Abstract: Obtaining high-quality models for nucleic acid structures by automated model building programs (AMB) is still a challenge. The main reasons are the rather low resolution of the diffraction data and the large number of rotatable bonds in the main chains. The application of the most popular and documented AMB programs (e.g., PHENIX.AUTOBUILD, NAUTILUS and ARP/wARP) may provide a good assessment of the state of the art. Quite recently, a cyclic automated model building (CAB) package was described; it is a new AMB approach that makes the use of BUCCANEER for protein model building cyclic without modifying its basic algorithms. The applications showed that CAB improves the efficiency of BUCCANEER. The success suggested an extension of CAB to nucleic acids—in particular, to check if cyclically including NAUTILUS in CAB may improve its effectiveness. To accomplish this task, CAB algorithms designed for protein model building were modified to adapt them to the nucleic acid crystallochemistry. CAB was tested using 29 nucleic acids (DNA and RNA fragments). The phase estimates obtained via molecular replacement (MR) techniques were automatically submitted to phase refinement and then used as input for CAB. The experimental results from CAB were compared with those obtained by NAUTILUS, ARP/wARP and PHENIX.AUTOBUILD.

Keywords: nucleic acids; molecular replacement; phase refinement; automated model building

1. Introduction

Automated model building (AMB) programs try to replace the visual interpretation of the three-dimensional electron density map, which is usually time consuming and subjective, with automatic procedures to speed up the structure determination process and to minimize the modelling errors.

Several successful and well-documented AMB programs are available for proteins (among others, we cite BUCCANEER [1], ARP/wARP [2], PHENIX.AUTOBUILD [3]). Equivalent tools for nucleic acids exist, but most of them are still in progress. Indeed, quite often, such AMB programs aid in detecting errors in crystallographic models [4], or extend and rebuild existing nucleotides chains [5] or perform semi-automatic building [6].

Because the number of solved nucleic acid structures is rapidly increasing, more efforts were spent recently on the specific difficulties in the electron density interpretation due to lower resolution data [7] and the large number of rotatable bonds in the main chain (two in the protein main chain, six in nucleic acids). As a consequence, the conformation at low resolution is often ambiguous, particularly for large nucleic acid structures, and is typically determined at resolutions worse than 2.5Å. It is not uncommon that phosphate and base planes are reliably located, but sugars and part of the backbone are not seen
at all [8]. Rebuilding and refining current models is often a time-consuming manual practice, so AMB promise to save time.

In spite of the above limitations, complete or almost complete AMB packages for nucleic acids exist: PHENIX.AUTOBUILD [3], ARP/wARP [9] and NAUTILUS [10]. These programs build nucleotide chains in a rather automatic manner. All such AMB procedures are based on an intensive use of prior crystallochemical knowledge. Indeed, nucleotides contain three rigid groups: the pentose sugar, the phosphate group and the base. Different programs use different tools for the AMB: some programs use the planarity of the base (e.g., in ARP/wARP), others exploit the sugar and the phosphate groups as the first tool for identifying the main chain (e.g., in NAUTILUS). All try to extend possible nucleotide chains and match the built chains to the nucleotide sequence. Models so found may be refined via REFMAC [11] or PHENIX.REFINE [12]; usually, calculations are iterated to obtain more complete models. It is typical that a good percentage of nucleotides are correctly built in the final model.

Recently, the cyclic automated model building (CAB) package for protein-automated model building was described [13]. CAB uses BUCCANEER in a cyclic procedure aimed at increasing its rate of success and the quality of the provided molecular models. In other words, CAB wraps around BUCCANEER. This program is itself cyclic; a standard BUCCANEER run performs five cycles of model building and 10 cycles of model refinement via REFMAC. CAB is highly automated and not very time consuming because BUCCANEER is fast, efficient, simple to use, and rather insensitive to the resolution limit of the data. CAB was tested over 81 protein structures solved via molecular replacement, anomalous dispersion and ab initio methods. The results showed that CAB gave more complete and accurate structures compared to the conventional use of BUCCANEER.

The success of CAB for proteins suggests that similar improvements may be made for nucleic acids. Here, we test whether or not the cyclic application of NAUTILUS by CAB improves the completeness and accuracy of the resulting nucleic acid structures. As was done for proteins, CAB does not modify the basic NAUTILUS algorithms at all; it has to be considered as a tool that allows the synergistic combination of NAUTILUS with some supplementary algorithms, offering more chances for the correct interpretation of the electron density maps. Indeed, in our experience, a starting set with the smallest average phase error is not always the most successful when an AMB program is applied; often, it is the variety of starting sets which allows CAB to improve NAUTILUS results.

The criteria we will use for making cyclic NAUTILUS applications cannot coincide with the criteria used for making cyclic BUCCANEER because of the strong differences between protein and nucleic acids crystallochemistry. We describe, in Section 2, the algorithms introduced into CAB for the location of the heavy atoms contained into ligands; in Section 3, we describe the recursive CAB algorithms for nucleic acids and, in Section 4, the experimental tests where we compare the results obtained by CAB with those obtained by NAUTILUS, ARP/wARP and PHENIX.AUTOBUILD.

2. CAB Algorithm for Locating Ligand Heavy Atoms

We suppose that a set of observed structure factor amplitudes with refined $\phi_r$ phases and $w_r$ weights are available as the starting point for any AMB application. They were first obtained by applying REMO09 [14] to the test structures, and then refined via the SYNERGY approach [15]. The name of the latter procedure arises from the fact that it combines mainstream phase refinement procedures (DM by Cowtan [16]) and out-of-mainstream phase refinement techniques. SYNERGY includes free lunch [17,18], low density Fourier transform [19], vive la difference [20,21], phantom derivative [22,23] and phase-driven model refinement [24]. SYNERGY, as well as REMO09, was included in a modified version of SIR2014 [25].

In nucleic acid structures, sometimes the ligands’ scattering power is a not negligible part of the total scattering power, and often ligands contain (or are constituted by) heavy atoms. If no ligand is taken into account, the final $R$ value may be large even if the nucleic acid model is good. This is mainly due to the fact that any AMB program is more focused on building nucleic acid models than defining the ligand substructure. Because the latter contribution does not enter into the model structure
factor calculation, an additional systematic discrepancy occurs between the observed structure factor amplitudes and the calculated nucleic acid amplitudes, thereby causing larger values of $R$ (and of $R_f$) and, therefore, a larger distrust of the user.

We then decided to modify the standard CAB approach by searching for ligands, including heavy atoms, in the first step. In this step, another task may also be accomplished: the identification of the P atoms belonging to the nucleic acid backbone. This decision is supported by the observation that often the average phase error corresponding to $\phi_r$ phases (say $<|\Delta \phi_r|>$) is low, even if the average phase error at the molecular replacement level is large. This is mainly due to the effectiveness of the SYNERGY step. Indeed, after SYNERGY it may be easier to locate heavy atoms and also to recognize a good percentage of the P atomic positions.

