JOURNAL OF CHEMOMETRICS
J. Chemometrics 2004; **18**: 37–42
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.842

A dedicated generalized Procrustes algorithm for consensus molecular alignment

Jacques J. F. Commandeur[1], Pieter M. Kroonenberg[2]* and William J. Dunn III[3]

[1]Boerhaavelaan 80, NL-2334 ES Leiden, The Netherlands
[2]Department of Education, Leiden University, Wassenaarseweg 52, NL-2333 AK Leiden, The Netherlands
[3]Department of Medicinal Chemistry, University of Illinois at Chicago, 833 S Wood Street, Chicago, IL 60612, USA

Received 10 August 2003; Revised 18 December 2003; Accepted 24 December 2003

Recently the idea of using generalized Procrustes analysis for aligning sets of molecules was introduced using standard algorithms. In this paper it is shown that, by tailoring the algorithm to this specific problem, a great gain in computational speed and memory efficiency can be obtained, but even more importantly, by using rotations without reflection, changes in chirality of molecules can be prevented, which was not previously possible. Copyright © 2004 John Wiley & Sons, Ltd.

KEYWORDS: molecular alignment; algorithms; generalized Procrustes analysis; rotation; reflection; cocaine; chirality; QSAR

1. INTRODUCTION

One of the tasks in three-dimensional quantitative structure–activity relationship (3D-QSAR) analyses of structure–activity data is to optimally align several compounds in data sets. In a previous paper, Kroonenberg *et al*. [1] proposed to base this alignment process on generalized Procrustes analysis (GPA). Their approach is a geometric one, as the proposed alignment was based on the relative positions of common atoms in three dimensions, and it resulted in a consensus alignment that used all molecules in the data set and avoided the bias introduced in the pairwise alignment strategy which is commonly used (for a review see Reference [2]). In their paper the authors provide the rationale for this strategy and discuss in detail the chemical background to the alignment problem.

Generalized Procrustes analysis (see Reference [3] for an overview), named after the innkeeper in Greek mythology who shaped the torsos of his guests in various ways so as to give them an optimal fit to their beds, is often used to align configurations from different analyses to assess them with respect to each other. The technique is frequently used in the social and behavioral sciences, but has seen even more use in sensory perception [4]. Procrustes analysis has been used in QSAR studies to compare different sets of properties of diverse compounds in a similarity study [5], and the technique has also been used in X-ray crystallography to align two molecules in a pairwise manner [6]. Apart from the work of Kroonenberg *et al*. [1], GPA has not been used for geometric alignment of sets of molecules, but independently Robert and Carbó-Dorca [6] hinted that GPA might be useful for such alignments.

For the application of GPA to the general alignment problem, each molecule is represented by a matrix in which the rows are the atoms, with mostly a different number of them for each molecule, and the columns are their co-ordinates in three-dimensional space. Commandeur [7] developed a variant of GPA that will handle differing numbers of rows in different samples. This feature makes it possible to use the technique for aligning molecules that have some common atoms and some non-common or unique ones, i.e. for partial alignment (see also Reference [8]). Given that a compound is rigid, aligning on common atoms automatically orients the geometric arrangements of the other atoms of the molecules so that they can be assessed with respect to each other and to atoms in other molecules.

To realize an alignment of the molecules in a study, at least three common atoms have to be chosen for the alignment rule. Furthermore, it has to be decided whether it is desirable to align (1) on a limited number of atoms or (2) on all atoms that are common to all molecules or (3) on all atoms that are common to all molecules as well as on those atoms that are common to two or more but not all molecules. When aligning on a limited number of atoms, their choice is of crucial importance and should be determined on chemical grounds. This type of choice is not considered in the present paper.

1.1. Chirality

When aligning molecules, there is a potential problem in conserving their chirality, where chirality refers to the 'handedness' of molecules and is related to the order in which the bonds are arranged around atoms. Even though two molecules may fit perfectly onto each other after reflection, such

*Correspondence to: P. M. Kroonenberg, Department of Education, Leiden University, Wassenaarseweg 52, NL-2333 AK Leiden, The Netherlands.
E-mail: KROONENB@fsw.leidenuniv.nl

Figure 1. Left version and right version of CHFClBr molecule.

reflections are not allowed, because the absolute stereochemistry of molecules must be maintained. Receptors are chiral and thus stereoisomers can have different biological properties. Therefore, either a proper selection of the atoms on which the molecules are to be aligned has to be made such that reflections do not occur, or the algorithm which optimally aligns the molecules should not permit the reflections themselves. In this respect, two situations are relevant when aligning molecules. The first is the situation where chiral molecules are included in the set to be aligned; the second is where the molecules in the set are non-chiral and reflections should be avoided to prevent them becoming chiral. In many cases a solution could be to include a judicious selection of atoms in the alignment set so that reflection cannot occur. Unfortunately, in very simple molecules this might not always be possible. For example, for the CHFClBr molecules in Figure 1 the only atoms available for alignment are the H and C atoms, because including any of the other atoms will destroy the stereochemistry of one of the molecules.

