DIBER: protein, DNA, or both?

Grzegorz Chojnowski
MPG-PAN Junior Research Group
International Institute of Molecular and Cell Biology,
4 Ks. Trojdena Street, 02-109 Warsaw, Poland

Institute of Experimental Physics, University of Warsaw,
93 Żwirki i Wigury Street, 02-089 Warsaw, Poland
gchojnowski@iimcb.gov.pl, grzegorz.chojnowski@fuw.edu.pl

Matthias Bochtler
MPG-PAN Junior Research Group,
International Institute of Molecular and Cell Biology,
4 Ks. Trojdena Street, 02-109 Warsaw, Poland

Schools of Chemistry and Biosciences,
Cardiff University,
Park Place, CF10 3AT Cardiff, United Kingdom
mbochtler@iimcb.gov.pl, bochtlerm@cardiff.ac.uk

Abstract

The program DIBER (an acronym for DNA and FIBER) requires only native diffraction data to predict whether a crystal contains protein, B-form DNA, or both. In standalone mode, the classification is based on the cube root of the reciprocal unit cell volume and the largest local average of diffraction intensities at 3.4 Å resolution. In combined mode, the PHASER rotation function score (for the 3.4 Å shell and a canonical B-DNA search model) is also taken into account. In standalone (combined) mode, DIBER classifies $87.4 \pm 0.2\%$ ($90.2 \pm 0.3\%$) of the protein, $69.1 \pm 0.3\%$ ($78.8 \pm 0.3\%$) of the protein-DNA and $92.7 \pm 0.2\%$ ($90.0 \pm 0.2\%$) of the DNA crystals correctly. Reliable predictions with a correct classification rate above 80% are possible for $36.8 \pm 1.0\%$ ($60.2 \pm 0.4\%$) of the protein, $43.6 \pm 0.5\%$ ($59.8 \pm 0.3\%$) of the protein-DNA and $83.3 \pm 0.3\%$ ($82.6 \pm 0.4\%$) of the DNA structures. Surprisingly, selective use of the diffraction data in the 3.4 Å shell improves the overall success rate of the combined mode classification. An open-source CCP4/CCP4i compatible version of DIBER is available from the authors' website at http://diber.iimcb.gov.pl/ and is subject to the GNU Public License.

1 Introduction

Structural studies of protein-nucleic acid complexes require the co-crystallization of both components. If a tight protein-DNA complex is not available for crystallization, uncertainty about the crystal content often remains until the structure is finally solved. Part of the difficulty is due to the surprising observation that DNA can be required for crystallization without getting incorporated into the crystal, perhaps by perturbing the pH of the buffer or by other indirect effects [15]. In principle, the crystal content could be clarified by spectroscopic methods, but the equipment for such measurements is often unavailable. Alternatively, crystals can be washed, dissolved and analyzed by gel electrophoresis with appropriate staining, but this method is destructive and does not always provide a clear-cut answer. On the one hand, components of the crystal can go unnoticed if crystals are small and detection efficiency is limited. On the other hand, components can be falsely diagnosed if they stick to the crystal surface