A useful premise for the location of heavy atoms is the following: the number and positions of the heavy atoms are unknown, while their atomic species are assumed to be known due to the known chemical composition of the ligands. Only atoms with an atomic number equal or larger than 20 are considered “heavy”. The above condition is suggested by the following criterion: heavy atoms are not subjected to constraints or restraints during the least-squares refinement, so the optimization of their positions is feasible only if the atomic species is sufficiently heavy. If ligands contain two or more heavy atom species, we associate them with the heaviest atomic species. The following automatic algorithm is used:

(i) An observed electron density map is calculated by using $\phi_r$ phases and $w_r$ weights. The first candidates for being P atoms in the target structure are the P atoms of the model structure, as refined by SYNERGY. If no heavy atoms are present in ligands, steps (ii) and (iii) are skipped;

(ii) The highest $N$ peaks (where $N = 30 \times$ number of nucleotides in the target sequence) are selected and sorted with respect to the intensity (I). All peaks closer than DIST from model atoms are rejected, where DIST corresponds to the covalent radius of the heaviest species in the target. Non-crystallographic occupancy $I(i)/I(1)$ and the heaviest atomic species in the target are associated to the $i$th peak;

(iii) The structure parameters of 10 peaks with the largest $I(i)/I(1)$ values are refined, one atom at a time, together with the nucleic acid model previously determined, by five REFMAC cycles. Then, a new $R$ value is obtained. If it is smaller than the previous one, the heavy atom is accepted as a reliable candidate of the heavy atom substructure and the next peak is processed; otherwise, the procedure for locating heavy atoms stops.

At the end of the steps (ii) and (iii), a list of heavy atom candidates are available. The final $R$ (and $R_f$) values and the new phase estimates (denoted by $\phi_{rh}$) are expected to be better than the corresponding values obtained at the end of SYNERGY. Correspondently, the new average phase error (say $<|\Delta \phi_{rh}|>$) is expected to be smaller.

Heavy atom candidates added to the SYNERGY model are used to calculate new structure factors by which a new electron density map is calculated. Steps ii) and iii) are repeated and a new model is obtained, which now includes the final estimates of P and heavy atoms’ positions. A new phase set ($\phi_{rh}$) and a new average phase error ($<|\Delta \phi_{rh}|>$) correspond to such models.

There is a special reason why P atoms’ positions are examined. In an experimental version of NAUTILUS (Cowtan, personal communication 2020), it is possible to use two new tools to improve the NAUTILUS default model building: a more extended library of nucleic acid model structures and the previous knowledge of the positions of the triples (O$_{3'}$, P, O$_{5'}$). A good percentage of such triples may be correctly identified when SYNERGY and the above described algorithm for heavy atom location end with a small $R$ value. Then, a list of triples (O$_{3'}$, P, O$_{5'}$) is automatically passed to NAUTILUS, which may build the final model of the target nucleic acid more efficiently.
3. The Recursive Algorithm

We briefly summarize, in this section, the main CAB algorithms that make the application of NAUTILUS cyclic. The different characteristics of nucleic acids with respect to proteins suggest that CAB algorithms cannot be the same for the two types of structures.

We suppose that REMO09 molecular replacement techniques were applied to the test structures, and then the corresponding phases were refined by SYNERGY. Let \( \phi_p \) and \( w_p \) be phases and weights at the end of the above procedure: they constitute the input for NAUTILUS, CAB, ARP/wARP and PHENIX.AUTOBUILD. Furthermore, let \( \phi_b \) and \( w_b \) be phases and weights obtained after the first application of NAUTILUS. We divided the CAB procedure into the following steps:

STEP 1: When a ligand contains heavy atoms, the procedure for locating them starts (see Section 2). \( \phi_{rh} \) and \( w_{rh} \) are phases and weights corresponding to the combination of the nucleic acid model and the heavy atoms; \( R_{rh} \) is the corresponding crystallographic residual and \( R_{fb} \) the \( R_{free} \) value. \( \phi_{rh} \) and \( w_{rh} \) coincide with \( \phi_p \) and \( w_p \) when ligand heavy atoms are not found;

STEP 2: \( \phi_p \) and \( w_{rh} \) are used automatically to start the first NAUTILUS application, which provides a new molecular model of the nucleic acid. The Fourier inversion of the new model leads to a set of calculated structure factors to which the contribution of the ligand heavy atoms is added. Let \( R_b \) and \( R_{fb} \) be the corresponding crystallographic residuals, \( \phi_b \) and \( w_b \) the corresponding model phases and weights. If \( R_b \) is smaller than 0.30, then CAB stops and the model is considered not worthy of further improvement;

STEP 3: The \( w_{rh} \) and \( w_b \) distributions are fitted through histogram matching, to put them on the same statistical basis. Then, the tangent

\[
\tan \phi_C = \frac{w_{rh} \sin \phi_{rh} + S_C \cdot w_b \sin \phi_b}{w_{rh} \cos \phi_{rh} + S_C \cdot w_b \cos \phi_b}
\]

is calculated, to derive a set of combined \( \phi_C \) phases. \( S_C \) is the parameter that defines how \( \phi_{rh} \) and \( \phi_b \) should be combined. If the \( \phi_b \) phases are supposed to be reliable, then \( S_C \) is expected to be large; if the user is not confident of their quality, then \( S_C \) has to be small. At this stage, the quality of the \( \phi_b \) phases may be estimated through the \( R_b \) value. We heuristically decided to linearly relate \( S_C \) to \( R_b \) via Equation (2) (owing to the different quality of the problem, this equation does not coincide with that used for proteins):

\[
S_C = 1.975 - 3.25R_b
\]

with the conditions that if \( R_b < 0.30 \), then \( S_C = 1 \), and if \( R_b > 0.5 \), then \( S_C = 0.35 \).

The reason is the following: when \( R_b \) is sufficiently small, then the \( \phi_b \) phases are expected to be reliable and their weights deserve to stay on the same scale of the \( \phi_{rh} \) phases. If \( R_b \) is large, then the contribution of the \( \phi_b \) phases to the tangent expression (1) has to be depleted. The weight

\[
w_C = \frac{1}{2}(t^2 + b^2)^{\frac{1}{2}}
\]

may be applied to the \( \phi_C \) phases.

However, if \( R_b \) is large, then it is very likely that the weakly weighted \( \phi_C \) phases are badly estimated. Accordingly, we decided to eliminate in Equation (1) a percentage (PERC) of the \( \phi_b \) phases (those with lower weights) defined by the following equation:

\[
PERC = 2.4R_{fb} - 0.84
\]

with the conditions that if \( R_{fb} > 0.6 \), then \( PERC = 0.60 \), and if \( R_{fb} < 0.35 \), no reflection is eliminated.