In more complex molecules where non-chiral atoms define the three-dimensional structure, a proper selection of atoms may prevent reflections. An example of a 'bad' choice for the standard algorithm is the phenyl ring in cocaine molecules, because the ring fits perfectly in two dimensions and therefore contains no information on the third dimension, so that rotations with reflections will not rule out the appearance of chiral molecules, as is demonstrated in Figure 2.

As shown in Reference [1], alignment can be done by using publicly available software; however, as mentioned in that paper, the standard GPA algorithms permit undesirable reflections of configurations. Moreover, as is explained in the Appendix of Reference [1], standard algorithms are unnecessarily computationally inefficient and are rather wasteful in their memory usage. The dedicated algorithm proposed in the present paper solves the reflection and the memory issue and also avoids several computational inefficiencies.

2. GENERALIZED PROCRUSTES ANALYSIS

In GPA the match between n configurations $\mathbf{X}_j$, each of order $p \times m$, is investigated under all relative Euclidean distance preserving transformations. In the present case the $\mathbf{X}_j$ contain the geometric configuration of molecule j described by the co-ordinates of its atoms (rows) on three co-ordinate axes (columns). Relative Euclidean distance-preserving transformations are orthonormal transformations (i.e. rotations and reflections), translations and isotropic scaling factors. Since consensus molecular alignment requires absolute Euclidean distance preserving transformations, isotropic scaling factors are undesirable and will therefore not be considered in the present paper.

In the case of consensus molecular alignment, not all atoms are common to all molecules. Formally, this can be 'translated' into a matching problem where some rows of the configurations are missing and others are not (see Reference [1] for details). Gower [9] first solved the problem of matching n configurations subject to relative Euclidean distance-preserving transformations and coined the term *generalized Procrustes analysis*. Ten Berge [10] improved Gower's method for the determination of isotropic scaling factors and orthonormal transformations. Commandeur [7] showed how to handle the situation of missing rows in GPA, of which a condensed overview is provided first.

2.1. Generalized Procrustes analysis with missing rows

The dimensionality (m) in molecular consensus alignment is by definition always equal to three, but in our theoretical exposition we will describe the m-dimensional case for greater generality.

Let $\mathbf{M}_j$ be a diagonal matrix of order $p \times p$ with ones on the diagonal if the corresponding rows in $\mathbf{X}_j$ are not missing, and with zeros on the diagonal for rows which are missing. Also, let $\mathbf{1}$ be an appropriately sized column vector consisting of ones, $\mathbf{R}_j$ be an unknown orthonormal (rotation) matrix of order $m \times m$, and $\mathbf{u}_j$ be an unknown translation vector of order $m \times 1$. Let the n translation vectors $\mathbf{u}_j$ be collected in the $nm \times 1$ partitioned column vector $\mathbf{u}$, and the orthonormal matrices $\mathbf{R}_j$ in the $nm \times m$ partitioned matrix $\mathbf{R}$, then the least squares criterion or loss function for evaluating the match between n given—and possibly incomplete—configurations $\mathbf{X}_j$ in GPA is defined as

$$f(\mathbf{u}, \mathbf{R}, \mathbf{Z}) = \sum_{j=1}^{n} \mathrm{tr}[(\mathbf{X}_j - \mathbf{1}\mathbf{u}_j^{\mathrm{T}})\mathbf{R}_j - \mathbf{Z}]^{\mathrm{T}} \times \mathbf{M}_j[(\mathbf{X}_j - \mathbf{1}\mathbf{u}_j^{\mathrm{T}})\mathbf{R}_j - \mathbf{Z}] \quad (1)$$

In (1), $\mathbf{Z}$ is an unknown group or centroid configuration of order $p \times m$.

As discussed in Reference [7] for fixed $\mathbf{R}$ and $\mathbf{Z}$, and keeping all but the jth translation vector $\mathbf{u}_j$ fixed, the conditional

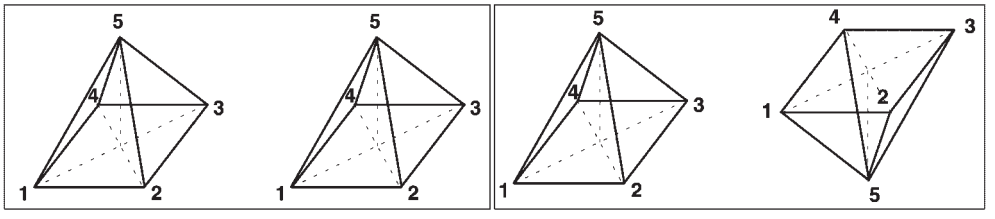

Figure 2. Original configurations (left). Configurations after a Procrustes analysis in which the common atoms consisted of the bases of the two pyramids (right). Distances are unaffected.

