Published online 12 August 2016 *Nucleic Acids Research, 2016, Vol. 44, No. 17* ***8479–8489***
doi: 10.1093/nar/gkw717

Conformation-dependent restraints for polynucleotides: I. Clustering of the geometry of the phosphodiester group

Marcin Kowiel[1,2], Dariusz Brzezinski[3] and Mariusz Jaskolski[1,4,*]

[1]Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan 61-704, Poland, [2]Department of Organic Chemistry, Poznan University of Medical Sciences, Poznan 60-780, Poland, [3]Institute of Computing Science, Poznan University of Technology, Poznan 60-965, Poland and [4]Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan 61-614, Poland

Received May 29, 2016; Revised August 05, 2016; Accepted August 06, 2016

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

ABSTRACT

The refinement of macromolecular structures is usually aided by prior stereochemical knowledge in the form of geometrical restraints. Such restraints are also used for the flexible sugar-phosphate backbones of nucleic acids. However, recent highly accurate structural studies of DNA suggest that the phosphate bond angles may have inadequate description in the existing stereochemical dictionaries. In this paper, we analyze the bonding deformations of the phosphodiester groups in the Cambridge Structural Database, cluster the studied fragments into six conformation-related categories and propose a revised set of restraints for the O-P-O bond angles and distances. The proposed restraints have been positively validated against data from the Nucleic Acid Database and an ultrahigh-resolution Z-DNA structure in the Protein Data Bank. Additionally, the manual classification of PO_4 geometry is compared with geometrical clusters automatically discovered by machine learning methods. The machine learning cluster analysis provides useful insights and a practical example for general applications of clustering algorithms for automatic discovery of hidden patterns of molecular geometry. Finally, we describe the implementation and application of a public-domain web server for automatic generation of the proposed restraints.

INTRODUCTION

The nucleotide unit, with six rotatable single bonds and the possibility of different sugar puckers, is a very flexible entity. In nucleic acid structures, its geometry is locked by the canonical structural constraints, such as base pairing and stacking. Yet, the variability of the sugar-phosphate backbone conformation is manifested by the existence of the canonical forms of the double helix and of their numerous variants. In RNA structures, the variability is on one hand limited by the additional O2′ hydroxyl group, but on the other hand greatly augmented in the single-stranded forms, occasionally also found in DNA.

The refinement of nucleic acid structures at lower resolution, or in the presence of disorder, requires the use of stereochemical restraints. Standard libraries of such restraints have been prepared by Parkinson *et al.* (1), Clowney *et al.* (2) and Gelbin *et al.* (3). An attempt to take account of the conformational variability of the nucleic acid chains has also been made, and is reflected in the definition of two types of PO_4 bond angles (*small* and *large*) included in the dictionary compiled by Gelbin *et al.* (3). However, this solution is quite crude, mainly because of the limited number of examples (13 cases) in the underlying study. Moreover, even though the authors recognized that the O1/O2-P-O angles (labeling as in Figure 1) have a bimodal distribution (*small*, *large*), they were unable to conclusively predict where and when the *small* or *large* angles would occur. They were, however, able to discover a bimodal distribution of the C5′-O5′-P-O3′ (τ_5) torsion angle and a trimodal distribution of the C3′-O3′-P-O5′ (τ_3) angle.

A recent extremely accurate study of the crystal structure of Z-DNA at 0.55 Å resolution (4) confirmed that the phosphate bond angles might indeed need to be revisited. In addition, that study also pointed out that the angles at the glycosidic bond of the *syn*-conformation purine units have inadequate description in the stereochemical dictionaries (to be discussed in a separate paper). The existing restraint libraries also show inconsistencies. For example, in the applications of SHELXL (5) the RNA restraints in use for the phosphodiester group still contain values from the compilation by Parkinson (1), while the proposed DNA restraints use the revised values of Gelbin (3).

[*]To whom correspondence should be addressed. Tel: +48 61 829 1274; Fax: +48 61 829 1555; Email: mariuszj@amu.edu.pl

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

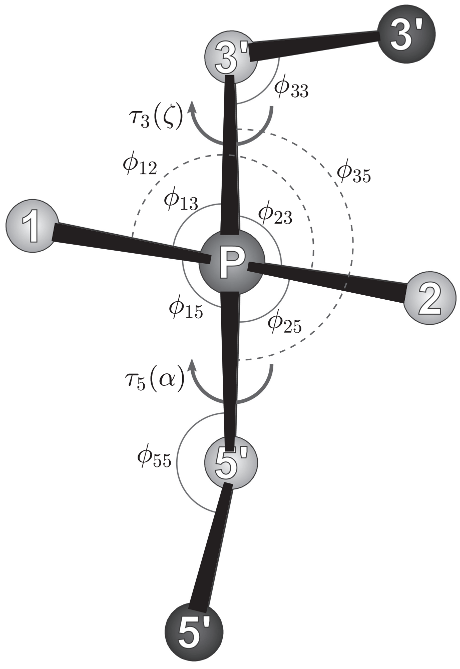

Figure 1. Atom and angle labeling. Standard atom labeling scheme of the phosphodiester C-O-PO_2-O-C group. The C-O-P-O torsion angles are denoted as C3′-O3′-P-O5′ (τ_3, ζ in standard nucleotide nomenclature) and C5′-O5′-P-O3′ (τ_5, α). The Ox-P-Oy bond angles are labeled φ_{xy} as follows: φ_{12} = O1-P-O2, φ_{13} = O1-P-O3′, φ_{15} = O1-P-O5′, φ_{23} = O2-P-O3′, φ_{25} = O2-P-O5′, φ_{35} = O3′-P-O5′, whereas the two P-O-C angles are labeled φ_{33} = P-O3′-C3′ and φ_{55} = P-O5′-C5′.

In this paper, we analyze bonding deformations of the phosphodiester fragments (C-O-PO_2-O-C) found in the Cambridge Structural Database (CSD) (6) and manually cluster them into six conformation-related categories. Based on the discovered categories, we propose a set of revised restraints for the bond angles and distances within the phosphodiester group. In a parallel study, we use machine learning algorithms to automatically discover the clustering patterns of the PO_4 geometry. More precisely, we employ four basic types of clustering algorithms, namely partitioning, hierarchical, density-based and graph-based (7). A comparison of the manual and automatic approaches shows that the results are consistent, suggesting potential applications of machine learning for the discovery of hidden molecular geometry patterns. Finally, the resulting revised restraints for the phosphodiester group geometry are validated using data retrieved from the Nucleic Acid Database (NDB) (8) and the ultrahigh-resolution Protein Data Bank (PDB) (9) Z-DNA structure 3P4J (4). To facilitate the use of the proposed restraints, we present a web server that automatically creates restraint scripts for a given .pdb file and illustrate its practical use with an example. Although Karplus *et al*. have proposed conformation-dependent restraints for proteins before (10–12), to the best of our knowledge the present work is the first application of analogous ideas to nucleic acids.

It might appear that the phosphate group is a picayune trifle not worth a deeper study. This is not true. The phosphodiester group is a key link of nucleic acid backbone conformation and in addition it is the most electron-rich (heavy) moiety of nucleic acid structures. Its proper geometrical parametrization is therefore very important for models refined against X-ray diffraction data.

MATERIALS AND METHODS

Selection of CSD fragments

The geometrical analyses were carried out for a set of small-molecule structural data retrieved from the CSD database version 5.36 using CONQUEST (13). Since there were only 10 oligonucleotide crystal structures, the CSD was queried for C-O-PO_2-O-C fragments in all crystal structures refined to $R \leq 7.5\%$. The anionic form of the phosphodiester group was guaranteed by specific exclusion of H atoms at the oxygen atoms. This approach led to the selection of 204 structures with 238 phosphate groups. The query parameters are summarized in Supplementary Table S1.

The O3′-C3′ and O5′-C5′ bonds in that CSD sample were not very good representatives of the sugar-phosphate linkages in DNA and RNA. To obtain a more adequate representation, we calculated the statistics of the P-O3′, O3′-C3′ and P-O5′, O5′-C5′ bond distances based on two supplementary CSD searches, in which the phosphodiester fragment was linked to ribose or deoxyribose at either the C3′ atom (for the P-O3′, O3′-C3′ bond length calculation) or C5′ (P-O5′, O5′-C5′) (Supplementary Table S1).