Equation (3) is equivalent to assigning \( w_b = 0 \) to 60% of the weakest estimates from NAUTILUS when \( R_{fb} \) is equal or larger than 0.6, and to assigning \( w_b = 0 \) to the 12% of the weakest estimates from NAUTILUS when \( R_{fb} = 0.40 \). \( \phi_C \) phases and \( w_C \) weights thus obtained are used as input values for the
next NAUTILUS run to produce new $\phi_b$ phases and $w_b$ weights, which are again combined according to Equation (1) with $\phi_{rh}$ phases and $w_{rh}$ weights in up to six cyclic NAUTILUS runs. The procedure stops if $R_b < 0.30$;

STEP 4: The cyclic procedure described in STEP 3 is a useful tool that offers a variety of electron density maps to NAUTILUS algorithms, to increase the chance of a good interpretation. The six maps, however, are close to each other: a large variety of maps could make success easier. A total of 12 supplementary cycles are thus introduced in the procedure. The $\phi_b$ phases and $w_b$ weights obtained at the cycle $n$ are combined via a tangent expression with the $\phi_c$ phases and $w_c$ weights obtained at the end of the $(n - 1)$-th tangent cycle. The procedure stops if $R_b < 0.30$.

At the end of the above procedure, the minimum $R_b$ value is selected (and denoted as $R_C$); the corresponding model is considered the most accurate.

A special case, not very rare in nucleic acid crystallography, occurs when the model and target sequences are identical or differ by one nucleotide. Among the 29 test structures, seven cases (4xqz, 5ihd, 5jua, 5nt5, 5t4w, 2a0p, 5tpg) have identical nucleotides and four cases (3n4o, 1iha, 2fd0, 4enc) differ by one nucleotide. In the latter case, the lengths of the two sequences may be the same or may differ by one nucleotide. If $R_{rh} < 0.35$ and $R_{rh} < R_C$, then SYNERGY and STEP 1 models are preferred to the CAB model (see Section 4).

4. Applications

In a previous paper [26], we selected from the Protein Data Bank (PDB) 38 nucleic acid structures for which phase solution attempts were made via molecular replacement (MR) techniques: we downloaded the observed diffraction data, unit cell dimensions, the space group symmetry, nucleotide sequence, and the structural models. We submitted the test structures to default runs of REMO09; for nine of them, REMO09 did not provide a sufficiently good model (i.e., the average phase error for such structures was larger than 80°). The remaining 29 structures, quoted in Table 1, are used as test cases for our applications: the first 16 of them are DNA, the other 13 are RNA fragments. For each structure, we show their PDB code (PDB), the space group (SG), and the data resolution (RES) in Table 1. The number of nucleotides in the asymmetric unit is reported in the form $n \cdot N$, where $n$ is the number of chains in the asymmetric unit and $N$ is the number of nucleotides per chain. $N$ is replaced by a sum of two numbers if two chains with a different type or number of nucleotides are present. The model used in the MR step (column model) is reported as $p \cdot CODE$, where CODE is the PDB code of the molecular fragment and $p$ is the number of fragments originally used in the MR process. The information on the ligands is given in the corresponding column, in the form of $m \cdot CODE$, where CODE is the PDB code of a ligand and $m$ is the number of ligands in the asymmetric unit of the target structure. The chemical formula for each ligand CODE is also specified at the bottom of Table 1, from which the possible presence of heavy atoms may be deduced.

Model phases obtained by REMO09 and refined by SYNERGY ($\phi_r$ and $w_r$, respectively) constitute the input for CAB, ARP/wARP and PHENIX.AUTOBUILD. Because we are interested in procedures for the automatic crystal structure solution, we used default directives for all three AMB programs. The available documentation for these programs suggested the following instructions (adapted to the 3eil test structure, as an example):

for CAB-NAUTILUS

```
nautilus_pipeline–stdin < 3eil_nautilus.inp, where 3eil_nautilus.inp contains
seqin 3eil.pir
mtzin 3eil_synergy.mtz
colin-fo F, SIGF
colin-phifom PHIC, FOM
colin-free FreeR_flag
pdbin 3eil_po3o5_out.pdb
```
for ARP/wARP:
```
auto_nuce.sh
```
```
datafile 3eil_synergy.mtz
nucleotides 72
fp F sigfp SIGF sigfp FOM FOM
```
for PHENIX.AUTOBUILD:
```
phenix.autobuild write_run_directory_to_file = 3eil_phenix.log
seq_file = 3eil.pir
input_data_file = 3eil_synergy.mtz
input_labels = “F SIGF PHIC FOM HLA HLB HLC HLD FreeR_flag”
chain type = DNA
ncs_copies = 1
nproc = 12
```

We are conscious that the above instructions do not correspond to the optimized ways of applying the mentioned AMB programs. Indeed, any of them may be more effective if suitable instructions are introduced to treat special types of DNA or RNA and/or to explore different building approaches. The directives we used are only simple tools for automatic runs, which, if successful, constitute important achievements by themselves.

Table 1. For the 29 nucleic acid test structures, the following abbreviations are used: Protein Data Bank (PDB) for the structure code, space group (SG) and the diffraction data resolution (RES) (Å). The number of nucleotides in the asymmetric unit (nN) is reported by the symbol n·N, where n is the number of chains per asymmetric unit, N is the number of nucleotides per chain. N is replaced by a sum of two numbers if chains with different type or different number of nucleotides are present. The model used in the MR step (column model) is reported as p·CODE, where CODE is the PDB code of the molecular fragment and p is the number of fragments originally used in the MR process. The information on the ligands is given in the corresponding column, in the form of m·CODE, where CODE is the PDB code of a ligand, and m is the number of ligands in the asymmetric unit of the target structure. The chemical formula for each ligand CODE is also specified at the bottom of the Table, from which the presence of heavy atoms may be deduced.

| PDB   | SG  | RES  | nN   | Model | Ligand(s)     |
|-------|-----|------|------|-------|----------------|
| 3ce5  | I4  | 2.50 | 2.12 | 1khp  | 2K + BRA       |
| 3eil  | P32 | 2.60 | 6.12 | 3463d | 7Mn            |
| 3n4o  | P22  | 2.90 | 2.12 | 1dnh  | 2B7C + HT      |
| 3tok  | C2  | 1.74 | 10+10| 2org  | Na             |
| 4gsg  | C2  | 2.00 | 2(10+10)| 2.9org | UCL            |
| 5dwx  | P42  | 2.23 | 1.10 | 3qnm  | Ba + RKF       |
| 4sqz  | P22  | 2.15 | 8.6  | 2.5ihd| 6-Cu + 4-Ca + 7-CI + MES + MOH |
| 5dax  | P42  | 2.71 | 24+8 | 1kFl  | K              |
| 5m5 [35] | R3  | 2.46 | 2.12 | 476d  | 8-Ca + 2-1W5   |
| 5m4d [33] | P22  | 1.57 | 4.6  | 2.2drg| 4-Cu + 2-Ca + 2-2OP + SIN |
| 5m44 [36] | P22  | 2.00 | 2.12 | 1d29  | Mg + CI        |
| 5j44 [37] | R3  | 2.17 | 2.12 | 463d  | 4-Ca + 2-1W5 + 2-1WA |
| 5wrt [38] | P32  | 1.89 | 2.12 | 5mvl  | 3-Co           |
| 5mt5 [39] | P22  | 2.30 | 2.12 | 1d29  | Na + CAC       |
| 5x4w [40] | P22  | 2.30 | 2.12 | 5ua   | DAP            |
| 1ba4 [41] | C2  | 1.60 | 2.9  | 165d  | 2-CI + 2-BRU + 2-RHD |
| 1z7f [42] | P32  | 2.10 | 3.16 | 1ym  | 2Sr            |
when detected after the SYNERGY step. The results obtained without such tools were poorer and are
not shown for brevity.