global minimum of (1) is obtained for

$$\mathbf{u}_j = \frac{(\mathbf{X}_j - \mathbf{Z}\mathbf{R}_j^{\mathrm{T}})^{\mathrm{T}}\mathbf{M}_j\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{M}_j\mathbf{1}} \tag{2}$$

Substituting all the $\mathbf{u}_j$ defined in (2) into (1) gives

$$f(\mathbf{R},\mathbf{Z}) = \sum_{j=1}^{n} \mathrm{tr}(\mathbf{X}_j\mathbf{R}_j - \mathbf{Z})^{\mathrm{T}}\mathbf{C}_j(\mathbf{X}_j\mathbf{R}_j - \mathbf{Z}) \tag{3}$$

where the

$$\mathbf{C}_j = \mathbf{M}_j\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{M}_j}{\mathbf{1}^{\mathrm{T}}\mathbf{M}_j\mathbf{1}}\right) \tag{4}$$

are symmetric and idempotent centering matrices, leaving (1) to be minimized over only two sets of unknowns: **R** and **Z**.

For fixed **R**, Commandeur [7] showed that the conditional global minimum of (3) with respect to **Z** is attained for

$$\mathbf{Z} = \mathbf{C}^{-}\sum_{j=1}^{n}\mathbf{C}_j\mathbf{X}_j\mathbf{R}_j \tag{5}$$

where

$$\mathbf{C} = \sum_{j=1}^{n}\mathbf{C}_j \tag{6}$$

and $\mathbf{C}^{-}$ of order $p \times p$ is the Moore–Penrose inverse of the sum of the centering matrices $\mathbf{C}_j$ in (4). The reason why a generalized inverse must be determined in order to solve for unknown **Z** is that

$$\begin{aligned}\mathbf{C}\mathbf{1} &= \sum_{j=1}^{n}\mathbf{M}_j\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{M}_j}{\mathbf{1}^{\mathrm{T}}\mathbf{M}_j\mathbf{1}}\right)\mathbf{1} \\ &= \sum_{j=1}^{n}\mathbf{M}_j\mathbf{1} - \sum_{j=1}^{n}\mathbf{M}_j\mathbf{1} = \mathbf{0}\end{aligned} \tag{7}$$

showing that **C**, the sum of the centering matrices $\mathbf{C}_j$, has no proper inverse. In Section 4 we will have more to say about this issue.

Substitution of (5) into (3) yields a loss function that no longer contains matrix **Z**,

$$\begin{aligned}f(\mathbf{R}) = &\sum_{j=1}^{n}\mathrm{tr}\mathbf{R}_j^{\mathrm{T}}\mathbf{X}_j^{\mathrm{T}}\mathbf{C}_j\mathbf{X}_j\mathbf{R}_j \\ &- \mathrm{tr}\left(\sum_{j=1}^{n}\mathbf{C}_j\mathbf{X}_j\mathbf{R}_j\right)^{\mathrm{T}}\mathbf{C}^{-}\left(\sum_{j=1}^{n}\mathbf{C}_j\mathbf{X}_j\mathbf{R}_j\right)\end{aligned} \tag{8}$$

leaving the original loss function (1) to be minimized over only one set of unknowns.

The minimization of (8) is equivalent to the maximization of

$$g(\mathbf{R}) = \mathrm{tr}\left(\sum_{j=1}^{n}\mathbf{C}_j\mathbf{X}_j\mathbf{R}_j\right)^{\mathrm{T}}\mathbf{C}^{-}\left(\sum_{j=1}^{n}\mathbf{C}_j\mathbf{X}_j\mathbf{R}_j\right) \tag{9}$$

under the constraint that the n square matrices $\mathbf{R}_j$ are orthonormal, i.e.

$$\mathbf{R}_j^{\mathrm{T}}\mathbf{R}_j = \mathbf{R}_j\mathbf{R}_j^{\mathrm{T}} = \mathbf{I}_m \quad \text{for} \quad j = 1,\ldots,n \tag{10}$$

Considering only one particular $\mathbf{R}_j$, it can be verified that the maximization of (9) is equivalent to the maximization of

$$h(\mathbf{R}_j) = \mathrm{tr}\mathbf{R}_j^{\mathrm{T}}\mathbf{B}_j \tag{11}$$

with

$$\mathbf{B}_j = \mathbf{X}_j^{\mathrm{T}}\mathbf{C}_j\mathbf{C}^{-}\sum_{i\neq j}\mathbf{C}_i\mathbf{X}_i\mathbf{R}_i \tag{12}$$

Function (11) is globally maximized subject to (10) by

$$\mathbf{R}_j = \mathbf{P}\mathbf{Q}^{\mathrm{T}} \tag{13}$$

where **P** and **Q** (both of order $m \times m$) hold the left and right singular vectors of the singular value decomposition

$$\mathbf{B}_j = \mathbf{P}\boldsymbol{\Phi}\mathbf{Q}^{\mathrm{T}} \tag{14}$$

For a proof we refer to References [10,11].