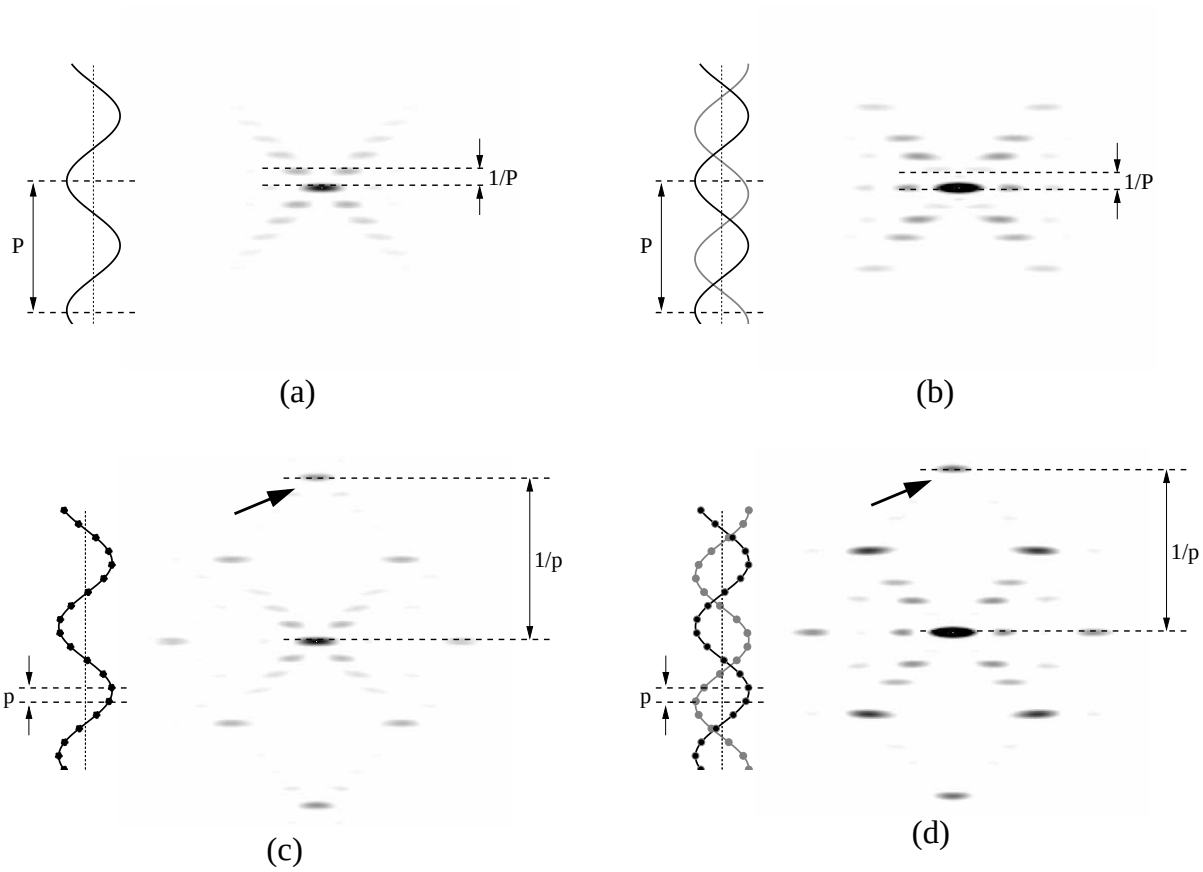

Figure 1: Real and reciprocal space representations of continuous and discontinuous helices and double-helices. All calculations were done with pitch $P = 34$ Å and (average) helix radius $r = 7.0$ Å. Axial distance between pearls in (c) and (d) was $p = 3.4$ Å. Layers have finite width because only two turns of the helix were used for the numerical calculations. The arrows highlight the characteristic 3.4 Å peak.

without being incorporated in the lattice. Clearly, a method that could distinguish between protein crystals, DNA crystals and crystals of both components based on the diffraction data alone (in the absence of any phase information) would be highly desirable.

Crystals that contain only DNA typically have much smaller unit cells than crystals that contain protein (either alone or in combination with DNA) and are therefore easily identifiable. It is much harder to distinguish protein crystals from crystals that contain protein and DNA, because their cell dimensions are typically comparable. We reasoned that the presence of double stranded B-DNA (dsDNA) should be deducible from the characteristic features of its Fourier transform, even though the latter is sampled by the reciprocal lattice in three-dimensional diffraction experiments. The key features of the Fourier transform of ds-DNA are well known [4, 10, 7]. The modulus is approximately cylinder symmetric. Slices that contain the reciprocal space helix axis reveal a cross at low resolution and a strong maximum at 3.4 Å resolution. This maximum is known as the meridional peak in fiber diffraction for its location in a typical setup. It is due to in-phase scattering of all DNA nucleotide pairs (which are related by 3.4 Å shifts along the helix axis and irrelevant rotations) (Fig. 1).