For terminal phosphates, we queried the CSD for C-O-$PO_3{}^{2-}$ fragments. We explicitly excluded H atoms by allowing only one bond for each of the -PO_3 oxygen atoms.

Selection of NDB fragments

Initially, we tried to find NDB structures with crystallographic resolution higher than 1 Å and $R \leq 7.5\%$. Surprisingly, there were only 20 crystal structures in the NDB fulfilling these criteria, mostly of dinucleotides. Even though the *R*-factor limit might seem very restrictive for macromolecular structures, we were interested in high quality crystal structures, and such an *R* factor is routinely achievable for small-molecule structures of up to ∼100 non-H atoms. It is worth noting that nearly half of the structures found in this search were solved in the 1970's and 1980's; only four of those structures were published after 2000. Therefore, the number of high-quality nucleic acid structures in the NDB is not much higher than at the time the compilations by Parkinson *et al.* (1), Clowney *et al.* (2) and Gelbin *et al*. (3) were published. With respect to the CSD the situation is dramatically different.

In view of this deficiency, we decided to relax the *R*-factor criterion to $R \leq 10.0\%$ at the validation step and to include a few extra oligonucleotide structures from the CSD, not present in the NDB. This led to 36 crystal structures, 9 from the CSD and 27 from NDB, with 126 independent phosphodiester moieties. For most of those structures it is not possible to determine whether they were refined with or without geometrical restraints, which is disappointing because for objective geometry validation one would like to use unbiased (unrestrained) information. On the other hand, the situation (even with the presence of restraints) may be not as bad as it would appear because refinement against high

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

resolution data is usually able to override the information injected to the system by geometrical restraints (14).

Outlier detection

To minimize the standard deviations in the target bond distances and angles, a modified *Z*-score test (15) was used to identify and reject outliers. In this test, a data item x_i is treated as an outlier if $|M_i| > 3.5$. M_i is calculated as follows:

$$MAD = median\{|x_i - \tilde{x}|\}$$
$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

where $\tilde{x}$ denotes the median of the sample. In the applications described in the subsequent sections, when an example was earmarked as an outlier for at least one of the calculated parameters, the entire entry was removed from all calculations as potentially contaminated by a gross error.

The geometrical characteristics of the PO_4 group

Within a phosphodiester PO_4 group, two oxygen atoms (labeled O3′ and O5′) have single bonds to carbon substituents, while the other two (O1 and O2) are terminal (Figure 1). Unlike in nucleic acids, where O3′ and O5′ are well defined, in the data obtained from the CSD the initial atom labels are arbitrary, the only restrictions being that: (i) atoms O3′/O5′ are bonded to carbon atoms, and (ii) atoms O1/O2 are bonded only to the central P atom. Even if we adhere to the IUB convention of nucleotide atom numbering (clockwise O1-O3′-O2-O5′ sequence in the Fischer projection of the P atom), there is still an atom labeling ambiguity, namely: (i) the labels attached to the O3′ and O5′ atoms are chosen arbitrarily and (ii) a small-molecule moiety must also be considered after the application of inversion symmetry, which is equivalent to swapping the O1/O2 labels and changing the signs of the torsion angles. Taking into account the O3′/O5′ and O1/O2 ambiguities gives four possible labeling schemes of the C-O-PO_2-O-C fragment (Table 1).

Only the O1-P-O2 (φ_{12}) and O3′-P-O5′ (φ_{35}) angles (Figure 1) are independent of the labeling ambiguity; the other bond and torsion angles and bond lengths may need to be permuted if the assigned atom labeling scheme is changed. The labeling ambiguity has two serious drawbacks for a statistical analysis: (i) in the original labeling, it assigns atom labels in an arbitrary way; (ii) if considered in all permissible permutations, it introduces the same data item into the statistics several times, and the count may depend on the site symmetry of a given phosphate group. The following clustering approach was designed to overcome these labeling problems.

Manual phosphodiester classification

Analyzing the angular distributions of the data retrieved from the CSD, we observed cases with unusual values of the torsion angles τ_3 and τ_5 (defined in Figure 1). More precisely, our sample contains torsion angles, not described by Parkinson *et al.* (1) or Gelbin *et al.* (3), corresponding to situations where the C-O-PO_2-O-C fragment is part of a cyclic system. In view of this observation, we sorted the 238 examples into two classes: where the above fragment is part of a macrocycle (*ring*, *R*), and where it is not (*acyclic*, *A*). These two main subpopulations (*R, A*) are referred to as *classes*.

The *R* class was further divided into four *categories*, labeled *R5*, *R6*, *R7* and *R8*, where the C-O-PO_2-O-C fragment is closed into, respectively, a five-, six-, seven- or eight-membered ring. Interestingly, each category has a characteristic pattern of the τ_3/τ_5 torsion angles. Within the *A* class, containing *acyclic* linear systems, all the τ_3 or τ_5 torsion angles were found in the +*sc*, –*sc* or *ap* regions described by Parkinson *et al.* (1) and Gelbin *et al.* (3) [the descriptors of torsion angle ranges follow the IUB recommendations: *c* - clinal (torsion angle 90 ± 60°), *p* - periplanar (0 or 180 ± 30°), *s* - syn (0 ± 90°), *a* - anti (180 ± 90°)]. However, we noticed that either both torsion angles have the same sign and are in the +*sc*/+*sc* or –*sc*/–*sc* regions, or one is in the *ap* region and the other in an *sc* region (Figure 2). Following this observation, we divided the *A* class into two categories termed: *acyclic symmetric* (*AS*) (both angles in +*sc* region or both angles in –*sc* region) and *acyclic asymmetric* (*AA*), where one angle is in the *ap* region. It is of note that the *acyclic* (*A*) class is the most interesting one since it corresponds to situations found in linear DNA and RNA oligo/polymers. In summary, we consider six torsion angle categories, as presented in Table 2.

A detailed analysis of the proposed categorization revealed that for each category only certain combinations (referred to as *groups*) of the τ_3/τ_5 torsion angle values and signs are observed in the CSD data. These combinations are summarized in Table 2 and schematically marked in Supplementary Figure S1, whereas the concrete τ_3/τ_5 values found in the CSD search are shown in Figure 2A. We note that in five-membered rings (*R5*), τ_3 and τ_5 are close to 0° making any combination of signs possible (+/–, –/+, +/+, –/–). Although the limited sample size of the *R5* category (9 examples) does not allow a full discrimination analysis, it appears that the +/+ and –/– combinations should be the most frequent ones, by analogy to the *R7* category where the ring is also odd-membered. Within the *R6* and *R8* categories the fragment is almost symmetric: φ_{13} is close to φ_{15} and φ_{23} to φ_{25} (Supplementary Figure S1A). Within the seven-membered (*R7*) and *acyclic symmetric* (*AS*) categories, the bond angles within two pairs ($\varphi_{13}/\varphi_{25}$ and $\varphi_{15}/\varphi_{23}$) are also close to each other (Supplementary Figure S1B). In the *AA* category, one torsion angle (τ_3 or τ_5) is close to 180° (Supplementary Figure S1C).