Table 2. NAUTILUS and cyclic automated model building (CAB) results. For the 29 test structures, PDB is their PDB code, \( \langle | \Delta \phi_f | \rangle > ^{O} \) is the average phase error at the end of SYNERGY, \( R_f \) and \( R_f \) are the crystallographic residuals (for all of the data) and the \( R_{free} \) value. The corresponding \( \phi_f \) phases are the input for NAUTILUS and CAB. \( R_f \), \( R_{free} \), \( MA_N \) and \( MA_{MN} \) are the \( R \), \( R_f \), \( MA \) and \( MA_{MN} \) values obtained after the application of NAUTILUS, \( R_C \), \( R_{JC} \), \( MA_C \) and \( MA_{MC} \) are the corresponding values obtained at the end of \( CAB. \) \( R \) and \( MA \) values are percentages.

| PDB | \( <| \Delta \phi_f |>^O \) | \( R_f \) | \( R_N \) | \( R_{free} \) | \( MA_N \) | \( MA_{MN} \) | \( R_C \) | \( R_{JC} \) | \( MA_C \) | \( MA_{MC} \) |
|------|-----------------|--------|--------|----------------|----------------|----------------|--------|--------|--------|--------|
| 3ce5 | 50              | 41     | 43     | 54             | 59             | 56             | 53     | 53     | 41     | 18     |
| 3ei1 | 46              | 31     | 36     | 47             | 50             | 59             | 49     | 36     | 38     | 76     |
| 3n4o | 33              | 23     | 26     | 44             | 45             | 45             | 36     | 23     | 26     | 91     | 69     |
| 3tok | 49              | 35     | 35     | 57             | 58             | 44             | 15     | 52     | 56     | 72     | 24     |
| 4gsq | 53              | 34     | 38     | 45             | 45             | 47             | 9      | 42     | 46     | 44     | 17     |
| 4ms5 | 59              | 46     | 64     | 56             | 57             | 50             | 0      | 4      | 37     | 41     | 78     | 57     |
| 4qz | 48              | 32     | 35     | 58             | 58             | 50             | 22     | 27     | 30     | 80     | 94     |
| 5dwx | 58              | 41     | 44     | 57             | 58             | 18             | 5      | 48     | 59     | 32     | 25     |
| 5i4s | 35              | 25     | 29     | 36             | 37             | 59             | 49     | 35     | 34     | 82     | 51     |

In Table 2, we show the experimental results obtained by NAUTILUS and CAB, both obtained by using the new NAUTILUS library and the knowledge of the positions of the triples (O\(_3^\prime\), P, O\(_3^\prime\)), when detected after the SYNERGY step. The results obtained without such tools were poorer and are not shown for brevity. \( \langle | \Delta \phi_f |>^O \) is the average phase error at the end of SYNERGY, \( R_f \) and \( R_f \) are the corresponding crystallographic residual (for all of the data) and the \( R_{free} \) value. \( R_N \), \( R_{free} \), \( R_C \) and \( R_{JC} \) are the \( R \), \( R_f \), \( MA \) and \( MA_{MN} \) values obtained after the application of NAUTILUS, \( R_C \), \( R_{JC} \), \( MA_C \) and \( MA_{MC} \) are the corresponding values obtained at the end of \( CAB. \) \( R \) and \( MA \) values are percentages.
Table 2. Cont.

| PDB   | $<\Delta \phi_i>$ | $R_i$ | $R_f$ | $R_N$ | $R_FN$ | $MA_N$ | $MA_{MN}$ | $RC$ | $R_{JC}$ | $MA_C$ | $MA_{MC}$ |
|-------|-------------------|-------|-------|-------|--------|--------|------------|------|--------|--------|-----------|
| 5ihd  | 39                | 34    | 36    | 51    | 52     | 50     | 39         | 25   | 29     | 100    | 92        |
| 5ju4  | 26                | 26    | 28    | 37    | 37     | 95     | 83         | 26   | 28     | 100    | 100       |
| 5lj4  | 29                | 25    | 29    | 44    | 48     | 86     | 58         | 41   | 45     | 82     | 58        |
| 5mvt  | 28                | 29    | 28    | 38    | 37     | 82     | 79         | 31   | 31     | 95     | 92        |
| 5nt5  | 24                | 27    | 28    | 46    | 47     | 86     | 64         | 27   | 28     | 100    | 99        |
| 5t4w  | 25                | 25    | 29    | 43    | 42     | 86     | 64         | 25   | 29     | 100    | 96        |
| 1iha  | 41                | 34    | 35    | 36    | 37     | 94     | 77         | 23   | 25     | 88     | 81        |
| 1z7f  | 34                | 32    | 34    | 42    | 43     | 69     | 71         | 30   | 30     | 100    | 100       |
| 2a0p  | 31                | 27    | 35    | 32    | 39     | 100    | 93         | 27   | 35     | 100    | 99        |
| 2fd0  | 33                | 32    | 36    | 37    | 38     | 89     | 78         | 32   | 36     | 95     | 85        |
| 2pn4  | 40                | 34    | 40    | 41    | 48     | 87     | 68         | 36   | 41     | 86     | 74        |
| 3d2v  | 57                | 47    | 51    | 49    | 51     | 34     | 29         | 49   | 50     | 32     | 30        |
| 3fs0  | 63                | 42    | 47    | 40    | 41     | 68     | 51         | 29   | 33     | 89     | 86        |
| 4enc  | 28                | 25    | 28    | 36    | 39     | 83     | 74         | 25   | 28     | 98     | 95        |
| 4enc  | 29                | 31    | 39    | 37    | 46     | 94     | 55         | 32   | 41     | 94     | 63        |
| 4enc  | 31                | 32    | 36    | 35    | 39     | 74     | 51         | 34   | 39     | 74     | 53        |
| 4enc  | 34                | 34    | 36    | 30    | 33     | 99     | 88         | 30   | 33     | 99     | 88        |
| 6az4  | 51                | 36    | 40    | 28    | 30     | 87     | 63         | 28   | 30     | 87     | 63        |

Table 2 suggests the following conclusions:

1. $MA_N$ and $MA_C$ only deal with the quality of the P chains; their usefulness as figures of merit has to be confirmed by $MA_{MN}$ and by $MA_{MC}$, respectively, which define the overall quality of the structural model. Even if there is a good correlation between $MA$ and $MA_M$ for all the tested AMB programs, their indications do not always agree;

2. The inequality $MA_N > MA_C$ is rare (only in four cases, 5lj4, 1iha, 2pn4, 3d2v), and in all cases $MA_{MN} \leq MA_{MC}$;