The transverse (perpendicular to the helix axis) and longitudinal (along the helix axis) profile of the characteristic 3.4 Å peak can be analyzed ei-

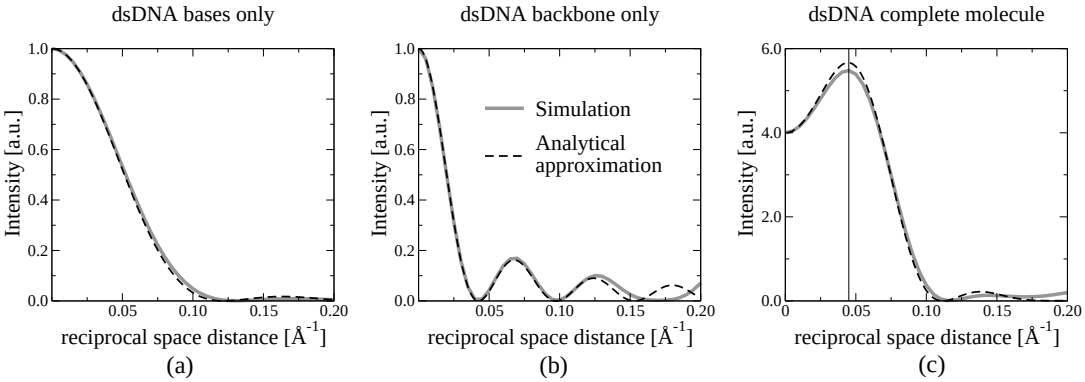

Figure 2: Transverse intensity profile of the 3.4 Å peak of dsDNA. The scattering of dsDNA (a) bases, (b) backbone and (c) complete molecule was estimated analytically (broken lines, formulas 3, 4 and 5 in Appendix A) and calculated numerically with cylindrical averaging (continuous gray lines) for a 10 base pair helix. The vertical line in panel (c) indicates the location of the maximum.

ther numerically and/or analytically (Figs. 2 and 3). The detailed calculations are presented in Appendix A. The transverse profile does not depend on helix length and has a complicated shape (Fig. 2). It can be attributed to coherent superposition of the structure factors of DNA bases (Fig. 2a) and backbone (Fig. 2b). Interference is destructive on-axis due to the location of phosphates half-way between bases in the axial direction. However, radial dependencies of base and backbone scattering differ. Therefore, the contributions reinforce each other at a radial distance $R = 0.04$ Å^{-1} off axis (Fig. 2c). Fortuitously, the 0.08 Å^{-1} separation between the maxima is approximately the inverse of the 12 Å radius of the DNA helix, and therefore perfectly in agreement with the reciprocity of real and reciprocal space dimensions. The longitudinal profile of the 3.4 Å resolution peak can be calculated like the width of the first maximum in a multiple-slit diffraction experiment. The half-width at half-maximum of approximately $\frac{1}{2}(3.4\ \text{Å})^{-1} \approx 0.15$ Å^{-1} divided by the number of base pairs in the dsDNA helix can be confirmed numerically and by more detailed analytical calculations (Fig. 3).