Since (9) cannot be solved for the n $\mathbf{R}_j$ simultaneously, an alternating least squares algorithm [12] is used where (9) is consecutively being maximized for $j = 1,\ldots,n$ and **R** is updated after each step. This process is repeated until n steps jointly fail to raise (9) above some threshold value. It can be shown that (9) will increase at each step until the algorithm convergences, albeit not necessarily to the global maximum of (9) (see Reference [10] for details).

At convergence the centroid configuration **Z** can be calculated from (5) and the translation vectors can then be obtained from (2). Because the whole GPA solution is unique up to a simultaneous orthonormal transformation of all optimally transformed configurations, the whole solution is generally rotated to the principal components of **Z**. If **K** denotes the $m \times m$ matrix of eigenvectors from the eigenvalue–eigenvector decomposition

$$\mathbf{Z}^{\mathrm{T}}\mathbf{C}\mathbf{Z} = \mathbf{K}\boldsymbol{\Lambda}\mathbf{K}^{\mathrm{T}} \tag{15}$$

then the principal component orientation is achieved by computing **ZK** and $(\mathbf{X}_j - \mathbf{1}\mathbf{u}_j^{\mathrm{T}})\mathbf{R}_j\mathbf{K}$.

As discussed in Reference [7], the total sum of squares of the n configurations about the origin can be partitioned into two parts:

$$\sum_{j=1}^{n}\mathrm{tr}\mathbf{X}_j^{\mathrm{T}}\mathbf{C}_j\mathbf{X}_j = \mathrm{tr}\mathbf{Z}^{\mathrm{T}}\mathbf{C}\mathbf{Z} + f(\mathbf{u},\mathbf{R},\mathbf{Z}) \tag{16}$$

The first part, $\mathrm{tr}\mathbf{Z}^{\mathrm{T}}\mathbf{C}\mathbf{Z}$, is the proportion of the total sum of squares accounted for by the GPA model, while the second part, $f(\mathbf{u}, \mathbf{R}, \mathbf{Z})$, is a residual sum of squares (SS). This means that the term

$$\mathrm{SS}_{\mathrm{fit}} = \frac{\mathrm{tr}\mathbf{Z}^{\mathrm{T}}\mathbf{C}\mathbf{Z}}{\sum_{j=1}^{n}\mathrm{tr}\mathbf{X}_j^{\mathrm{T}}\mathbf{C}_j\mathbf{X}_j} \tag{17}$$

satisfies $0 \leqslant \mathrm{SS}_{\mathrm{fit}} \leqslant 1$ and can be interpreted as the proportion of explained variation in the GPA solution.

2.2. The GPA algorithm and molecular alignment

In the context of consensus molecular alignment there are two problems with the algorithm outlined above. Firstly, the solution (13) incorporates both rotations and reflections, with the possibility of the undesirable effect of transforming molecules into their chiral counterparts in order to obtain an optimal match. Therefore in Section 3 a procedure is discussed for obtaining the optimum of (11) under the additional restriction that $\mathbf{R}_j$ is a *pure rotation* matrix.

Secondly, as discussed in Reference [1], each of the n molecules $\mathbf{X}_j$ may contain many unique atoms, so that the number

of rows of the configurations (p) can easily become very large (see also Section 5.2). Amongst others, this has the drawback that the computation of the generalized inverse of the $p \times p$ matrix $\mathbf{C}$ in (6)—even though it has only to be calculated once—can become very time-consuming. In Section 4 we show how this problem can be handled by restricting the analysis to the non-unique atoms of the molecules only.

3. THE PURE ROTATION OPTIMIZATION PROCEDURE

Letting $|\mathbf{A}|$ denote the determinant of a square matrix $\mathbf{A}$, it follows from (10) that

$$|\mathbf{R}_j\mathbf{R}_j^{\mathrm{T}}| = |\mathbf{I}_m| = 1 \tag{18}$$

the determinant of $\mathbf{I}_m$ being the product of its eigenvalues. Because $\mathbf{R}_j$ is a square matrix and, moreover,

$$|\mathbf{R}_j\mathbf{R}_j^{\mathrm{T}}| = |\mathbf{R}_j|^2 \tag{19}$$

it follows from (18) and (19) that

$$|\mathbf{R}_j| = \pm 1 \tag{20}$$

If the determinant of $\mathbf{R}_j$ equals 1, the orthonormal matrix is a pure rotation matrix; if its determinant equals -1, then the orthonormal matrix also involves a reflection. In order to avoid a reflection, (11) must be maximized subject to *two* restrictions:

$$\mathbf{R}_j^{\mathrm{T}}\mathbf{R}_j = \mathbf{R}_j\mathbf{R}_j^{\mathrm{T}} = \mathbf{I}_m \tag{21}$$

and

$$|\mathbf{R}_j| = 1 \tag{22}$$

Together, conditions (21) and (22) guarantee that $\mathbf{R}_j$ is a pure rotation matrix and therefore that no stereoisomers of the original molecules can be obtained.