3. For a high percentage of test structures, the quality of the NAUTILUS model is largely improved by CAB (examples are not given for brevity). Frequently, quite poor initial models are transformed by CAB into almost complete models. These cases correspond to poor values of $MA_N$ and $MA_{MN}$, and to large values of $MA_C$ and $MA_{MC}$;

4. In all test cases, $R_N \geq R_C$. That increases the confidence of CAB users in the quality of the built model. In some cases, the final residuals are large because of the unmodeled contribution to the diffraction from ligands that are missing in the model;

5. Eleven test cases (in bold) represent situations where the model and target sequences are identical or differ by only one residue and where the conditions $R_h < 0.35$ and $R_h < R_C$ are satisfied. The program automatically checks the sequence relationships and verifies if the numerical conditions are satisfied. As stated before, in all eleven cases, the program automatically chooses the models at the end of STEP 1 rather than the final CAB models. As an example, in Figure 1, we show the structure 2fd0, for which $MA_C = 0.95$ and $MA_{MC} = 0.85$. The CAB model is on the left and the published model on the right; the main difference concerns a nucleotide close to a chain terminal.
In Table 3, we quote $\text{MA}_C$ and $\text{MA}_{MC}$ values obtained with and without the application of the algorithm for the sequence control, to allow the reader to understand how these values differ from each other. It is easily seen that $\text{MA}_C$ and $\text{MA}_{MC}$ values without the control are much worse;

Table 3. $\text{MA}_C$ and $\text{MA}_{MC}$ for the eleven structures for which model and target sequences are equal or differ in one position, and for which the conditions $R_{rh} < 0.35$ and $R_{rh} < R_C$ are satisfied. **WITH** and **WITHOUT** indicate if the control on the sequences has been applied or not. **MA** values are percentages.

| PDB  | WITH $\text{MA}_C$ | WITH $\text{MA}_{MC}$ | WITHOUT $\text{MA}_C$ | WITHOUT $\text{MA}_{MC}$ |
|------|------------------|------------------|------------------|------------------|
| 3n4o | 91               | 69               | 77               | 38               |
| 4xqz | 80               | 94               | 43               | 30               |
| 5ihd | 100              | 92               | 70               | 47               |
| 5ju4 | 100              | 100              | 95               | 83               |
| 5nt5 | 100              | 99               | 100              | 87               |
| 5t4w | 100              | 96               | 91               | 64               |
| 1iha | 88               | 81               | 81               | 77               |
| 2a0p | 100              | 99               | 100              | 99               |
| 2fd0 | 95               | 85               | 95               | 81               |
| 4enc | 98               | 95               | 83               | 78               |
| 5tgp | 100              | 100              | 100              | 75               |

(6) CAB (and NAUTILUS, of course) usually fails when $|\Delta\phi_r|^\circ$ is close or larger than 50° (this is the case for 3ce5, 3tok, 4gsg, 4ms5, 5dwx, 3d2v), even if two cases can be found in which it has success (3eil and 3fs0). This error limit is usually exceeded when CAB is applied to proteins.

In Table 4 we show, for all test structures, the values of $R$, $R_f$, $\text{MA}$ and $\text{MA}_M$ obtained after the application of ARP/wARP (say $R_A$, $R_{fA}$, $\text{MA}_A$ and $\text{MA}_{MA}$, respectively), and the analogous values obtained by the application of PHENIX.AUTOBUILD (say $R_P$, $R_{fP}$, $\text{MA}_P$ and $\text{MA}_{MP}$, respectively). Comparing the quartet $R_A$, $R_{fA}$, $\text{MA}_A$ and $\text{MA}_{MA}$ with the quartet $R_P$, $R_{fP}$, $\text{MA}_P$ and $\text{MA}_{MP}$ clearly suggests the larger effectiveness of PHENIX.AUTOBUILD: usually $R_P < R_A$, $R_{fP} < R_{fA}$, $\text{MA}_P > \text{MA}_A$ and $\text{MA}_{MP} > \text{MA}_{MA}$. The cpu time, however, is much larger for PHENIX.AUTOBUILD.
Table 4. ARP/wARP and PHENIX.AUTOBUILD experimental results for the 29 test structures. PDB is their PDB code; $|\Delta\phi_r|$ is the average phase error available at the end of the SYNERGY refinement process. The corresponding $\phi_r$ phases are the input for ARP/wARP and PHENIX.AUTOBUILD. $R_A$, $R_{\phi_A}$, $M_A$ and $M_{MA}$ are the $R$, $R_{\phi}$, $M$ and $M_{M}$ values obtained at the end of the automatic runs of ARP/wARP; $R_P$, $R_{\phi_P}$, $M_P$ and $M_{MP}$ are the corresponding values obtained at the end of the automatic runs of PHENIX.AUTOBUILD. $R$ and $M$ values are percentages.

| PDB  | $|\Delta\phi_r|$ (°) | $R_A$ | $R_{\phi_A}$ | $M_A$ | $M_{MA}$ | $R_P$ | $R_{\phi_P}$ | $M_P$ | $M_{MP}$ |
|------|---------------------|------|-------------|------|---------|------|-------------|------|---------|
| 3ce5 | 50                  | 53   | 56          | 23   | 11      | 45   | 47          | 50   | 40      |
| 3e1l | 46                  | 48   | 56          | 26   | 15      | 43   | 47          | 53   | 57      |
| 3n4o | 33                  | 33   | 52          | 64   | 25      | 33   | 37          | 82   | 57      |
| 3tok | 49                  | 52   | 53          | 28   | 10      | 45   | 47          | 94   | 34      |
| 4gsg | 53                  | 37   | 43          | 44   | 16      | 38   | 38          | 39   | 17      |
| 4ms5 | 59                  | 0    | 0           | 0    | 0       | 48   | 53          | 44   | 29      |
| 4xqz | 48                  | 53   | 54          | 13   | 5       | 57   | 60          | 10   | 11      |
| 5dwx | 56                  | 40   | 47          | 36   | 10      | 49   | 53          | 27   | 27      |
| 5fs4 | 35                  | 34   | 44          | 50   | 22      | 36   | 39          | 50   | 49      |
| 5ihd | 39                  | 51   | 51          | 10   | 5       | 52   | 56          | 25   | 19      |
| 5j4u | 26                  | 49   | 58          | 59   | 14      | 35   | 33          | 100  | 84      |
| 5j4j | 29                  | 41   | 52          | 55   | 22      | 40   | 41          | 68   | 65      |
| 5mvt | 28                  | 45   | 51          | 50   | 22      | 46   | 44          | 91   | 60      |
| 5na5 | 24                  | 35   | 48          | 91   | 43      | 35   | 38          | 100  | 84      |
| 5lw4 | 25                  | 31   | 45          | 91   | 49      | 31   | 33          | 95   | 83      |
| 1ha  | 41                  | 41   | 41          | 75   | 51      | 36   | 33          | 81   | 64      |
| 1z7l | 34                  | 40   | 46          | 69   | 32      | 35   | 36          | 91   | 82      |
| 2a0p | 31                  | 39   | 53          | 86   | 40      | 31   | 37          | 93   | 93      |
| 2fd0 | 33                  | 45   | 52          | 73   | 30      | 37   | 36          | 95   | 80      |
| 2pn4 | 40                  | 47   | 55          | 32   | 13      | 42   | 48          | 57   | 52      |
| 3d2v | 57                  | 56   | 57          | 6    | 3       | 47   | 48          | 26   | 23      |
| 3fs0 | 63                  | 0    | 0           | 0    | 0       | 29   | 34          | 74   | 69      |
| 4enc | 28                  | 33   | 46          | 79   | 34      | 40   | 41          | 71   | 67      |
| 5kvj | 49                  | 39   | 55          | 59   | 20      | 35   | 40          | 84   | 63      |
| 5l4o | 40                  | 44   | 53          | 46   | 16      | 45   | 50          | 54   | 49      |
| 5rz6 | 45                  | 34   | 38          | 53   | 29      | 35   | 37          | 78   | 56      |
| 5gp  | 26                  | 45   | 51          | 86   | 45      | 34   | 33          | 100  | 89      |
| 5uz6 | 34                  | 34   | 40          | 91   | 53      | 33   | 33          | 91   | 82      |
| 6az4 | 51                  | 42   | 46          | 38   | 15      | 39   | 40          | 67   | 53      |