In this work, we present the CCP4-compatible, GPL-licensed program DIBER (an acronym for DNA and FIBER), which takes 3D-diffraction data as input and predicts whether a given crystal contains protein only, protein-DNA, or DNA only. DIBER is intended to search for double stranded B-DNA and not the rarer A- or Z-DNA forms of double stranded DNA, double stranded RNA, nor for any single stranded nucleic acid. The program quantifies the intensity average in regions of reciprocal space that could represent the characteristic 3.4 Å dsDNA peak. Moreover it takes into account the reciprocal unit cell size represented by the cube root of its volume. Assuming equal a priori probabilities for protein only, protein-DNA and DNA only crystals, the program forecasts the crystal content, and assesses the confidence of the prediction. A graph with the reflection averages in a thin resolution shell around 3.4 Å is also produced. Regions of exceptionally strong signals may correlate with the position of the dsDNA characteristic peak, and may indicate the double helix orientation (up to the usual ambiguity of hand). It must be stressed, however, that this information should be taken with a grain of salt, because DIBER was not written for this purpose and because the feature has not yet been benchmarked.

2 Methods

2.1 Training and test data

Crystal structures solved at 3.0 Å resolution or better were downloaded from Protein Data Bank

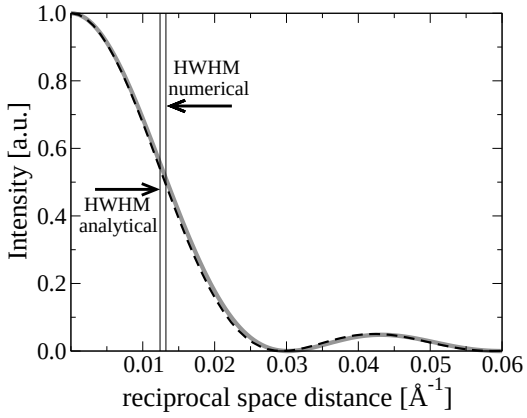

Figure 3: Longitudinal intensity profile of the 3.4 Å peak of dsDNA. The analytical (broken line, formula 8 in Appendix A) and numerical (continuous gray line) results apply to a complete 10 base pair dsDNA helix. Vertical lines and arrows indicate the estimates for the half-width at half-maximum (HWHM) according to formulas 8 and 11 in Appendix A.

(PDB, release date 23 March 2009) together with corresponding experimental diffraction data. Duplicates or near duplicates (90% sequence identity cut-off) were removed from the set. Structures containing RNA or nucleic acids of less than 2 standard Watson-Crick base pairs (as recognized by 3DNA [11]) were removed as well. The final set contained 10580 protein only, 791 protein-DNA and 258 DNA only crystal structures. Protein-DNA structures were further subdivided into 762 B-DNA structures (containing at least 2 neighbouring base pairs of double stranded B-DNA) and 29 others. Similarly, DNA only structures were partitioned into 151 B-DNA structures and 107 others. All DNA containing structures were checked for continuous helices. For every double stranded DNA molecule, the centroids of the four terminal nucleotides of both ends were calculated, stored in pairs of spatially close ends, and expanded by crystallographic symmetry. DNA was classified as continuous if at least one centroid on each end was within 5 Å of another centroid. This procedure should treat DNA duplexes with sticky ends correctly. However, it does not take into account unusual arrangements (like DNA on histones). Therefore, the set was also manually curated. Finally, we identified structures with translational non-crystallographic symmetry in all three sets (protein only, protein-DNA and DNA only). These were defined by the presence of strong off-origin peaks in the native Patterson maps (above 40% of the origin peak height). All reported calculations are based on experimental diffraction data. Structural information was only used to select and classify datasets according to their macromolecule content.

The support vector machine algorithm assumes that the classifier will operate on data drawn from the same distribution as the training data. In DIBER, equal a priori probabilities for obtaining DNA, protein-DNA and DNA crystals are assumed, which is not reflected by the actual numbers of available datasets for the three classes. Therefore we had to rebalance the data artificially. For initial tests, we randomly pruned protein datasets and replicated DNA datasets so that their numbers matched the number of protein-DNA structures. For the final optimization, we took the opposite approach, and replicated protein-DNA and DNA structures until their numbers were equal to the number of protein structures in the set. The classification performance was estimated using a repeated stratified sub-sampling validation procedure. Classifiers were trained with equal numbers of structures from each class (roughly 50% of instances). The remaining structures (in general, unequally distributed among classes) were used for testing. The average error rate of 100 training and testing cycles was used as an estimate of the true error rate. All graphs were prepared with the GRACE (http://plasma-gate.weizmann.ac.il/Grace/) or Matplotlib [9] software.

