

Global PCA.—To find and investigate the main directions of variations in our force field spreadsheet, a PCA was performed on all the data. The score plot in Fig. 8a displays the main dispersion of the force field performances as a function of PC1 and PC2. The plot has only projected down 30% of the total variation in the data, but it sums up the most important features. The plot indicates that FF-11 and FF-13 are extreme force fields which differ in each of the two main variation directions from the bulk of the force fields. Their positions are a reflection of the fact that the two force fields have the largest variations from the mean values of the variables and indicate that especially FF-11 is a quite different force field, as it deviates strongly from the other force fields in the main variation direction (PC1). It is also apparent from the score plot that the two different versions of MM3 (FF-5 and FF-9) perform practically in an identical manner (they are completely superimposed in the score plot, Fig. 8a). More interestingly, the score plot reveals that the Macromodel parameterization of MM3 labeled MM3* (FF-15 and FF-16), where the original bond dipoles have been changed to atomic charges, is changing the force field performances significantly from that of the original MM3 force field. This is indicated by the large gap between FF-5/FF-9 and FF-15/FF-16. Finally, the score plot reveals that all the AMBER-based force fields (FF-1/FF-6/FF-8/FF-10/FF-12 and FF-14) are grouping together near the PCA model center. This is further evidenced in Fig. 8b, which displays the main dispersion of the force field performances as a function of PC2 and PC3. In this plot all the AMBER-inherited force fields (plus FF-4 which is also AMBER based) cluster in a group indicated by an ellipse. Such clustering indicates that re-parameterization with special emphasis on carbohydrates does not move the force field's performance far