PHENIX.AUTOBUILD and CAB results may be easily compared via their corresponding quartet ($R$, $R_{\phi}$, $M$, $M_{M}$). Usually $R_P > R_C$, $R_{\phi_P} > R_{\phi_C}$, $M_P < M_C$, $M_{MP} < M_{MC}$, but there are also few cases in which PHENIX.AUTOBUILD alone performs better.

Figures 2 and 3 synthetically represent the results quoted in Tables 2 and 4. Figure 2 shows the MA values corresponding to the default application of NAUTILUS, CAB, ARP/wARP and PHENIX.AUTOBUILD. In this condition, ARP/wARP seems the least efficient program: $M_A > M_C$ only in one case (5dwx) and $M_A =$ $M_C$ also in one case (4gsg). The NAUTILUS and PHENIX.AUTOBUILD lines are closer to the CAB line. For NAUTILUS, $M_{AN} > M_C$ in four cases (5kvj, 5l4o, 5uz6, 6az4) and $M_{AN} =$ $M_C$ in three cases (1ha, 2pn4, 3d2v). For PHENIX.AUTOBUILD, $M_P > M_C$ in only two cases (3ce5, 3tok) and $M_P =$ $M_C$ in four cases (5ju4, 5nt5, 2fd0, 5gp).
As previously stated, MA values are not in themselves indisputable estimates of the quality of the built models, because they register only the correctness of the P atoms. $M_{M}$ may be considered a more general figure of merit involving all the non-H atoms. The $M_{M}$ values obtained by the four tested programs are plotted in Figure 3. A common feature, no matter the algorithm used for AMB, is that usually $MA > M_{M}$; the P positions are more easily located than the other atoms.

Even in this case, ARP/wARP seems the least efficient program, while the NAUTILUS and PHENIX.AUTOBUILD lines are closer to the CAB line, but the quality of the CAB models is markedly
higher. In most cases, the CAB percentage of non-H atoms at a distance less than 0.6 Å from the published positions is greater than 50.

A final observation is mandatory. This paper is mainly concerned with the full automation of the model building tools. However, the role of CAB for nucleic acids in the present scientific panorama may be better appreciated by including it in Table 5, where the most popular automated or semi-automated tools for model building are cited.

Table 5. Most popular automated or semi-automated tools for model building of nucleic acids. SUB1: automated model building into electron density map from sequence; SUB2: guided semi-automated model building into electron density maps; SUB3: completing and rebuilding existing models into electron density maps; SUB4: building models from sequences without electron density.

| SUB1            | SUB2                  | SUB3                     | SUB4               |
|-----------------|-----------------------|--------------------------|--------------------|
| NAUTILUS [10]   | RCRANE [6] in COOT [54]| NAFIT, NABUILD in LAFIRE [5] | AMBER [55]         |
| ARP/wARP [9]    |                       | ERRASER [56]             | FARFAR [57,58]     |
| PHENIX.AUTOB    |                       |                          | ROSETTA [59]       |
| NUT/DHL/RSR     |                       |                          | 3DNA [62]          |
| CAB             |                       |                          |                    |

5. Discussion

The CAB approach, originally designed for making the BUCCANEER application to proteins cyclic, was modified for use as an AMB tool for nucleic acids. In the new CAB version, we included NAUTILUS; the purpose was to improve the AMB effectiveness without changing NAUTILUS algorithms.

We applied CAB to a set of 29 nucleic acids (DNA and RNA) and compared the models thus obtained with those available after the mere application of NAUTILUS. We also applied ARP/wARP and PHENIX.AUTOBUILD to the same set of test structures. The procedures were fully automatic: a set of default instructions were given as inputs to any AMB program. Obviously, more appropriate input directives may improve the experimental results described in this paper. The results thus obtained show that the CAB cyclic approach remarkably increases NAUTILUS effectiveness and it is quite competitive with ARP/wARP and PHENIX.AUTOBUILD.

The AMB programs tested in this paper clearly show that their efficiency for nucleic acids is much smaller than for proteins. This partly depends on the particular difficulties to overcome for nucleic acids (see Section 1), but also on the smaller efforts spent in this field. This conclusion is supported by the following observation: quite often, SYNERGY ends with $<|\Delta \phi_r|> \leq 40^\circ$ (16 times out of 29). This situation is usually very favourable for AMB programs when applied to proteins; on the contrary, Tables 2 and 3 show that MA and MA_M values are often far from the expected values. Further efforts are needed for a complete and satisfactory AMB automation. Some of these efforts may be spent on improving the $\phi_r$ phases, but most of them should concern the improvement of the AMB algorithms.

CAB for nucleic acids is part of an experimental version of SIR2014. Its full use requires that an experimental version of NAUTILUS, on which it is based, is also available. Hopefully, CAB will be released in late 2020.

Author Contributions: Conceptualization, G.L.C. and C.G.; methodology, C.G.; software, G.L.C.; validation, M.C.B., B.C. and G.P.; writing—review & editing, M.C.B., B.C. and G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We thank Kevin Cowtan for his illuminating discussions and for allowing us to use the experimental version of NAUTILUS. We also thank Blaine Mooers for his friendly and precious advice.

Conflicts of Interest: The authors declare no conflicts of interest.
Abbreviations

AMB  automated model building.
MR   molecular replacement.
R    crystallographic residual between observed and calculated structure factor amplitudes (for all of the experimental data).
Rf   cross validation R-value for the free data set [63].
MA   ratio “number of residues with P atoms within 0.6Å distance from the published positions/number of residues in the asymmetric unit”, according to the published sequence. It is an indication of the accuracy of the model.
MAm  ratio “number of non-hydrogen atoms within 0.6Å distance from published positions/number of non-hydrogen atoms in the asymmetric unit”.