The solution to the maximization of (11) subject to (21) and (22) is due to Gower [13]. Let

$$\hat{\mathbf{R}}_j = \mathbf{P}\mathbf{Q}^{\mathrm{T}} \tag{23}$$

where $\mathbf{P}$ and $\mathbf{Q}$ (both of order $m \times m$) hold the left and right singular vectors from the singular value decomposition defined in (14). When (23) satisfies (22), we are done. On the other hand, when $|\hat{\mathbf{R}}_j| = -1$, that column of either matrix $\mathbf{P}$ or matrix $\mathbf{Q}$ (which one of the two is immaterial) must be reflected which corresponds with the smallest singular value on the diagonal of $\mathbf{\Phi}$ in (14).

4. THE DEDICATED GPA ALGORITHM

Since unique atoms always fit perfectly in the GPA solution, they contribute nothing whatsoever to the value of loss function (1) and therefore do not need to enter the GPA itself.

4.1. GPA on the set of non-unique atoms

The computational effort required in a GPA of molecules each containing many unique atoms can be considerably reduced by restricting the analysis to the non-unique atoms of the molecules only. Formally, this is achieved by selecting those rows of the n molecules corresponding to the elements of the $p \times 1$ vector

$$\mathbf{w} = \sum_{j=1}^{n} \mathbf{M}_j\mathbf{1} \tag{24}$$

(where matrices $\mathbf{M}_j$ are defined in (1)) which satisfy

$$w_i \neq 1, \quad i = 1, \ldots, p \tag{25}$$

We will denote the number of elements of (24) satisfying the latter inequality with s, so that s is the number of non-unique atoms and $t = p - s$ is the total number of unique atoms.

Also, let $\mathbf{M}_j^*$ and $\mathbf{X}_j^*$ of order $s \times s$ and $s \times m$ denote the submatrices of $\mathbf{M}_j$ and $\mathbf{X}_j$ corresponding to the elements of (24) satisfying (25) respectively. Then (4) and (6) become

$$\mathbf{C}_j^* = \mathbf{M}_j^*\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{M}_j^*}{\mathbf{1}^{\mathrm{T}}\mathbf{M}_j^*\mathbf{1}}\right) \tag{26}$$

and

$$\mathbf{C}^* = \sum_{j=1}^{n} \mathbf{C}_j^* \tag{27}$$

respectively and the computation of the Moore–Penrose inverse of the $p \times p$ matrix $\mathbf{C}$ in (6) is reduced to a potentially much smaller problem of order $s \times s$.

Moreover, in the special situation where atoms are either common to *all* molecules or unique to one molecule (i.e. the situation where vector $\mathbf{w}$ in (24) consists of s elements equal to m, and t elements equal to 1), the analysis can even be restricted to n configurations $\mathbf{X}_j^*$ containing no missing rows at all. In that case the computation of the Moore–Penrose inverse of (27) becomes especially easy:

$$\mathbf{C}^{*-} = \left(\sum_{j=1}^{n} \mathbf{C}_j^*\right)^{-} = (n\mathbf{J})^{-} = \frac{1}{n}\mathbf{J} \tag{28}$$

where

$$\mathbf{J} = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\mathbf{1}} \tag{29}$$

is the complete centering matrix of order $s \times s$, so that in this situation the computation of an inverse is not necessary. The last equality in (28) holds true because $\mathbf{J}$ in (29) is a symmetric and idempotent matrix satisfying $\mathbf{J}\mathbf{J} = \mathbf{J}$ (see e.g. Reference [14], pp. 32–33).

For given configurations $\mathbf{X}_j^*$ of order $s \times m$ we now have to maximize

$$g(\mathbf{R}) = \mathrm{tr}\left(\sum_{j=1}^{n} \mathbf{C}_j^*\mathbf{X}_j^*\mathbf{R}_j\right)^{\mathrm{T}}\mathbf{C}^{*-}\left(\sum_{j=1}^{n} \mathbf{C}_j^*\mathbf{X}_j^*\mathbf{R}_j\right) \tag{30}$$

subject to (21) and (22). This can be achieved by applying the procedures presented in Sections 2 and 3 to (30). At convergence of the algorithm the centroid configuration can be computed as

$$\mathbf{Z}^* = \mathbf{C}^{*-}\sum_{j=1}^{n} \mathbf{C}_j^*\mathbf{X}_j^*\mathbf{R}_j \tag{31}$$

and

$$\mathbf{u}_j^* = \frac{(\mathbf{X}_j^* - \mathbf{Z}^*\mathbf{R}_j^{\mathrm{T}})^{\mathrm{T}}\mathbf{M}_j^*\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{M}_j^*\mathbf{1}} \tag{32}$$

then yields the translation vectors.