In order to establish the target values and standard uncertainties of PO_4 restraints, we categorized the phosphodiester fragments and averaged the angle and bond distance values within each category. All groups in one category represent the same kind of geometry thus contributing to improved statistics. The categorization of fragments in the *R* class was done by counting the ring atoms, whereas fragments in the *A* class were assigned to categories by manually comparing the torsion angles τ_3 and τ_5 to the characteristic templates defined in Table 2. However, at this point we cannot calculate the bond angle and bond distance statistics within a category directly because the atom labels may come from one of the four allowed permutations presented in Table 1. Therefore, we have to relabel the atoms in each exam-

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

Table 1. The four allowed label permutations of a phosphodiester group and the labeling of torsion angles, bond angles and bond lengths in relation to the initial labeling of atoms

Initial Permutation	O1	O2	O3′	O5′	τ_3	τ_5	φ_{12}	φ_{13}	φ_{15}	φ_{23}	φ_{25}	φ_{35}	φ_{33}	φ_{55}	P-O1	P-O2	P-O3′	P-O5′	O3′-C3′	O5′-C5′
(12)(35)	O2	O1	O5′	O3′	τ_5	τ_3	φ_{12}	φ_{25}	φ_{23}	φ_{15}	φ_{13}	φ_{35}	φ_{55}	φ_{33}	P-O2	P-O1	P-O5′	P-O3′	O5′-C5′	O3′-C3′
(35)	O1	O2	O5′	O3′	$-\tau_5$	$-\tau_3$	φ_{12}	φ_{15}	φ_{13}	φ_{25}	φ_{23}	φ_{35}	φ_{55}	φ_{33}	P-O1	P-O2	P-O5′	P-O3′	O5′-C5′	O3′-C3′
(12)	O2	O1	O3′	O5′	$-\tau_3$	$-\tau_5$	φ_{12}	φ_{23}	φ_{25}	φ_{13}	φ_{15}	φ_{35}	φ_{33}	φ_{55}	P-O2	P-O1	P-O3′	P-O5′	O3′-C3′	O5′-C5′

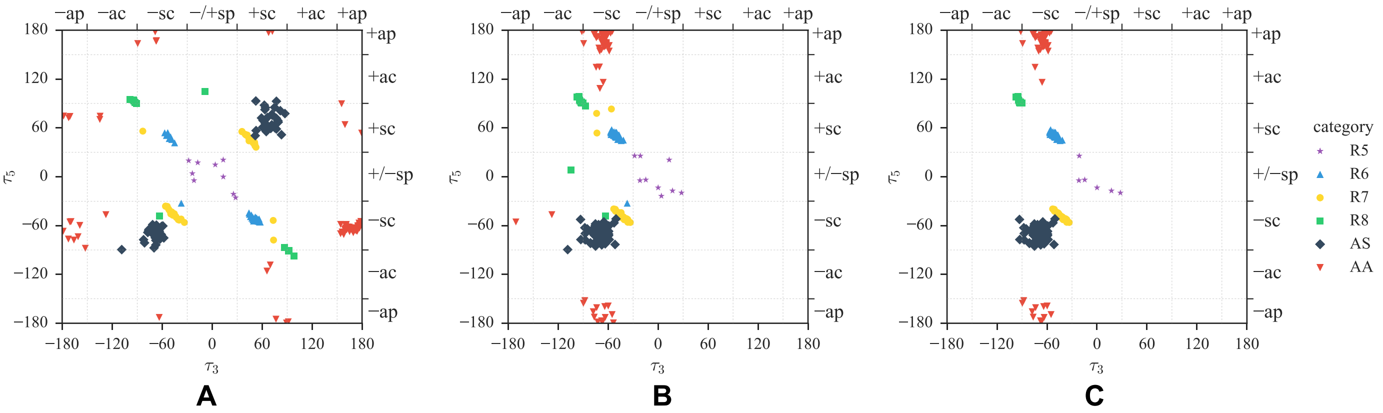

Figure 2. C-O-P-O-C Torsion angle distribution. The τ_3 (x-axis) and τ_5 (y-axis) torsion angles (°), with subdivision into categories, marked by symbols explained in the legend. The examples of the phosphodiester group found in the CSD are shown as (**A**) the raw values retrieved by the initial query; (**B**) after removing the labeling ambiguity; and (**C**) after the outlier rejection procedure. The τ_3 and τ_5 angles are defined in Figure 1. The torsion angle ranges (dash lines) are according to IUB.

Table 2. Division of the CSD C-O-PO$_2$-O-C fragments into: classes, categories and groups

Class	Category	Groups			
R	*5*	*–sp/-sp*	*+sp/+sp*		
	6	*–sc/+sc*	*+sc/–sc*		
	7	*–sc/–sc*	*+sc/+sc*		
	8	*–c/+c*	*+c/–c*		
A	*S*	*–sc/–sc*	*+sc/+sc*		
	A	*ap/–sc*	*–sc/ap*	*ap/+sc*	*+sc/ap*

ple within a given category to conform to a unique situation. We note here the obvious fact that the labeling ambiguity affects the atom and angle order, but not the numerical values. The idea of the relabeling method is to find such an atom order in each case (example) that will give the most consistent values for the same bond angle labels in each category. This way, we reduce a category to a single permutation and are able to calculate the means and standard deviations of the bond angles and bond distances. Such means and standard deviations are more accurate as they are calculated using all examples within a category.

To explain the above procedure, let us analyze an example. Suppose we have correctly assigned CSD examples to the *AA* category. Due to the labeling ambiguity, we have four possible permutations within *AA* and we are unable to distinguish their bond angles in advance. Therefore, for each example we choose such an atom labeling scheme that will give the smallest Euclidean distance to the mean values of the bond angles φ_{12}, φ_{13}, φ_{15}, φ_{23}, φ_{25}, φ_{35} within the *AA* category. At the beginning, we need a starting mean value for each of these angles, and for this purpose we select one of the four ambiguously labeled angles (φ_{13}, φ_{15}, φ_{23} or φ_{25}) as a reference angle. For instance, if we use φ_{13} as the reference, the first step of the procedure for each example will permute the labels in such a way that φ_{13} is the smallest angle. By setting (arbitrarily) φ_{13} to be the smallest angle, we can obtain a good approximation of the mean φ angle values for one of the four permutations. Next, the procedure iteratively refines this approximation by, once again, permuting the atom labels of each example and choosing the permutation that minimizes the Euclidean distance to the mean angles φ_{12}, φ_{13}, φ_{15}, φ_{23}, φ_{25}, φ_{35} from the previous step. Because the mean values and standard deviations may be distorted if the category contains low-quality or noisy examples, the mean values for the next step are calculated only for a subset of examples that are not marked as outliers (Figure 2B and C). The minimization procedure is repeated until all φ angle mean values converge, i.e. their values no longer change. This iterative procedure can be considered to be a variation of the k-means algorithm (7).

The above procedure is repeated in each category for all four label permutations, so that each of the φ_{13}, φ_{15}, φ_{23}, φ_{25} angles is the smallest. Therefore, for each of the six categories (*R5, R6, R7, R8, AS, AA*) we obtain four restraint sets (Supplementary Table S2). It may seem that it is hard to choose the proper permutation in practice. However, based on the torsion angles it is possible to select exactly one permutation in the *AA* category, since one group matches one permutation. Although in the *AS* category one group matches two permutations, based on the analysis of the NDB data we were able to determine the default values for this category (highlighted in bold in Supplementary Table S2).

In practical applications of the restraints, toward the end of structure refinement, we recommend to select the permu-

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

tation that fits the current geometry in the best way (i.e. has the smallest Euclidean distance to the φ angles); such a selection can be performed automatically using the web server presented in this paper. A practical example illustrating the use of the proposed restraints during a crystal structure refinement is presented in the following sections.

Machine learning methods for automatic clustering

To validate the proposed phosphodiester categorization in an objective way (i.e. without human prejudice, positive or negative), we used standard machine learning procedures to automatically cluster the data retrieved from the CSD. This alternative and independent analysis, when compared with the above manual approach, allowed us to verify if the patterns of PO_4 geometry can be detected automatically, using only the bond angles as the categorization criteria.

Machine learning is a field of computer science that develops algorithms that learn from and make predictions on data. One of the most common tasks in machine learning is *clustering*, where the goal is to partition a set of examples into meaningful groups of similar objects, called *clusters* (7,16). Clustering algorithms are divided into four types, as *partitioning*, *hierarchical*, *density-based* and *graph-based* approaches (7). To validate the proposed phosphodiester categorization, we investigated the use of four different algorithms, called PAM, AHC, DBSCAN and SC, representing, respectively, the four types of clustering approaches.

Partitioning Around Medoids (PAM) is a clustering algorithm that minimizes the sum of distances of objects within k clusters (17). PAM is a modification of the k-means method that assigns examples to medoids (medians) instead of centroids (means).