References

1. Cowtan, K.D. The Buccaneer software for automated model building. Tracing protein chains. Acta Cryst. 2006, D62, 1002–1011. [CrossRef]
2. Langer, G.; Cohen, S.X.; Lamzin, V.S.; Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. Nat. Protoc. 2008, 3, 1171–1179. [CrossRef] [PubMed]
3. Terwilliger, T.C.; Grosse-Kunstleve, R.W.; Afonine, P.V.; Moriarty, N.W.; Zwart, P.H.; Hung, L.-W.; Read, R.J.; Adams, P.D. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Cryst. 2008, D64, 61–69. [CrossRef]
4. Wang, X.; Kapral, G.; Murray, L.; Richardson, D.; Richardson, J.; Snoeying, J. RNABC: Forward kinematics to reduce all-atom steric clashes in RNA backbone. J. Math. Biol. 2008, 56, 253–278. [CrossRef]
5. Yamashita, K.; Zhou, Y.; Tanaka, I.; Yao, M. New model-fitting and model-completion programs for automated iterative nucleic acid refinement. Acta Cryst. 2013, D69, 1171–1179. [CrossRef]
6. Keating, K.S.; Pyle, A.M. RCrane: Semi-automated RNA model building. Acta Cryst. 2012, D68, 985–995. [CrossRef]
7. Keating, K.S.; Pyle, A.M. Semi-automated model building for RNA crystallography using a directed rotameric approach. Proc. Natl. Acad. Sci. USA 2010, 107, 8177–8182. [CrossRef]
8. Murray, L.J.; Arendall, W.B.; Richardson, D.C.; Richardson, J.S. RNA backbone is rotameric. Proc. Natl. Acad. Sci. USA 2003, 100, 13904–13909. [CrossRef]
9. Hattne, J.; Lamzin, V.S. Pattern-recognition-based detection of planar objects in three-dimensional electron-density maps. Acta Cryst. 2008, D64, 834–842. [CrossRef]
10. Cowtan, K. Automated nucleic acid chain tracing in real time. IUCR J. 2014, 1, 387–392. [CrossRef]
11. Murshudov, G.N.; Vagin, A.A.; Lebedev, A.; Wilson, K.S.; Dodson, E.J. Efficient anisotropic refinement of macromolecular structures using FFT. Acta Cryst. 1999, D55, 247–255. [CrossRef]
12. Afonine, P.V.; Grosse-Kunstleve, R.W.; Echols, N.; Headd, J.J.; Moriarty, N.W.; Mustyakimov, M.; Terwilliger, T.C.; Urzhumtsev, A.; Zwart, P.H.; Adams, P.D. Towards automated crystallographic structure refinement with phenix.refine. Towards automated crystallographic structure refinement with phenix.refine. Acta Cryst. 2012, D68, 352–367. [CrossRef]
13. Burla, M.C.; Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C.; Polidori, G. CAB: A cyclic automatic model building procedure. Acta Cryst. 2018, D74, 1096–1104. [CrossRef]
14. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C.; Mazzone, A.; Siliqi, D. Molecular replacement: The probabilistic approach of the program REMO09 and its applications. Acta Cryst. 2009, A65, 512–527. [CrossRef]
15. Burla, M.C.; Cascarano, G.L.; Giacovazzo, C.; Polidori, G. Synergy among phase-refinement techniques in macromolecular crystallography. Acta Cryst. 2017, D73, 877–888. [CrossRef]
16. Cowtan, K. Fast Fourier feature recognition. Acta Cryst. 2001, D57, 1435–1444. [CrossRef]
17. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; De Caro, L.; Giacovazzo, C.; Siliqi, D. Phasing at resolution higher than the experimental resolution. Acta Cryst. 2005, D61, 556–565. [CrossRef]
18. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; De Caro, L.; Giacovazzo, C.; Siliqi, D. Ab initio phasing at resolution higher than experimental resolution. Acta Cryst. 2005, D61, 1080–1087. [CrossRef]
19. Giacovazzo, C.; Siliqi, D. Improving Direct Methods phases by Heavy-Atom information and Solvent Flattening. *Acta Cryst.* 1997, D53, 789–798. [CrossRef]

20. Burla, M.C.; Caliandro, R.; Giacovazzo, C.; Polidori, G. The difference electron density: A probabilistic reformulation. *Acta Cryst.* 2010, A66, 347–361. [CrossRef]

21. Burla, M.C.; Giacovazzo, C.; Polidori, G. From a random to the correct structure: The VLD algorithm. *J. Appl. Cryst.* 2010, 43, 825–836. [CrossRef]

22. Giacovazzo, C. Solution of the phase problem at non-atomic resolution by the Phantom Derivative method. *Acta Cryst.* 2015, D71, 483–512. [CrossRef] [PubMed]

23. Giacovazzo, C. From direct-space discrepancy functions to crystallographic least squares. *Acta Cryst.* 2015, D71, 36–45. [CrossRef]

24. Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C. Phase improvement via the Phantom Derivative technique: Ancils that are related to the target structure. *Acta Cryst.* 2016, A72, 551–557. [CrossRef]

25. Burla, M.C.; Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; Cuocci, C.; Giacovazzo, C.; Mallamo, M.; Mazzone, A.; Polidori, G. Crystal structure determination and refinement via SIR2014. *J. Appl. Cryst.* 2015, 48, 306–309. [CrossRef]

26. Burla, M.C.; Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C.; Polidori, G. How far are we from automatic crystal structure solution via molecular-replacement techniques? *Acta Cryst.* 2020, D76, 9–18. [CrossRef]

27. Campbell, N.H.; Parkinson, G.N.; Reszka, A.P.; Neidle, S. Structural basis of DNA quadruplex recognition by an acridine drug. *J. Am. Chem. Soc.* 2008, 130, 6722–6724. [CrossRef]

28. Millonig, H.; Pous, J.; Gouyette, C.; Subirana, J.A.; Campos, J.L. The interaction of manganese ions with DNA. *J. Inorg. Biochem.* 2009, 103, 876–880. [CrossRef]

29. Juan, E.C.M.; Shimizu, S.; Ma, X.; Kurose, T.; Haraguchi, T.; Zhang, F.; Tsunoda, M.; Ohkubo, A.; Sekine, M.; Shibata, T.; et al. Insights into the DNA stabilizing contributions of a bicyclic cytosine analogue: Crystal structures of DNA duplexes containing 7,8-dihydropyrido [2,3-d]pyrimidin-2-one. *Nucleic Acids Res.* 2010, 38, 6737–6747. [CrossRef]

30. Carter, M.; Ho, P.S. Assaying the energies of biological halogen bonds. *Cryst. Growth Des.* 2011, 11, 5087–5095. [CrossRef]

31. Carter, M.; Voth, A.R.; Scholfield, M.R.; Rummel, B.; Sowers, L.C.; Ho, P.S. Enthalpy-entropy compensation in biomolecular halogen bonds measured in DNA junctions. *Biochemistry* 2013, 52, 4891–4903. [CrossRef] [PubMed]

32. Hall, J.P. The Effects of Disubstitution on the Binding of Ruthenium Complexes to DNA. Ph.D. Thesis, University of Reading, Reading, UK, 2014.