4.2. Adapting the unique atoms to the GPA solution for the non-unique atoms

After the GPA of these reduced configurations we need to reinstate the unique atoms of the molecules in the obtained solution. Letting $\mathbf{K}^*$ denote the $m \times m$ matrix of eigenvectors from the eigenvalue–eigenvector decomposition

$$\mathbf{Z}^{*\mathrm{T}}\mathbf{C}^*\mathbf{Z}^* = \mathbf{K}^*\mathbf{\Lambda}^*\mathbf{K}^{*\mathrm{T}} \tag{33}$$

where $\mathbf{Z}^*$ is now of order $s \times m$, this is achieved by computing

$$\mathbf{Y}_j = (\mathbf{X}_j - \mathbf{1}\mathbf{u}_j^{*\mathrm{T}})\mathbf{R}_j\mathbf{K}^* \quad \text{for} \quad j = 1, \ldots, n \tag{34}$$

for the $p \times m$ matrices $\mathbf{X}_j$ containing the co-ordinates of the complete molecules (including their unique atoms). Finally, using the full $p \times p$ matrices $\mathbf{M}_j$ defined in (1),

$$\mathbf{Z}^+ = \left(\sum_{j=1}^{n} \mathbf{M}_j\right)^{-1} \sum_{j=1}^{n} \mathbf{M}_j\mathbf{Y}_j \tag{35}$$

yields the co-ordinates of the complete centroid configuration of order $p \times m$. It may be noted that the proportion of explained variation calculated as

$$\frac{\mathrm{tr}\mathbf{Z}^{+\mathrm{T}}\mathbf{C}\mathbf{Z}^+}{\sum_{j=1}^{n} \mathrm{tr}\mathbf{X}_j^{\mathrm{T}}\mathbf{C}_j\mathbf{X}_j} \tag{36}$$

is identical to the proportion computed in (17).

5. RESULTS

We will consider three types of improvements of the dedicated algorithm (reflection-free rotation, gain in execution speed, and reduction of memory usage) in turn. To illustrate them, the same set of cocaine derivatives was used as discussed by Kroonenberg *et al.* [1] (see their Table 1). This set consisted of 13 cocaine molecules each of which has a tropane and a phenyl ring (see Figure 3), and compared with the other molecules, cocainehin and sulphurhin have a number of additional atoms between the rings. The 13 molecules had 43, 40, 42, 40, 43, 44, 47, 47, 47, 47, 42, 45 and 41 atoms respectively, or a total of 568 atoms. All computations were carried out with Fortran90 programs compiled under Windows, which are available from the second author (see also his website: http://three-mode.leidenuniv.nl). A user-friendly interface for running the complete analysis is under construction.

5.1. Reflection-free rotation

The set of atoms proposed for the alignment of the cocaine molecules consisted of the $s = 6$ carbon atoms of the phenyl ring (see Figure 3). Because a phenyl ring is rigid, the co-ordinates after alignment should be very similar, and the six carbon atoms were chosen because these atoms should maintain the nearly identical geometry in all molecules if aligned properly. Thus misalignments should be easily identified if the atoms of the phenyl ring in all molecules were not nearly perfectly superimposed onto each other. The alignment was also chosen because it contained a plane of symmetry, which would make it easier to identify problems with the translation and/or rotations of atoms that are not used for alignment.

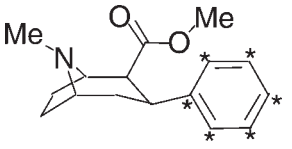

Figure 3. Alignment on the carbon atoms of the phenyl ring of cocaine molecules.

When analyzed with the standard GPA algorithm, the alignment procedure resulted in the alignment given in Plate 1A, while the analysis with the dedicated algorithm yielded the alignment given in Plates 1B and 1C. In Plate 1A, two of the molecules have been transformed into their chiral variants, as is evident from the ester group sticking out at the top, while with the dedicated algorithm the chirality of all molecules has been preserved. What is also evident is that the tropane rings and ester groups of two molecules (cocainehin and sulphurhin) are not well aligned, and this is due to the additional atoms between the tropane and the phenyl rings.

5.2. Gain in execution speed

Let r denote the total number of atoms in the molecules and s the number of atoms used for alignment (see also Section 4.1). If the s atoms are common to all n molecules, then the number of rows of the configurations required in the standard GPA algorithm equals $p = r - s(n-1)$. In the example discussed in Section 5.1, this amounts to $p = 568 - (6)(12) = 496$ rows for each configuration. Thus each configuration has $s = 6$ rows of common atoms plus 490 rows of the unique atoms of all molecules together.