Agglomerative Hierarchical Clustering (AHC) is one of the most popular hierarchical approaches (7). The algorithm joins the closest examples (preliminary clusters), one by one, into k larger clusters. For this purpose, the algorithm requires a *merging* strategy that defines how distance is calculated for clusters with more than one example. Since AHC is sensitive to noise, we additionally extended the algorithm by preliminary anomaly detection, using the method of Loreiro *et al.* (18), which performs a hierarchical clustering of the data into nc groups and removes those that have less than t examples. As our goal was to eliminate only the most distant outliers, the parameters used were always $nc = D/10$ and $t = 2$, where D is the number of examples.

Density-Based Spatial Clustering for Applications with Noise (DBSCAN) is an algorithm that locates regions of high density that are separated by regions of low density (19). DBSCAN is one of the few algorithms that detect outliers while clustering, and for this purpose requires the definition of a neighborhood *Eps* (maximum distance between neighbors), and the minimum number of examples required to form a cluster *MinPts*.

Spectral clustering (SC) can be considered a representative of graph-based clustering approaches (20). It creates a similarity graph where edges connect an example and its p nearest neighbors. Next, SC creates k clusters by finding the minimal number of edge-cuts that divide the similarity graph into k separate graphs.

As input, all four algorithms used a dissimilarity matrix, which defines the distance between each pair of examples. The distance between a pair of PO_4 fragments was defined as the Euclidean distance between the corresponding O-P-O bond angles (φ_{12}, φ_{13}, φ_{15}, φ_{23}, φ_{25}, φ_{35}). Owing to the atom labeling ambiguity, this actually gives four possible values, defined by the four possible label permutations. In our implementation, we chose the smallest of these four values as the distance between a pair of fragments.

The algorithms were tested with several sets of parameters to find the clustering with the best silhouette coefficient. *Silhouette coefficient* (21) is an evaluation measure which promotes cohesive and well-separated clusters. Apart from optimizing the silhouette coefficient, we additionally rejected clusterings that removed more than 20% of the data as outliers. This restriction was introduced because DBSCAN would mark even 95% of the data as outliers for some parametrizations. The parameter optimization involved finding the best: number of clusters $k \in [2, 12]$ for PAM, AHC and SC; $Eps \in [0.05, 1.20]$ and $MinPts \in [2, 4]$ for DBSCAN; *merging* ∈ {average, maximum} for AHC; and $p \in \{5, 7, 10\}$ for SC. Finally, the clustering selected by the parameter optimization procedure was refined by outlier removal using the M_i test.

RESULTS AND DISCUSSION

The proposed geometry restraints

The new phosphodiester restraints have been derived from our manual statistical analysis of 238 C-O-PO_2-O-C fragments retrieved from the CSD, with additional data used for the O3′/O5′ linkages and terminal phosphates, as specified in Supplementary Table S1. The geometrical parameters τ (torsion angles) and φ (bond angles) are defined in Figure 1, whereas the conformational categories and groups are summarized in Table 2.

The resulting mean values for the torsion angles, bond angles and bond lengths, are summarized in Table 3. The P-O3′, O3′-C3′ and P-O5′, O5′-C5′ bond distances and φ_{33}, φ_{55} bond angles should be taken from Table 4. A full list of the restraints, taking into account all possible permutations, can be found in Supplementary Table S2.

The mean values from the Parkinson library (1) are mostly to be compared with the values found in the *AS* category, as the mean values of the τ_3, τ_5, φ_{12} and φ_{35} angles are very similar in both categorizations. However, the angles φ_{13}, φ_{15}, φ_{23}, φ_{25} in the Parkinson library are all ∼108°, which is approximately the mean value of the two smaller (104.5, 105.2°) and two larger (110.3, 111.5°) angles from the *AS* category. Gelbin *et al.* noticed one more value for the τ_3 torsion angle (163.1°) and differentiated between the *small* and *large* values for the $\varphi_{13}/\varphi_{23}$ (105.2/110.5°) and $\varphi_{15}/\varphi_{25}$ (105.7/110.7°) angles. The values are roughly in agreement with those found in the *AS* category, but the previous authors were unable to correlate the placement of the *small* or *large* angles with the sign and values of the τ_3 and τ_5 torsion angles.

The *AA* category was not recognized by any of the previous authors. The mean values of the φ_{12}, φ_{35} angles in the *AA* category are smaller than in the *AS* category (Table 3). The change in conformation also affects the φ_{13}, φ_{15}, φ_{23},

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

Table 3. CSD-derived mean values and standard deviations (in parentheses, in units of the last significant digit of the mean value) for the torsion and bond angles (in °) and for the P-O1, P-O2, P-O3′, P-O5′ bond distances (in Å), for the *R5*, *R6*, *R7 R8*, *AS*, *AA* categories found by manual classification (upper part) and for clusters A–D found by the best machine learning algorithm AHC with the following parameters: $k = 4$, *merge* = average (lower part)

	N	τ_3 (ζ)	τ_5 (α)	φ_{12}	φ_{13}	φ_{15}	φ_{23}	φ_{25}	φ_{35}	φ_{33}	φ_{55}	P-O1	P-O2	P-O3′	P-O5′
Parkinson		−70.8(48) 80.7(143)	−74.7(98) 81.0(121)	119.6(15)	107.7(32)	108.1(29)	108.3(32)	108.3(27)	104.0(19)	119.7(12)	120.9(16)	1.485(17)	1.485(17)	1.607(12)	1.593(10)
Gelbin	13-30*	−70.8(48)$^{-sc}$ 80.7(143)$^{+sc}$ 163.1(6)^{ap}	-74.7(98)$^{-sc}$ 81.0(121)$^{+sc}$	119.6(15)	110.5(11)^{L} 105.2(22)^{S}	110.7(12)^{L} 105.7(9)^{S}	110.5(11)^{L} 105.2(22)^{S}	110.7(12)^{L} 105.7(9)^{S}	104.0(19)	119.7(12)	120.9(16)	1.485(17)	1.485(17)	1.607(12)	1.593(10)
Category															
R5	6 [9]	−2(19)	-6(15)	117.2(8)	109.3(7)	111.1(10)	110.8(6)	109.9(5)	96.6(4)	111.3(18)	111.9(6)	1.483(4)	1.480(4)	1.613(8)	1.607(3)
R6	35 [36]	−51(4)	52(3)	118.4(11)	107.0(7)	108.0(5)	109.5(7)	109.9(9)	102.9(9)	115.8(25)	116.2(27)	1.483(9)	1.488(7)	1.604(10)	1.600(8)
R7	39 [45]	−44(5)	-48(5)	119.7(12)	104.9(9)	110.8(6)	111.9(6)	106.3(5)	101.9(9)	119.0(21)	117.7(17)	1.482(7)	1.472(9)	1.616(12)	1.616(8)
R8	9 [12]	−94(2)	94(3)	120.5(8)	105.1(6)	105.7(4)	109.8(5)	109.4(6)	105.2(5)	122.7(14)	122.4(15)	1.472(8)	1.477(6)	1.605(7)	1.607(7)
AS	55 [60]	−70(10)	-69(9)	119.9(16)	104.5(9)	110.3(8)	111.5(11)	105.2(8)	104.2(15)	121.5(30)	121.6(28)	1.484(12)	1.478(10)	1.599(16)	1.601(16)
AA	63 [76]	−67(7)	171(15)	117.6(12)	106.2(11)	110.2(13)	112.2(10)	109.3(9)	99.9(7)	120.8(11)	120.3(19)	1.487(10)	1.483(10)	1.580(10)	1.603(11)
Cluster															
Cluster A (*R6*/*R8*)	50 [52]	−55(35)	53(37)	118.9(13)	106.5(10)	107.6(11)	109.6(7)	109.8(10)	103.3(12)	118.0(42)	118.3(42)	1.480(10)	1.485(9)	1.603(10)	1.600(10)
Cluster B (*AS*/*R7*)	95 [101]	−59(16)	-60(13)	119.8(15)	104.6(9)	110.7(8)	111.6(10)	105.8(8)	103.2(17)	120.5(29)	120.0(31)	1.483(11)	1.476(10)	1.607(17)	1.606(16)
Cluster C (*AA*)	61 [70]	−66(7)	172(13)	117.6(12)	106.2(11)	110.2(13)	112.3(10)	109.2(07)	99.9(6)	120.8(11)	120.2(17)	1.487(10)	1.483(10)	1.579(10)	1.603(11)
Cluster D (*R5*)	6 [8]	−7(24)	1(20)	117.2(9)	109.5(5)	110.5(8)	111.1(8)	110.1(9)	96.4 (5)	110.9(16)	111.9(7)	1.483(5)	1.479(8)	1.611(5)	1.610(8)