2.2 Program implementation

DIBER is written in C/C++ and comprises less than 3000 lines of newly written source code. The program relies extensively on CCP4 [5] and CLIPPER [6] libraries to handle keyword parsing, crystal symmetry issues and diffraction data formats. In addition, the support vector machine LIBSVM libraries [3] are used for training and decision making. DIBER does not include any routines

of the molecular replacement program PHASER [12], but provides an interface to run this program to obtain a score (the likelihood-enhanced fast rotation function rescored with full likelihood target).

2.3 Anisotropy correction

Overall anisotropy was corrected in all DIBER modes. For calculation of local averages, the CLIPPER routines were used, because the resolution dependence of scaling is smooth [6]. Scaling factors were applied to all diffraction data. However, they were calculated without the data in the 3.37 Å to 3.43 Å resolution shell to avoid any degradation of the helix signal. In the PHASER assisted mode of DIBER, the rotation score was calculated after applying the PHASER anisotropy correction [12].

2.4 Normalization of structure factors and intensities

Normalization of structure factors poses similar problems like anisotropy correction. For the calculation of local averages, we used CLIPPER routines, which model the resolution dependence of the average intensity without dividing diffraction data into resolution shells (in order to avoid problems at low-resolution [2]). As for anisotropy correction, the normalization factors were calculated without the 3.4 Å resolution shell, but applied throughout.

2.5 The averaging region (standalone mode)

The size and shape of the characteristic 3.4 Å peak of dsDNA is shown in Figs. 2 and 3. In transverse direction, the profile can be roughly approximated by a step function. In the longitudinal direction, a Gaussian or quadratic function would be a better approximation. Nevertheless, considerations of computational efficiency suggested to use a step function also in this direction. The dimensions of the averaging cylinder (with axis pointing towards the origin of reciprocal space) were tuned to maximize the DIBER performance. The percentage of correctly classified datasets (at all costs) was taken as the criterion of success. A cylinder height of 0.04 $Å^{-1}$ and radius of 0.09 $Å^{-1}$ were found to be optimal.

2.6 The calculation of local averages (standalone mode)

The crystallographically independent part of the 3.4 Å resolution shell was sampled to determine the maximum local average. To cover this region evenly, we tested pre-computed icosahedral sphere coverings [8] with between 522 to 78032 sampling points. The set of 5072 points (corresponding approximately to 3° sampling) provided smooth graphics at acceptable computational cost and was used throughout. Diffraction data were expanded to space group $P1$ to avoid computationally expensive on the fly use of crystallographic symmetry.

2.7 Maximum of the likelihood-enhanced fast rotation function score (PHASER only mode)

PHASER 2.1.1 [12] was used for the molecular replacement calculations with a poly-adenine/poly-thymine dsDNA model generated with 3DNA [11]. The likelihood function was defined with default solvent-related parameters $B_{sol} = 300$ $Å^2$, $f_{sol} = 0.95$. Their exact values have only a minor influence on the classification performance. Parameters defining model quality (expected root-mean-square coordinate error RMS and a fraction of the scattering power f_p) were optimized with respect to the classification performance and set to $RMS = 0.5$ $Å^2$ and $f_p = 0.5$. As input for the classifier, the largest log-likelihood gain (LLG) score of the likelihood-enhanced rotation function of order 1 (LERF1) [14] after rescoring its top 100 solutions with the Sim maximum likelihood rotation function (MLRF) [13] was used. Only the diffraction data in the resolution range between 3.18 Å and 3.65 Å were used for calculations. Henceforth, we will for simplicity refer to this score as the rotation score.