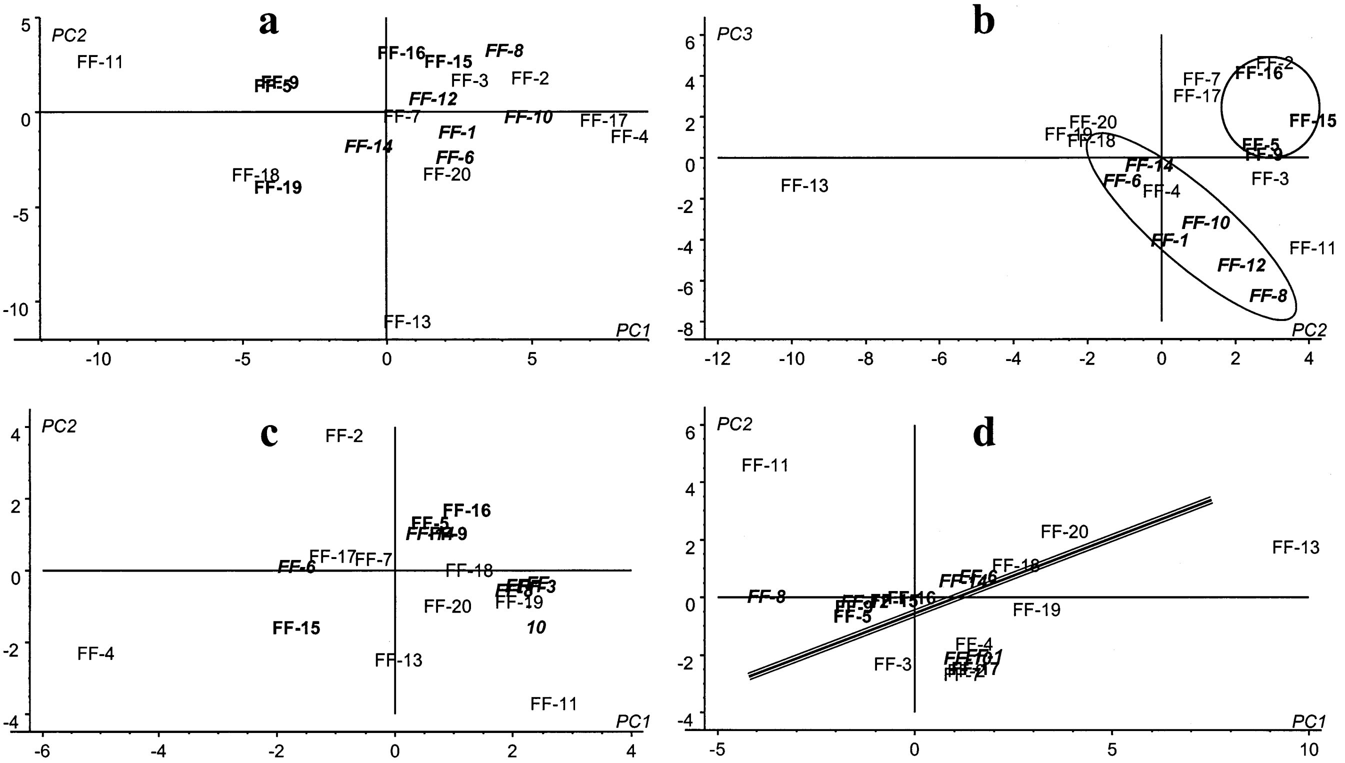

Fig. 8. PCA score plots. AMBER force fields are indicated in italic, MM3 force fields are indicated in bold, and all others are indicated with normal characters. (a) PCA on all data (78 variables). This score plot (PC1 vs. PC2) only contains (18 + 12%) 30% of the entire variation. (b) PCA on all data (78 variables). The score plot (PC2 vs. PC3) contains 12 + 11% of the entire variation. (c) PCA on energetic data (12 variables). The score plot (PC1 vs. PC2) contains 27 + 20% of the entire variation. (d) PCA on geometrical data minus torsional angles (18 variables). The score plot (PC1 vs. PC2) contains 46 + 18% of the entire variation.

from its heritage. Whether this is due to the potential energy functional form, its basic parameterization (e.g., van der Waals parameters of hydrogen and carbon), or to the use of the same type of energy minimization (algorithm and convergence criterion), is an interesting question which should be investigated further. In Fig. 8b we again find FF-13 as an extreme force field in the PC2 direction; this is not surprising if we compare with Fig. 8a, and, furthermore, we find that the MM3-based force fields are located in the same region, despite their relative strong differences along PC1.

The above-mentioned PCA was performed on all 78 variables in the test which contributed with equal weight, giving a relatively high influence of dihedral performance (46 out of 78 features). If we instead run a PCA on only the energetic information (12 out of 78 features), we obtain the score plot shown in Fig. 8c. We observe that FF-2, FF-4, and FF-11 are the extreme force fields spanning the force field variations. Moreover, we note that the Macromodel AMBER* parameterizations both with (FF-14) and without (FF-6) inclusion of GB/SA water solvation are separating rather significantly from the rest of the AMBER family (FF-1/FF-8/FF-10 and FF-12), which has clustered very tightly and included FF-3. Furthermore, we observe that the MM3 force fields FF-5 and FF-9, energetically speaking (in contrast to geometrically speaking), are very close to the Macromodel MM3* parameterization with a dielectric constant of 80 (FF-16), while the same force field with a dielectric constant of 1 (FF-15) behaves very differently.

Finally, we also ran a PCA on all geometric variables that were not dihedral angles (20 out of 78 features). The resulting score plot containing almost 60% of the total variation, Fig. 8d, displays an interesting division of the force fields. On one side of the line we find a very tight cluster containing the force fields FF-1, FF-2, FF-4, FF-7, FF-10, and FF-17; on the

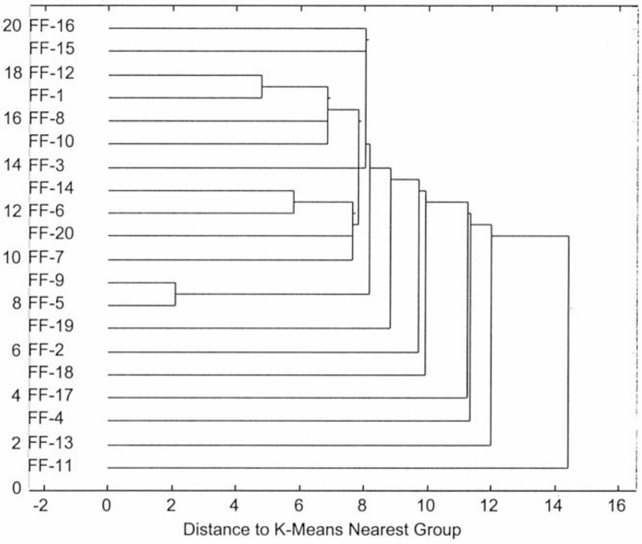

Fig. 9. Cluster analysis of the PCA score data. Dendrogram based on the distance to k-means nearest group.

other side of the line we find a stretched grouping containing all MM3-based force fields, plus FF-6, FF-14, and FF-18. The main reason for this grouping has to do with different modeling of β anomeric bonds and angles, but we have not investigated this special grouping in more detail.