For each cluster, the matching manual categories are given *in italics* in parentheses. The preferred values for the P-O3′/P-O5′/O3′-C3′/O5′-C5′ bond lengths and the $\varphi_{33,}$ φ_{55} angles for the *AS* and *AA* categories are in Table 4. N is the number of cases [number of cases before outlier rejection in square brackets]. The first two rows contain reference values from the Parkinson *et al.* (1) and Gelbin *et al.* (3) libraries. The following superscript annotations are used: ^{L}-large angle, ^{S}-small angle, $^{-sc}/^{+sc}/^{ap}$- ±synclinal or antiperiplanar conformation; *for each column a different number of examples was used by the original authors.

φ_{25} angles, as three out of the four angles in the *AA* category are larger than the corresponding angles in the *AS* category, and only the φ_{15} angle is about the same.

The *acyclic* class is the most interesting one from the point of view of restraints for the refinement of nucleic acid structures, but for generality and completeness we have calculated the mean values for the *R* class as well. In addition, those unusual cases may be actually quite useful for restraint definitions in exotic nucleic acid structures (22).

Since in the *R* class the O3′ and O5′ atoms are constrained by the ring structure, the O3′-P-O5′ angle shows high variability, from 96.6° in *R*5 to 105.2° in *R*8. In the *R6*/*R7* rings φ_{35} has intermediate values (102.9/101.9°). The φ_{12} angle increases with the ring size from 117.2° in *R5* to 120.5° in the *R8* category. The phosphate groups in the *R6* and *R8* categories have similar conformation since the torsion angles have the same combination of signs. On the other hand, the *R7* category is more similar to the *AS* category. A comparison of all the bond angles with reference to the τ_3 and τ_5 torsion angles is presented in Supplementary Figure S2A.

The proposed P-O3′, O3′-C3′ and P-O5′, O5′-C5′ restraints (Table 4) were calculated for the *AS* and *AA* categories separately. The O3′-C3′/O5′-C5′ bond lengths are between 1.422 and 1.438 Å. Unfortunately, the numbers of independent examples for the P-O3′, O3′-C3′ bonds in the *AS* category and for P-O5′, O5′-C5′ in the *AA* category are very limited (3 and 2, respectively); thus, the mean values may not be very accurate.

Supplementary Figure S2A provides clear evidence that in the phosphodiester moiety there is no functional relation between the torsion angles and the bond angles, i.e. the bond angles φ_{xy} cannot be computed using analytical functions of the torsion angles in a way similar to that used for conformation-dependent restraints in proteins (10–12). Instead, the proposed categories define six coherent and mostly well-separated clusters; a similar situation exists for the bond distances (Supplementary Figure S4A). Moreover, the histograms in Supplementary Figures S6 and S8 demonstrate that the bond angles are practically normally distributed within each category.

Automatic clustering of the phosphodiester groups

The goal of the automatic clustering experiment was to validate the manual categorization in a prejudice-free manner. In contrast to the manual categorization, where additional chemical knowledge was used, the non-supervised machine learning methods do not understand the meaning of chemical bonds or conformations. The best results, in terms of the silhouette coefficient, were obtained using the AHC algorithm with $k = 4$ and *merging* = average. The resulting clusters are presented in Table 3 and illustrated in Supplementary Figures S2B and S4B. Even though it is hard to generalize the results for other types of moieties or systems of atoms, there are several important conclusions from this exercise.

In the phosphodiester moiety, almost all geometrical differences are reflected in the angular parameters, while most of the bond distances remain unchanged upon torsion-angle variations. Thus, we were able to measure the similarity of pairs of fragments by the use of bond angles only.

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

Table 4. Mean values and standard deviations (in parentheses, in units of the last significant digit of the mean value) for the P-O3′, O3′-C3′ and P-O5′, O5′-C5′ bond lengths (Å) and the φ_{33} and φ_{55} angles (°) for PO_4-ribose and PO_4-deoxyribose structures in the CSD

	N	P-O3′	O3′-C3′	φ_{33}	N	P-O5′	O5′-C5′	φ_{55}
Parkinson	-	1.607(12)	1.423(14)^R 1.431(13)^D	119.7(12)	-	1.593(10)	1.440(16)	120.9(16)
Gelbin	8	1.607(12)	1.433(19)	119.7(12)	6	1.593(10)	1.440(16)	120.9(16)
Category								
AS	3 [7]	1.603(14)	1.438(7)	120.7(29)	13 [13]	1.594(9)	1.437(17)	119.3(15)
AA	42 [55]	1.601(8)	1.422(10)	120.2(15)	2 [3]	1.591(4)	1.428(13)	121.7(30)

N is the number of cases [number of cases before outlier rejection in square brackets]. The superscripts denote: ^{R}-ribose, ^{D}-2′-deoxyribose.

Such an approach may be applicable to other systems of atoms.

The SC and DBSCAN algorithms did not find reasonable clusters for the phosphodiester moiety. One of the reasons may be that the data set was contaminated by a high number of outliers. In our manual clustering, we marked as outliers and excluded ~13% of the cases. This is of special note, as the data were harvested from high-quality small-molecule CSD crystal structures. One can expect that with less accurate data, e.g. from the PDB, the fraction of outliers would be even higher (23). Another possible reason for the poor clustering results of SC and DBSCAN may be the distribution of the angle values. The clusters in the analyzed data are not always clearly separated and the number of examples in each cluster is highly variable. In the manual clustering experiment, the number of examples in the categories was from 6 to 63, and the best AHC clusters have from 6 to 95 examples. In conclusion, graph- and density-based algorithms should be aided by additional outlier detection and cluster imbalance methods, or not considered as the first choice for clustering structural chemistry data. PAM and AHC, on the other hand, were clearly superior at finding unevenly represented clusters, with AHC slightly better in handling outliers.

In our experience, in order to obtain good clusters, handling of outliers is crucial. Attempts at automatic clustering without outlier rejection usually led to very high numbers of clusters or strongly biased the results. Handling of outliers starts with very careful selection of the input data. The geometrical data should, therefore, come from high-quality structures.

The optimal number of clusters was chosen based on the silhouette coefficient. For the AHC algorithm, the best value of the coefficient was obtained for four clusters (*A*–*D*). In the manual clustering on the other hand, we distinguished six categories. This discrepancy might appear to question the agreement of the two approaches. However, in the manual clustering, we mainly focused on the analysis of the τ_3 and τ_5 torsion angles, and we used additional information about the topology (circular, linear) of the analyzed fragments. The AHC algorithm identified differences between the basic geometries, also found by the manual clustering (compare Supplementary Figure S2A and S2B). The AHC *cluster A* contains examples mostly from the *R6* and *R8* categories, *cluster B* from the *R7* and *AS* categories and *cluster C* from the *AA* category. Additionally, AHC separated the *R5* category (*cluster D)* from the other clusters although it has only six examples after outlier removal. Post analysis of the mean values obtained by the manual clustering revealed that the categories *R6* and *R8*, as well as *R7* and *AS*, are very similar, thus reducing the effective number of significantly different categories from six to four. From this point of view, we conclude that the automatic clustering by the AHC algorithm has been indeed quite successful.

In this study, we have exploited the manual categorization and used automatic clustering only for validation, because we had additional structural knowledge about the torsion angles and whether a given example was cyclic. However, in situations where such additional information is not available or is too complex for manual processing, the presented automatic clustering methodology should be capable of finding reasonable groupings. In conclusion, we recommend to further investigate the applicability of the proposed automatic clustering methodology to other molecular fragments and to consider it as an alternative to the most commonly used k-means algorithm.