Comparing the standard and the dedicated GPA algorithm, the gain in execution speed with the dedicated algorithm is primarily due to elimination of the inversion of the huge summed centering matrix $\mathbf{C}$ of order 496×496 defined in (6). In the dedicated algorithm there are only six common atoms, so that the summed centering matrix $\mathbf{C}^*$ has size 6×6, and in this case no explicit inversion is necessary because all molecules contain all six alignment atoms (see (28)). The execution speed of the standard algorithm was 27.30 s for the initialization, including the inversion of the summed centering matrix, and 2.85 s for the main algorithm, while the dedicated algorithm needed less than 0.01 s for the initialization and 0.05 s for the main iterations.

As a second example, for a comparable set of 20 amino acids the execution speed of the standard algorithm was 6.00 s for the initialization and 1.43 s for the main algorithm, while the dedicated algorithm required less than 0.001 s for both initialization and main iterations.

5.3. Gain in memory usage

Apart from an increase in computational speed by only using the common atoms for the Procrustes analysis proper, further gains can be made, especially in memory usage, by storing the data in a more efficient way. In particular, it was possible to reduce the memory requirement by storing only the molecules themselves, rather than the configurations with 496 rows of three co-ordinates. For the set of 13 cocaine derivatives with 496 rows per configuration there were a total of $496 \times 13 = 6448$ rows of three co-ordinates each, versus only 568 in the improved set-up.

For the 20 amino acids the standard algorithm required working with $20 \times 290 = 5800$ rows, while the new set-up only required 290 of them. These reductions not only have their effect in pure data storage, but also in the size of the arrays necessary for the computations.

For the basic memory allocation in the standard algorithm used on the cocaine molecules, 30 Mb basic storage was needed, while in the dedicated algorithm we could make do with 45 kb. The comparable figures for the amino acids were 15.5 Mb and 44 kb. These huge reductions in memory usage also made for more efficient computations. With more and larger molecules the comparative memory requirements of the standard algorithm will increase even further.

6. DISCUSSION

In this paper we have presented a dedicated generalized Procrustes algorithm for aligning molecules. Its special features are that it does not allow reflections during rotations, so that the stereochemistry of the molecules is preserved. Moreover, special measures have been taken in the program to reduce both execution times and memory usage. Our examples showed real gains on all accounts.

A matter not discussed in this paper, but crucial to practical applications of generalized Procrustes analysis for the alignment of sets of molecules, is the choice of common atoms for alignment. Even though this issue falls largely outside the scope of this paper, it may be remarked that, on different grounds, one can argue in favor of as many or as few atoms as possible to align on. With a few well-chosen atoms such as those in the phenyl ring, one can obtain a near perfect fit for the aligned atoms, and because of that examine in detail the differences and similarities between the other parts of the (rigid) molecules. On the other hand, one can get a good assessment of the overall similarities between the molecules by choosing, for instance, the phenyl ring, the tropane ring and the ester group for alignment, with the exception of cocainehin and sulphurhin for which only the phenyl ring would be fitted. Thus the purpose of the alignment and the type of molecules under consideration will determine the choice of alignment or common atoms.

REFERENCES

1. Kroonenberg PM, Dunn III WJ, Commandeur JJF. Consensus molecular alignment based on generalized Procrustes analysis. *J. Chem. Info. Comput. Sci.* 2003; **4**: 2025–2032.
2. Lemmen C, Lengauer T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Design* 2000; **14**: 215–232.
3. Gower JC, Dijksterhuis GB. *Procrustes Problems.* Oxford University Press: Oxford, 2004.
4. Dijksterhuis GB, Gower JC. The interpretation of generalized Procrustes analysis and allied methods. *Food Qual. Prefer.* 1992–93; **3**: 67–87.
5. Rose FS, Rahr E, Hudson BD. The use of Procrustes analysis to compare different property sets for the characterization of a diverse set of compounds. *Quant. Struct.–Act. Relat.* 1994; **13**: 152–158.
6. Robert D, Carbó-Dorca R. Anàlisi de Procrustes i alineament molecular. *Sci. Gerundensis* 1999; **24**: 175–181.
7. Commandeur JJF. *Matching Configurations.* DSWO Press: Leiden, 1991 (available from the website of The Three-Mode Company: http://three-mode.leidenuniv.nl).
8. Robinson DD, Lyne PD, Richards WG. Partial molecular alignment via local structure analysis. *J. Chem. Info. Comput. Sci.* 2000; **40**: 503–512.
9. Gower JC. Generalized Procrustes analysis. *Psychometrika* 1975; **40**: 33–51.
10. Ten Berge JMF. Orthogonal Procrustes rotation for two or more matrices. *Psychometrika* 1977; **42**: 267–276.
11. Ten Berge JMF. *Least Squares Optimization in Multivariate Analysis.* DSWO Press: Leiden, 1993.
12. Wold H. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, Krishnaiah PR (ed.). Academic Press: New York, 1966; 391–420.
13. Gower JC. Procrustes rotation problems. *Math. Sci.* 1976; **1**: 12–15.
14. Magnus JR, Neudecker H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley: Chicester, 1988.