As a final global analysis, we performed a cluster analysis of the score vectors from the PCA on the entire data set. The result in the form of a dendrogram is displayed in Fig. 9. The cluster analysis was performed on the 12 first-score vectors, which was the (leave one out validated) number of significant principal components in the PCA. The figure clearly demonstrates that the two MM3 force fields, FF-5 and FF-9, are by far the two most closely related. FF-1 and FF-12 and FF-6 and FF-14 are closely related; they are all AMBER-based force fields. Apparently, the GLYCAM parameterization (FF-1) is the one which is the closest to the native AMBER performance (FF-12), while the two Macromodels AMBER* behave similarly with (FF-14) or without (FF-6) GB/SA water solvation. Furthermore, we observe that FF-1 and FF-12 also cluster relatively closely with FF-8 and FF-10, which are the Glennon and Homans parameterizations of AMBER, respectively. Finally, we observe that the force fields FF-19, FF-2, FF-18, FF-17, FF-4, FF-13, and FF-11 all contribute with individual answers to the test problems (globally speaking) which are not in the same family as the answers from any of the other force fields.

4. Conclusions

The interest in carbohydrate modeling has led to the development of several sets of carbohydrate parameters that were intended to accommodate the special effects encountered in this family of molecules. The application of 20 of these force fields and/or sets of parameters to a series of seven test cases provides a fairly nonuniform picture of the potentiality of these parameter sets for giving a consistent image of structure and energy of carbohydrate molecules. Actually, the only agreement appears at the disaccharide level, when it comes to assessing the occurrence of the low-energy conformers. At this level of structural complexity, van der Waals interactions become the predominant forces for which proper parameterizations exist. However, the relative energy of these conformers varies greatly according to the parameter sets used. Therefore, the consideration of these relative energies in further calculations is a highly risky exercise.

When confronted with the task of selecting a force field, the inexperienced user should be aware that most molecular mechanics force fields have been developed to reproduce a given set of molecular features in a given environment for a given set of molecules with the aim of predicting molecular properties of related molecular systems for which detailed experimental data are not available. Therefore, particular care should be taken as to what level of extrapolation can be expected to be successful. Evidently, the high level of self-consistency of each development precludes the use of sets of parameters from different origins and/or the combination of force fields.

It is certainly true that there is a lack of suitable experimental data to use as benchmarks in testing of the force fields for carbohydrate modeling. The successful application of solvation with explicit solvent molecules may provide insight into the underlying physical properties of carbohydrates. Nevertheless, other basic features remain to be considered such as (i) how can one explore conformational hyperspace in an efficient way and (ii) which is the appropriate model to be used in simulating the observed spectroscopic properties? This means that validation of computer predictions requires a close interplay between experimentalists and theoreticians.

The result derived from the chemometric analysis provides a global view of the performances of the force fields and parameter sets for carbohydrates. Among the many conclusions that can be reached, the present analysis (i) provides an identification of the parameter sets that differ from the bulk, (ii) helps to establish the relationship that exists between the different parameter sets, (iii) provides indications for selecting different parameter sets to explore the force field dependency (or the lack thereof) of a given molecular modeling study.

Through the PCA we have created a force-field landscape on which the different force fields are related to each other on a relative scale. New carbohydrate force fields can easily be inserted into this landscape (PCA model) and related to the performance of existing force fields. If a scientist has a hypothesis on how a force field should perform on the proposed test systems, this hypothesis can with equal ease be inserted into the model as a reference center. And even better, we may one day be able to provide the true solution for the proposed test systems and, by inserting this in the model, be able to calculate the force field error as the distance to the true solution.

References

[1] R.H. Marchessault, S. Pérez, *Biopolymers*, 18 (1979) 2369–2374.

[2] E. Hough, H. Neidle, D. Rogers, P.G.H. Troughton, *Acta Crystallogr. Sect. B*, 29 (1973) 365–367.

[3] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Jr. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, *J. Am. Chem. Soc.*, 117 (1995) 5179–5197.

[4] D.A. Pearlman, D.A. Case, J.C. Caldwell, G.L. Seibel, U. Chandra Singh, P. Weiner, P.A. Kollman, AMBER 4.0, University of California, San Francisco, CA, 1991.

[5] S.W. Homans, *Biochemistry*, 29 (1990) 9110–9118.