We note that our use of automatic clustering was partially inspired by the classification of Zn coordination geometries presented by Yao *et al.* (24), even though the validity of the structural conclusions drawn by those authors has been questioned (25). However, in our method, we only use clustering algorithms, whereas Yao *et al*. proposed a complex processing pipeline that mixes clustering algorithms with classifiers and requires *a priori* knowledge of the data. We also provide an explicit definition of the distance between pairs of fragments, which was problematic in the paper by Yao *et al.* Furthermore, we searched for the best partitioning in terms of the silhouette coefficient, which optimizes cluster cohesion and separation. In summary, we believe that the automatic clustering procedure described in this paper presents a more general, straightforward and superior approach, applicable to other structural problems.

The case of terminal phosphomonoester groups

To complement the phosphodiester geometry, we also analyzed the geometry of phosphomonoesters, $C\text{-}O\text{-}PO_3^{2-}$, which in nucleic acids are normally attached at the terminal C5′ or C3′ atoms. The clustering of the $C\text{-}O\text{-}PO_3^{2-}$ geometry was carried out exclusively by the AHC machine learning algorithm, and is in a sense a test ground for this approach.

Chemical intuition would suggest that—linked at a single point of attachment—the phosphomonoester group will have covalent geometry largely independent of the type (label) of the C atom. In the following analysis of the phospho-

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

monoester geometry we have therefore labeled the C atom, for simplicity, as C5′. For terminal phosphate groups attached to other atoms (in particular C3′), the numbering of atoms in the restraint Supplementary Table S3 should be adjusted accordingly. Atom labeling in a terminal phosphate is presented in Supplementary Figure S1D. It is noted that this labeling is ambiguous in practice, as the O1, O2 and O3 atoms are usually labeled randomly.

The torsion angles C5′-O5′-P-Ox, where x = 1, 2, 3, were found to be either around ±60°/180°, or close to ±120°/0°. Intuitively, one would expect two types of covalent geometry, clustered with the, e.g. C5′-O5′-P-O3 torsion angle close to 180° or 0°. However, automatic clustering, analogous to that used for C-O-PO_2-O-C, revealed that there is only one cluster. Indeed, the problem lies in proper labeling of the O3 atom. Labeling the O3 atom consistently to make the C5′-O5′-P-O3 torsion angle the closest to 180° (in both groups), one gets a unique representation of the C-O-PO_3^{2-} geometry. Making C5′-O5′-P-O3 close to 180° is equivalent to labeling as O3 the oxygen atom that is farthest from C5′. The O1/O2 labels were assigned according to Fischer convention with the same caveat about ambiguity as for the phosphodiester group.

The mean values and standard deviations recommended as restraints for terminal phosphomonoesters are summarized in Supplementary Table S3. As expected, even though the O3 atom was singled out, the terminal P-O1/P-O2/P-O3 bonds are very similar (1.514(9)/1.520(9)/1.514(10) Å) while the P-O5′ bond is significantly longer (1.622(9) Å). The bond-angle geometry is more complicated, and reflects the symmetry (of O1/O2) relative to the uniquely defined O5′/O3 atoms. The O1-P-O5′ and O2-P-O5′ angles are comparable and smaller (107.5(7), 107.2(8)°) than O1-P-O3/O2-P-O3 (114.0(7)/112.8(10)°). The mean values of the remaining angles, O3-P-O5′, O1-P-O2 and P-O5′-C5′, span a large interval, 102.9(12), 111.7(10) and 119.0(22)°, respectively. In practice, in our application server we consider six possible permutations of the proposed restraints (Supplementary Table S4), because models in the PDB label the O1, O2, O3 atoms arbitrarily.

Validation of the restraints using NDB data

To further validate the derived mean values of the PO_4 geometry and to confirm that they can be used as restraints for the phosphodiester moieties of nucleic acids, we selected high-quality oligonucleotide structures, mostly from the NDB, and calculated for this sample the mean values for the torsion/bond angles and bond lengths within the categories and groups defined in Table 2.

The NDB sample means and standard deviations, after assigning the examples to the two groups in the *AS* category (*–sc*/*–sc*, +*sc*/+*sc*) and to three out of the four groups in the *AA* category (*ap*/*–sc*, *–sc*/*ap*, +*sc*/*ap*), are summarized in Supplementary Table S5. We did not find any cases in the NDB that would fall into the *AA* (*ap*/+*sc*) group. This may be a consequence of the still rather small number of PO_4 moieties in the NDB sample, or of the fact that the CSD sample represented a broader range of fragment conformations than what has been actually observed in nucleic acid structures to date.

We compare the quality of the proposed restraints with those of Gelbin *et al.* (3) in a histogram of absolute differences of bond angles and bond lengths, for each entry in the NDB set (Supplementary Figure S7). It shows that the angle restraints have been improved whereas the deviations of bond lengths are comparable. This confirms that the proposed categorization and the resulting sets of restraints can be safely used for the refinement of nucleic acid structures. The distributions of the bond angles and bond distances for the CSD and NDB data are compared in Supplementary Figures S2A and S3 (angles) and Supplementary Figures S4A and S5 (distances). We note that atom labels of the NDB fragments were not permuted since atom labeling is uniquely defined in this database.

The NDB τ_3/τ_5 torsion angles that are in the *ap* interval have a bimodal distribution, clustering around −140 and +170° (Supplementary Figure S3). No such division was observed for the data obtained from the CSD, which indicates that the torsion angle at −140° is characteristic of nucleic acids. In future, it may be necessary to further divide the *AA* category into eight groups instead of four, but at present there are not enough examples to calculate reliable statistics for such a subdivision, and such structures are completely absent from the CSD.

While to the first approximation the covalent geometry of organic molecules is considered to be robust and relatively constant, numerous analyses have shown that on closer scrutiny the covalent architecture is correlated with conformation. The variations are more distinct in bond angles, which are more easily deformed, but are also detectable in bond lengths. Similar conclusions have been reached for macromolecules. The present analysis, so far limited to the phosphodiester group, confirms that the same is true for nucleic acids.

The ultrahigh-resolution PDB structure 3P4J

The proposed new geometrical restraints for the phosphodiester moiety (Table 3) were compared with the corresponding geometrical parameters found in the ultrahigh-resolution (0.55 Å), unrestrained Z-DNA structure (4) with the PDB code 3P4J. In the 3P4J model there are 10 PO_4 moieties with 3 different combinations of the τ_3/τ_5 angles (Supplementary Table S6), representing the following groups: *AS*(+*sc*/+*sc*), *AA*(*–sc*/*ap*) and *AA*(+*sc*/*ap*). When the 3P4J model is compared with the currently used stereochemical targets, as included in REFMAC (26,27), the RMSD(angles) is 1.60° and it drops to 1.29° when the new phosphodiester targets are used.

It is important to stress that the 3P4J model was refined using the method of least-squares and without any geometrical restraints (4). The final geometrical parameters are therefore not biased by any *a priori* assumptions and are characterized not only by their numerical values but also by their estimated standard deviations.

Practical example

As an example, we present re-refinement of the 1.95 Å resolution protein–DNA complex (nuclear receptor bound to its response element hsp27 gene promoter) deposited in

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

the PDB with the accession code 2HAN (28). The 2HAN model was originally refined using REFMAC version 5.1.24 (26,27) and was published with R/R_{free} of 18.0/21.7% and RMSD values of 0.016 Å for bond lengths and 2.03° for bond angles (Table 5). That refinement used standard protein (29) and nucleic acid restraints, as included in REFMAC from the CCP4 suite (30). We note that the asymmetric unit of the 2HAN structure is composed of 165 amino acid residues and 40 nucleotides, with 38 phosphodiester groups, plus 4 zinc ions and 222 water molecules. In our re-refinement, we used 30 iterations of REFMAC version 5.8.135 as included in CCP4 version 7.0.011.