33. Rohner, M.; Medina-Molner, A.; Spingler, B. N,N,O and N,O,N meridional cis coordination of two guanines to copper(II) by d(CGCGCGCGG). *Inorg. Chem.* 2016, 55, 6130–6140. [CrossRef] [PubMed]

34. Russo Krauss, I.; Ramaswamy, S.; Neidle, S.; Parkinson, G.N. Structural insights into the quadruplex-duplex 3′ interface formed from a telomeric repeat: A potential molecular target. *J. Am. Chem. Soc.* 2016, 138, 1226–1233. [CrossRef] [PubMed]

35. Ahmad Sobri, A.F.; Brady, R.L. Non-natural DNA pair Z (6-amino-5-nitro-2[1H] pyridone heterocycle)-guanosine. 2016; Unpublished work.

36. Sbirkova, H.I.; Schivachev, B.L. Crystal structure of a DNA sequence d(CGTGAATTCACG) at 130K. *Bulg. Chem. Commun.* 2016, 48, 589–593.

37. Reichenbach, L.F.; Sobri, A.A.; Zaccai, N.R.; Agnew, C.R.J.; Burton, N.; Eperon, L.P.; de Ornellas, S.; Eperon, I.C.; Brady, R.L.; Burley, G.A. Structural basis of the mispairing of an artificially expanded genetic information system. *Chemistry* 2016, 1, 946–958. [CrossRef]

38. Hardwick, J.S.; Ptchelkine, D.; El-Sagheer, A.H.; Tear, I.; Singleton, D.; Phillips, S.E.V.; Lane, A.N.; Brown, T. 5-Formylcytosine does not change the global structure of DNA. *Nat. Struct. Mol. Biol.* 2017, 24, 544–552. [CrossRef]

39. Sbirkova-Dimitrova, H.I.; Shivachev, B.L. Crystal structure of the DNA sequence d(CGTGAATTCACG)2 with DAPI. *Acta Cryst.* 2017, 73, 500–504. [CrossRef]

40. Cruse, W.; Saludjian, P.; Neuman, A.; Prange, T. Destabilizing effect of a fluorouracil extra base in a hybrid RNA duplex compared with bromo and chloro analogues. *Acta Cryst.* 2001, D57, 1609–1613. [CrossRef]
42. Gherghe, C.M.; Krahn, J.M.; Weeks, K.M. Crystal structures, reactivity and inferred acylation transition states for 2′-amine substituted RNA. *J. Am. Chem. Soc.* 2005, 127, 13622–13628. [CrossRef]

43. Haeberli, P.; Berger, I.; Pallan, P.S.; Egli, M. Syntheses of 4′-thioribonucleosides and thermodynamic stability and crystal structure of RNA oligomers with incorporated 4′-thiocytosine. *Nucleic Acids Res.* 2006, 34, 2328–2339. [CrossRef]

44. Ennifar, E.; Paillart, J.C.; Bodlenner, A.; Walter, P.; Weibel, J.-M.; Aubertin, A.-M.; Pale, P.; Dumas, P.; Marquet, R. Targeting the dimerization initiation site of HIV-1 RNA with aminoglycosides: From crystal to cell. *Nucleic Acids Res.* 2006, 34, 2328–2339. [CrossRef]

45. Zhao, Q.; Han, Q.; Kissinger, C.R.; Hermann, T.; Thompson, P.A. Structure of hepatitis C virus IRES subdomain Ia. *Acta Cryst.* 2008, D64, 436–443. [CrossRef]

46. Thore, S.; Frick, C.; Ban, N. Structural basis of thiamine pyrophosphate analogues binding to the eukaryotic riboswitch. *J. Am. Chem. Soc.* 2008, 130, 8116–8117. [CrossRef]

47. Pitt, J.N.; Ferre-D’Amare, A.R. Structure-guided engineering of the regioselectivity of RNA ligase ribozymes. *J. Am. Chem. Soc.* 2009, 131, 3532–3540. [CrossRef]

48. Ren, A.; Rajashankar, K.R.; Patel, D.J. Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch. *Nature* 2012, 486, 85–89. [CrossRef]

49. Vorobiev, S.M.; Ma, L.-C.; Montelione, G.T. Crystal structure of the 16-mer double stranded RNA. 2016; Unpublished work.

50. Monestier, A.; Aleksandrov, A.; Coureux, P.D.; Panvert, M.; Mechulam, Y.; Schmitt, E. The structure of an E. coli tRNAf(Met) A1-U72 variant shows an unusual conformation of the A1-U72 base pair. *RNA* 2017, 23, 673–682. [CrossRef]

51. Huang, L.; Wang, J.; Wilson, T.J.; Lilley, D.M.J. Structure of the guanidine III riboswitch. *Cell Chem. Biol.* 2017, 24, 1407–1415. [CrossRef]

52. Zhang, W.; Huang, Z. DNA 8mer containing two 2SeT modifications. 2016; Unpublished work.

53. Zhang, W.; Tam, C.P.; Zhou, L.; Oh, S.S.; Wang, J.; Szostak, J.W. Structural Rationale for the Enhanced Catalysis of Nonenzymatic RNA Primer Extension by a Downstream Oligonucleotide. *J. Am. Chem. Soc.* 2018, 140, 2829–2840. [CrossRef]

54. Emsley, P.; Lohkamp, B.; Scott, W.G.; Cowtan, K. Features and development of Coot. *Acta Cryst.* 2010, D66, 486–501. [CrossRef]

55. Case, D.A.; Cheatham, T.E.I.I.I.; Darden, T.; Gohlke, H.; Luo, R.; Merz KMJr Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* 2005, 26, 1668–1688. [CrossRef]

56. Chou, F.C.; Sripakdeevong, P.; Dibrov, S.M.; Hermann, T.; Das, R. Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nat. Methods* 2013, 10, 74–76. [CrossRef]

57. Das, R.; Karanicolas, J.; Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* 2010, 7, 291–294. [CrossRef]

58. Cheng, C.Y.; Chou, F.-C.; Das, R. Modeling Complex RNA Tertiary Folds with Rosetta. *Methods Enzymol.* 2015, 553, 35–64. [CrossRef]

59. Leaver-Fay, A.; Tyka, M.; Lewis, S.M.; Lange, O.F.; Thompson, J.; Jacak, R.; Kaufman, K.W.; Renfrew, P.D.; Smith, C.A.; Sheffler, W.; et al. Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol.* 2011, 487, 545–574. [CrossRef]

60. Pavelcik, F.; Schneider, B. Building of RNA and DNA double helices into electron density. *Acta Cryst.* 2008, D64, 620–626. [CrossRef]

61. Pavelcik, F. Application of constrained real-space refinement of flexible molecular fragments to automatic model building of RNA structures. *J. Appl. Cryst.* 2012, 45, 309–315. [CrossRef]

62. Lu, X.L.; Olson, W.K. 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* 2003, 31, 5108–5121. [CrossRef]

63. Brünger, A.T. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 1992, 355, 472–475. [CrossRef]