2.8 Support vector machine for classification

All DIBER predictions were carried out with Support Vector Machine (SVM) classifiers implemented in the LIBSVM library [3], with input data scaled linearly from 0 to 1. Input data and kernel parameters were dependent on DIBER mode. In standalone mode, the cube root of the reciprocal unit cell volume and largest local intensity average were used. The kernel parameters were $\gamma = 0.02$, $C = 500.0$. In PHASER only mode, the rotation score replaced the largest local intensity average. Moreover, $\gamma = 0.01$ instead of $\gamma = 0.02$ was used. In combined mode, the cube root of the reciprocal unit cell volume, largest local intensity average and rotation score were input to the classifier, and kernel parameters were set to $\gamma = 2.0$, $C = 500.0$.

3 Results

3.1 DIBER overview

DIBER requires a (binary) CCP4 mtz file with diffraction data to at least 3.0 Å resolution. The program extracts three parameters, (a) the cube root of the reciprocal unit cell volume, (b) the largest local average of reflection intensities (c) a rotation score of a PHASER molecular replacement run. The DIBER classification of a crystal of unknown content can be carried out in standalone mode (parameters a and b), PHASER only mode (parameters a and c) or combined mode (all three parameters). In all three modes, DIBER predicts crystal content with the help of a support vector machine (SVM) classifier and estimates a probability for correct classification that depends on the actual parameters of the unknown structure. As a side product of the classification, DIBER also outputs a plot of local intensity averages in the thin 3.4 Å resolution shell on a stereographic net (and PHASER solutions if available). This information can be useful to derive the orientation of the DNA structure in the crystal.

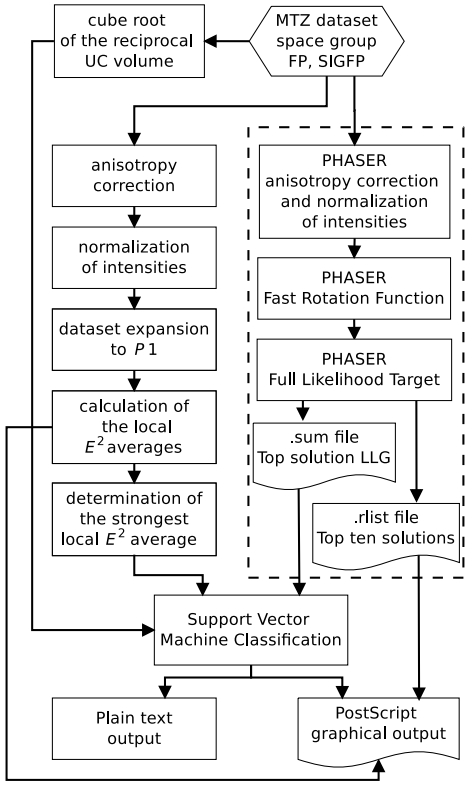

Figure 4: DIBER flow-chart. In standalone mode the predictions are based on the strongest average intensity at 3.4 Å resolution and the cube root of the reciprocal unit cell volume. As an option the prediction reliability may be improved by taking into account the PHASER rotation function score (dashed box).

3.2 Standalone search for the 3.4 Å peak of B-DNA

In standalone mode, DIBER performs a search in a thin reciprocal space shell around 3.4 Å resolution for a small region with many strong, neighbouring reflections. However, the notion of a strong reflection in a newly collected dataset is relative. Average intensities are resolution dependent, and even within a resolution shell, they can vary due to overall temperature factors anisotropy. DIBER corrects for these effects with a global anisotropy correction, followed by a normalization of diffraction intensities (Fig. 4). In order to detect the regions with strong reflections DIBER