First, we re-refined the structure with the REFMAC built-in restraints to obtain reference R/R_{free} of 19.57/22.41% and RMSD(bonds)/(angles) of 0.0200 Å/2.0958°. These values differ from those originally deposited due to changes between REFMAC versions but they are in agreement with the values reported in the current PDB validation report for 2HAN.

Next, we re-refined the 2HAN model again, this time with the option of providing custom external structural restraints (31). The external restraint file was generated using our web server (vide infra). With external restraints, the refinement results strongly depend on the adjustable parameters w, w_{ext} and κ, where w weights the contribution of the experimental data, w_{ext} adjusts the weights of the external restraints relative to other geometry components, and κ is the Geman–McClure robust estimation function parameter (31). In order to make the refinements with and without the proposed new PO_4 restraints comparable, we analyzed the impact of w, w_{ext} and κ on the R factor, R_{free} and RMSD(bonds).

Figure 3 shows the re-refinement results versus variation of the weighting parameters (w, w_{ext}), while Table 5 presents the statistics obtained using the optimal parameters. To find comparable w, w_{ext} and κ values for the refinements with and without external restraints, we first find the w value that gives RMSD(bonds) of ∼0.02 Å (green dash line in Figure 3A). This fixes the only parameter that is common to both refinements. Next, we select such a value of w_{ext} that gives the best R and R_{free}, and keeps RMSD(bonds) at 0.02 Å (solid green line in Figure 3B). In the analyzed example, changing κ between 0.5 and 5 had no effect on model statistics. Figure 3 illustrates the well-known fact that there is a trade-off between low R/R_{free} values and low RMSD from ideality, depending on the weight w, which defines the relative contributions of the experimental data and stereochemical information. Nevertheless, the presented plots show that with the new PO_4 restraints the model converges at better RMSD values for the same R/R_{free} (better agreement with the stereochemical targets without degrading the agreement with the experimental data), or better R/R_{free} factors at the same RMSD(bonds) level (the same stereochemical model quality with better agreement with the experimental data). This improvement is quite significant if we take into account that we have updated restraints for only 38 phosphodiester groups (228 bonds, 304 angles) in a 2387-atom structure with 3960 bond and 7115 angle restraints.

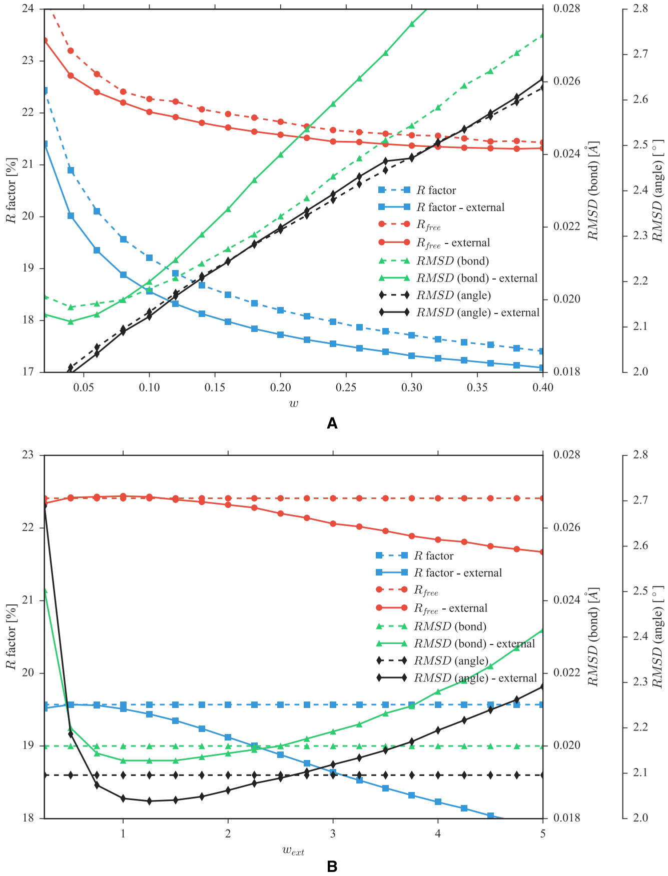

Figure 3. 2HAN Refinement parameters. R factor (blue lines, left y-axis), R_{free} (red lines, left y-axis) and RMSD(bonds) (green lines, right y-axis), RMSD(angles) (black lines, right y-axis) after 30 REFMAC iterations for the 2HAN model without (dashed lines) and with (solid lines) external PO_4 restrains, plotted against: (**A**) the weight parameter w ($w_{\mathrm{ext}} = 2.5$, $\kappa = 0.5$); (**B**) external weight scale w_{ext} ($w = 0.08$, $\kappa = 0.5$).

The two final models, after re-refinement with the standard restraints (R = 19.57%) and with the proposed restraints ($R = 18.88\%$) are comparable. The geometrical differences between these two models are consistent with the changes in the restraint targets and are, on average, 1.6° (φ angles) 2.4° (τ angles) and 0.010 Å (bonds). The largest bond angle difference was found for φ_{15} (6.8°) of residue 10, chain C, whereas the largest difference in torsion angles occurred at residue 6D, where τ_3 changed by 13.3° and was coupled to a visible change of the sugar pucker (as gauged by the pseudorotation parameters P/τ_m) from 145.0(7)/45.0(6)° to 123.5(6)/39.7(4)° [note the decrease of the standard deviations, estimated by the method of Jaskolski (32)].

The presented example of a re-refinement of a previously deposited model is not necessarily a typical application, since—in normal situations—the proposed restraints would also be used to aid the earlier stages of model building. Moreover, in practice it might be worthwhile to regenerate the external restraints after each refinement run, as entities may change their category assignment.

PO_4 restraint web server

A dedicated web server has been created for the generation of external PO_4 restraints for use in REFMAC, by analyz-

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

Table 5. *R*, R_{free}, RMSD(bonds) and RMSD(angles) for the 2HAN model deposited in the PDB and re-refined with REFMAC with and without external restraints

Refinement	R (%)	R_{free} (%)	RMSD(bonds) (Å)	RMSD(angles) (°)
2HAN as deposited in PDB†	18.0 (18.58)#	21.7 (22.01)#	0.016 (0.0178)#	2.029 (2.1037)#
2HAN without external restraints ($w = 0.08$)*	19.57	22.41	0.0200	2.0958
2HAN using external restraints ($w = 0.08$, $w_{\text{ext}} = 1.5$, $\kappa = 0.5$)*	19.35	22.39	**0.0196**	**2.0409**
2HAN using external restraints ($w = 0.08$, $w_{\text{ext}} = 2.5$, $\kappa = 0.5$)*	**18.88**	**22.20**	0.0200	2.0894

†REFMAC version 5.1.024. *REFMAC version 5.8.135. #Values in parentheses as reported by a zeroth cycle of REFMAC version 5.8.135.

ing each phosphodiester moiety in the input .pdb file and selecting the proper restraint category and permutation based on the values presented in Supplementary Table S2. More precisely, we apply a restraint directly if the analyzed moiety has its τ_3/τ_5 torsion angles within 4 standard deviations of a single value in Supplementary Table S2. If more than one restraint could be applied (within 4 standard deviations), we choose the one with the smallest Euclidean distance from the bond angles φ_{12}, φ_{13}, φ_{15}, φ_{23}, φ_{25}, φ_{35}. If the τ_3/τ_5 torsion angles are more than 4 standard deviations from all of the values in Supplementary Table S2, we take the restraint with the smallest Euclidean distance to the τ_3/τ_5 angles.

The server is freely available at http://achesym.ibch.poznan.pl/restraintlib/. It is recommended to regenerate the restraints every few cycles, as the phosphate groups may change their category during the refinement.