[6] S.N. Ha, A. Giammona, M. Field, J.W. Brady, *Carbohydr. Res.*, 180 (1988) 207–221.

[7] T.M. Glennon, Y.-J. Zheng, S.M. LeGrand, B.A. Shutzberg, K.M. Merz Jr., *J. Comput. Chem.*, 15 (1994) 1019–1040.

[8] R.J. Woods, R.A. Dwek, C.J. Edge, B. Fraser Reid, *J. Phys. Chem.*, 99 (1995) 3832–3846.

[9] F. Mohamadi, N.G.J. Richards, W.C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson, W.C. Still, *J. Comput. Chem.*, 11 (1990) 440–467.

[10] H. Senderowitz, W.C. Still, *J. Org. Chem*, 62 (1997) 1427–1438.

[11] W.C. Still, A. Tempczyk, R.C. Hawley, T. Hendrickson, *J. Am. Chem. Soc.*, 112 (1990) 6127–6129.

[12] M. Dauchez, P. Derreumaux, P. Lagant, G. Vergoten, *J. Comput. Chem.*, 16 (1995) 188–199.

[13] B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, M. Karplus, *J. Comput. Chem.*, 4 (1983) 187–217.

[14] M.L.C.E. Kouwijzer, P.D.J. Grootenhuis, *J. Phys. Chem.*, 99 (1995) 13426–13436.

[15] W.F. van Gunsteren, H.J.C. Berendsen, *Groningen Molecular Simulation (GROMOS) Library Manual*, BIOMOS, Nijenborgh 16, Groningen, The Netherlands, 1987.

[16] J.H.E. Koehler, W. Saenger, W.F. van Gunsteren, *Eur. Biophys. J.*, 25 (1987) 197–210.

[17] M. Clark, R.D. Cramer III, N. van Opdenbosch, *J. Comput. Chem.*, 62 (1989) 982–1012.

[18] S. Pérez, C. Meyer, A. Imberty, in A. Pullman, J. Jortner, B. Pullman (Eds.), *Modelling of Biopolymers Structures and Mechanisms*, Kluwer, Dordrecht, 1995, pp. 425–454.

[19] N.L. Allinger, Y.H. Yuh, J.-H. Lii, *J. Am. Chem. Soc.*, 111 (1989) 8551–8567.

[20] N.L. Allinger, M. Rahman, J.-H Lii, *J. Am. Chem. Soc.*, 112 (1990) 8293–8307.

[21] J. Fabricius, S.B. Engelsen, K. Rasmussen, *J. Carbohydr. Chem.*, 16 (1997) 751–772.

[22] S. Lifson, A. Warshel, *J. Chem. Phys.*, 49 (1968) 5116–5129.

[23] K. Rasmussen, S.B. Engelsen, J. Fabricius, B. Rasmussen, *Recent Experimental and Computational Advances in Molecular Spectroscopy*, NATO ASI Series C: Mathematical and Physical Sciences, Vol. 406, 1993, pp 381–419.

[24] S.L. Mayo, B.D. Olafson, W.A. Goddard, *J. Phys. Chem.*, 94 (1990) 8897–8909.

[25] S. Wold, K. Esbensen, P. Geladi, *Chemometr. Intell. Lab. Syst.*, 2 (1987) 37–52.

[26] D.L. Massart, L. Kaufman, *Interpretation of Analytical Data by the Use of Cluster Analysis*, Wiley, New York, 1983.

[27] D. Cremer, J.A. Pople, *J. Am. Chem. Soc.*, 97 (1975) 1354–1358.

[28] A.D. French, V. Tran, *Biopolymers*, 29 (1990) 1599–1611.

[29] J. Gruza, K. Koca, S. Pérez, A. Imberty, *J. Mol. Struct. (Theochem.)*, 424 (1998) 269–280.

[30] V. Smirnyagin, C.T. Bishop, *Can. J. Chem.*, 46 (1968) 3085–3090.

[31] G.A. Jeffrey, R. Taylor, *J. Comput. Chem.*, 1 (1980) 99–109.

[32] Y. Nishida, H. Ohrui, H. Meguro, *Tetrahedron Lett.*, 25 (1984) 1575–1578.

[33] J.W. Brady, *J. Am. Chem. Soc.*, 111 (1989) 5155–5165.

[34] A. Imberty, *Curr. Opin. Struct. Biol.*, 7 (1997) 617–723.