(A)

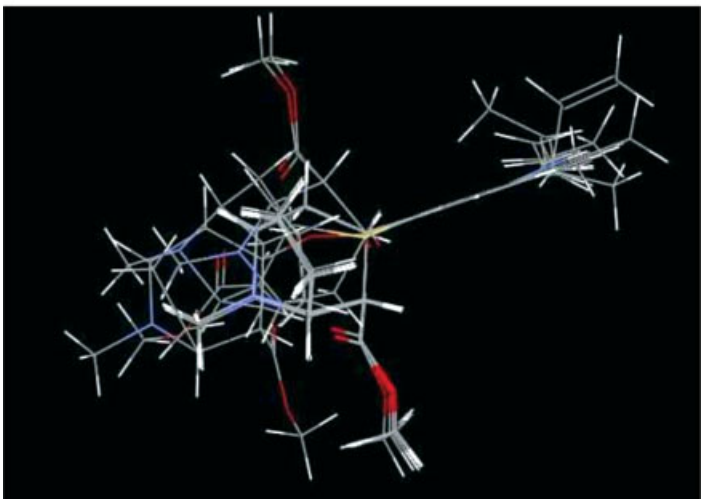

Plate 1A. Result of aligning the 13 cocaine molecules on the phenyl rings when reflections are not blocked. Two mirrored COOH ester groups can be clearly seen. The phenyl rings fit perfectly and are seen here as a line, because the plane of the phenyl rings is in the line of sight.

(B)

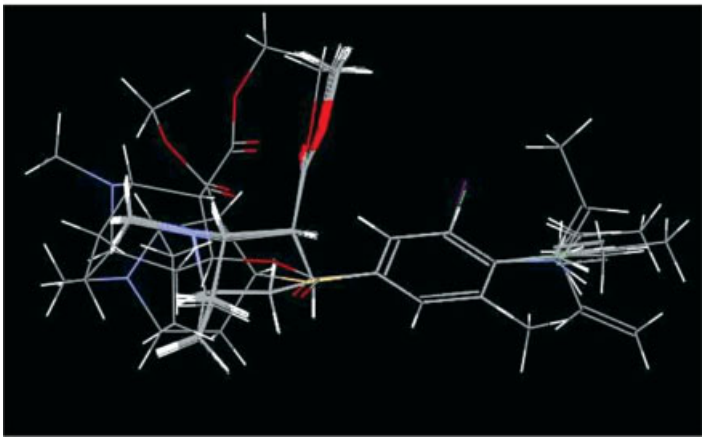

Plate 1B. Result of alignment without reflections. All phenyl rings are perfectly aligned (as in Plate 1A), and all but two tropane rings and two COOH ester groups (of cocainehin and sulphurhin) are also well aligned.

(C)

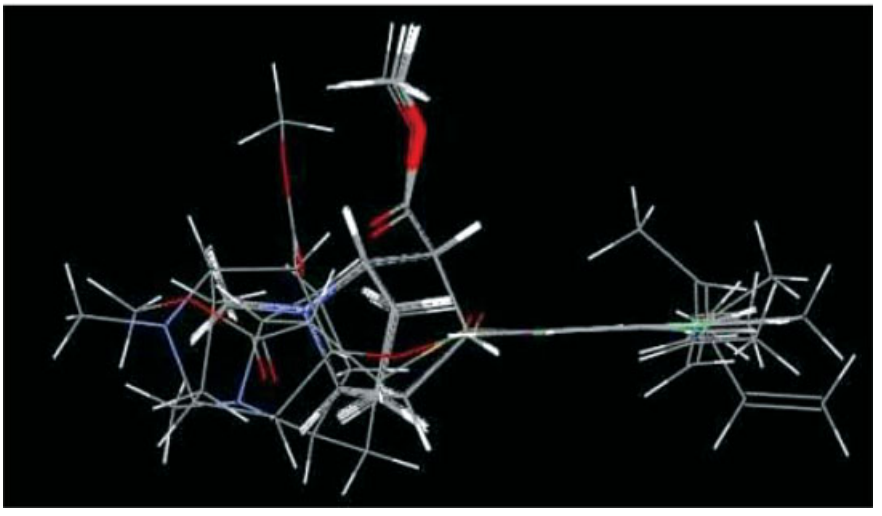

Plate 1C. Same solution as in Plate 1B, but again the plane of the phenyl rings is in the line of sight. The common position of the ester groups can now be seen more clearly.