CONCLUSIONS AND OUTLOOK

In the present work, we have revisited the standard restraints for the valence geometry of the phospho-di/monoester groups, with the purpose of proposing improved stereochemical targets for the refinement of nucleic acid structures. The new restraints are based on data harvested from the CSD and manually clustered into six categories: four cyclic (*R5*, *R6*, *R7*, *R8*) and two linear ones (*AS*, *AA*). In contrast to previous studies (1,3), we were able to correlate the O-P-O angles with the τ_3/τ_5 torsion angles of the phosphodiester backbone. Moreover, the *AA* category had not been described in the previous compilations. The proposed categorization was successfully validated using machine learning methods. When compared against NDB data and an ultrahigh resolution PDB structure, the new restraints are superior to those of Gelbin *et al.* (3). In a practical example, it was possible to reduce *R* and R_{free}, while preserving the same level of deviations of bond distances and angles from the targets. To facilitate the use of the proposed restraints, we have implemented a publicly available application server that automatically creates the PO_4 restraints for a given .pdb file.

This study is part of a wider program of revision of nucleic acid stereochemical restraints. Our future studies, already in progress, will deal with revised ribose and glycosidic stereochemistry. As more accurate examples are accumulated in the databases, we may revisit the geometry of the PO_4 group again, to deal with the possible further subdivision of the *AA* and *AS* categories.

Finally, we see great promise in the automated approach to database analysis and restraint discovery with the application of machine learning. The feasibility of this approach has been demonstrated as proof of principle in the present study and will be exploited in our future work.

AVAILABILITY

The input data, source codes, grid search parameters and reproducible experiment scripts in Python are available for download from http://www.cs.put.poznan.pl/dbrzezinski/software.php. A dedicated web server for the generation of external PO_4 restraints is available at http://achesym.ibch.poznan.pl/restraintlib/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors thank Dr Mirek Gilski and Dr Pawel Drozdzal for help with REFMAC refinement.

FUNDING

National Science Centre (NCN) [2013/10/M/NZ1/00251 to M.J.]; FNP START scholarship and Institute of Computing Science Statutory Funds [to D.B.]. Funding for open access charge: KNOW Project.
Conflict of interest statement. None declared.

REFERENCES

1. Parkinson,G., Vojtechovsky,J., Clowney,L., Brünger,A.T. and Berman,H.M. (1996) New parameters for the refinement of nucleic acid-containing structures. *Acta Crystallogr. D Biol. Crystallogr.*, **D52**, 57–64.
2. Clowney,L., Jain,S.C., Srinivasan,A.R., Westbrook,J., Olson,W.K. and Berman,H.M. (1996) Geometric parameters in nucleic acids: Nitrogenous bases. *J. Am. Chem. Soc.*, **118**, 509–518.
3. Gelbin,A., Schneider,B., Clowney,L., Hsieh,S., Olson,W.K. and Berman,H.M. (1996) Geometric parameters in nucleic acids: Sugar and phosphate constituents. *J. Am. Chem. Soc.*, **118**, 519–529.
4. Brzezinski,K., Brzuszkiewicz,A., Dauter,M., Kubicki,M., Jaskolski,M. and Dauter,Z. (2011) High regularity of Z-DNA revealed by ultra high-resolution crystal structure at 0.55 Å. *Nucleic Acids Res.*, **39**, 6238–6248.
5. Sheldrick,G.M. (2008) A short history of SHELX. *Acta Crystallogr. A*, **A64**, 112–122.
6. Allen,F.H. (2002) The Cambridge structural database: A quarter of a million crystal structures and rising. *Acta Crystallogr. B*, **B58**, 380–388.
7. Tan,P.N., Steinbach,M. and Kumar,V. (2005) *Introduction to Data Mining*. Pearson Addison Wesley, Boston.
8. Narayanan,B.C., Westbrook,J., Ghosh,S., Petrov,A.I., Sweeney,B., Zirbel,C.L., Leontis,N.B. and Berman,H.M. (2013) The Nucleic Acid Database: New features and capabilities. *Nucleic Acids Res.*, **42**, D114–D122.
9. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
10. Tronrud,D.E. and Karplus,P.A. (2011) A conformation-dependent stereochemical library improves crystallographic refinement even at

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024

atomic resolution. *Acta Crystallogr. D Biol. Crystallogr.*, **D67**, 699–706.

11. Moriarty,N.W., Tronrud,D.E., Adams,P.D. and Karplus,P.A. (2014) Conformation-dependent backbone geometry restraints set a new standard for protein crystallographic refinement. *FEBS J.*, **281**, 4061–4071.
12. Moriarty,N.W., Tronrud,D.E., Adams,P.D. and Karplus,P.A. (2015) A new default restraint library for the protein backbone in Phenix: A conformation-dependent geometry goes mainstream. *Acta Crystallogr. D Struct. Biol.*, **D72**, 176–179.
13. Bruno,I.J., Cole,J.C., Edgington,P.R., Kessler,M., Macrae,C.F., McCabe,P., Pearson,J. and Taylor,R. (2002) New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallogr. B*, **B58**, 389–397.
14. Jaskolski,M. (2013) High resolution macromolecular crystallography. In: Read,R, Urzhumtsev,AG and Lunin,VY (eds). *Advancing Methods for Biomolecular Crystallography*. Springer, Dordrecht, pp. 259–275.
15. Iglewicz,B. and Hoaglin,D. (1993) *How to detect and handle outliers*. ASQC Quality Press, Milwaukee.
16. Han,J. and Kamber,M. (2000) *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco.
17. Kaufman,L. and Rousseeuw,P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, NY.
18. Loreiro,A., Torgo,L. and Soares,C. (2004) *Outlier detectioon using clustering methods: a data cleaning application*. Proc. Data Mining Business Workshop.
19. Ester,M., Kriegel,H.-P., Sander,J. and Xu,X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proc. of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, pp. 226–231.
20. Leskovec,J., Rajaraman,A. and Ullman,J.D. (2014) *Mining of Massive Datasets*, 2nd edn, Cambridge University Press, Cambridge.
21. Rousseeuw,P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
22. Palm,G.J., Billy,E., Filipowicz,W. and Wlodawer,A. (2000) Crystal structure of RNA 3′-terminal phosphate cyclase, an ubiquitous enzyme with unusual topology. *Structure*, **8**, 13–23.
23. Deller,M.C. and Rupp,B. (2015) Models of protein–ligand crystal structures: trust, but verify. *J. Comput. Aided. Mol. Des.*, **29**, 817–836.
24. Yao,S., Flight,R.M., Rouchka,E.C. and Moseley,H.N. (2015) A less-biased analysis of metalloproteins reveals novel zinc coordination geometries. *Proteins*, **83**, 1470–1487.
25. Raczynska,J., Wlodawer,A. and Jaskolski,M. (2016) Prior knowledge or freedom of interpretation? A critical look at a recently published classification of 'novel' Zn binding sites. *Proteins*, **84**, 770–776.
26. Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.*, **D53**, 240–255.
27. Murshudov,G.N., Skubák,P., Lebedev,A.A., Pannu,N.S., Steiner,R.A., Nicholls,R.A., Winn,M.D., Long,F. and Vagin,A.A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.*, **D67**, 355–367.
28. Jakob,M., Kolodziejczyk,R., Orlowski,M., Krzywda,S., Kowalska,A., Dutko-Gwozdz,J., Gwozdz,T., Kochman,M., Jaskolski,M. and Ozyhar,A. (2007) Novel DNA-binding element within the C-terminal extension of the nuclear receptor DNA-binding domain. *Nucleic Acids Res.*, **35**, 2705–2718.
29. Engh,R.A. and Huber,R. (2001) Structure quality and target parameters. In: Rossman,MG and Arnold,E (eds). *International Tables for Crystallography*, Kluwer Academic Publishers, Dordrecht, pp. 382–392.
30. Winn,M.D., Ballard,C.C., Cowtan,K.D., Dodson,E.J., Emsley,P., Evans,P.R., Keegan,R.M., Krissinel,E.B., Leslie,A.G., McCoy,A. *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.*, **D67**, 235–242.
31. Nicholls,R.A., Long,F. and Murshudov,G.N. (2012) Low-resolution refinement tools in REFMAC5. *Acta Crystallogr D Biol. Crystallogr.*, **D68**, 404–417.
32. Jaskolski,M. (1984) A comparison of two methods for the calculation of pseudorotation parameters. *Acta Crystallogr. A*, **A40**, 364–366.

Downloaded from https://academic.oup.com/nar/article/44/17/8479/2468064 by guest on 09 April 2024