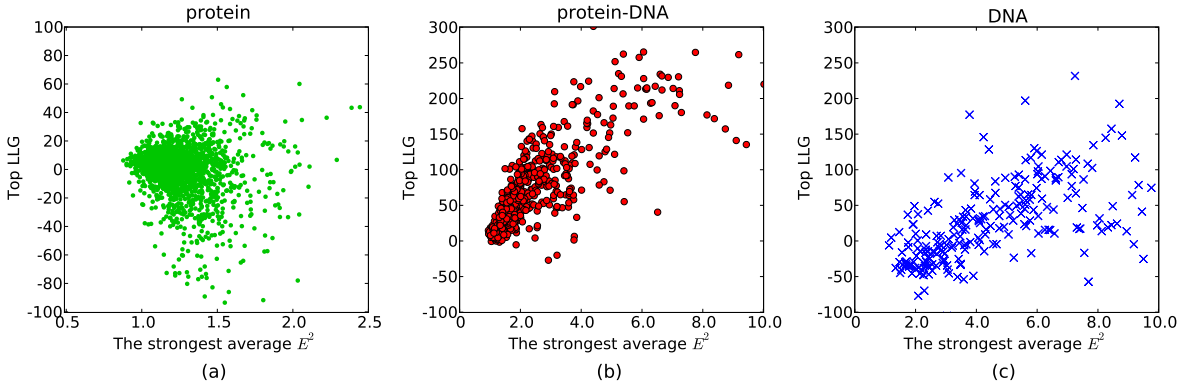

Figure 5: Correlation of the PHASER rotation function score (Top LLG) with the strongest local average of normalized intensity (E^2) for (a) protein only (green circles), (b) protein-DNA (red circles) and (c) DNA only (blue crosses) structures. The corresponding correlation coefficients are -0.10 for protein only, 0.82 for protein-DNA and 0.63 for DNA only structures.

calculates local averages within suitably oriented cylindrical disks (cylinder axis directed towards the origin of reciprocal space) placed at 3.4 Å resolution. The local averages are calculated with appropriate sampling of a crystallographically independent set of orientations, and the largest average is retained as the score for the classifier.

3.3 PHASER search for the 3.4 Å peak of B-DNA

The molecular replacement procedure implemented in PHASER determines the orientation of a search model according to maximum likelihood principles. For a correct orientation of a search model, the probability to observe the experimentally determined data is larger than for the same search model in random orientation. The correct orientation is found by maximizing the increase of (the logarithm of) this probability. Missing information about the model position and hence relative phases of the structure factor contributions of symmetry related molecules is treated with a random walk approximation. We reasoned that we could use the rotation search of PHASER to look for the characteristic 3.4 Å peak of dsDNA. As a dsDNA model, we used an idealized 11 base pair long dsDNA. As the score, we took the log likelihood gain (LLG) of the likelihood-enhanced fast rotation function after rescoring of the top 100 solutions with the full likelihood target (in

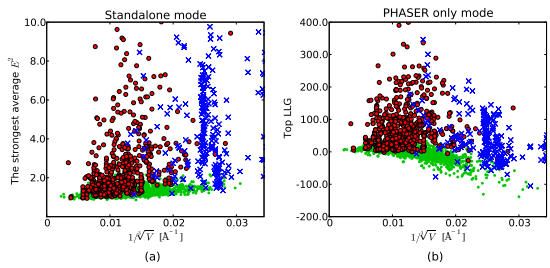

Figure 6: Scatter plot of the classifier input parameters for (a) standalone and (b) PHASER only mode of DIBER. Color codes are the same as in Fig. 5.

this paper referred to as the rotation score).

3.4 Correlation of the largest local intensity average and the rotation score

The largest local intensity average and the rotation score are both measures of the presence or absence of the 3.4 Å peak of B-DNA. In the absence of DNA, the two measures are essentially uncorrelated (Fig. 5a, correlation coefficient -0.10). In contrast, when DNA is present (either with protein or alone), the two measures detect the 3.4 Å peak and are clearly correlated, although not very strongly. The correlation coefficients are 0.82 for

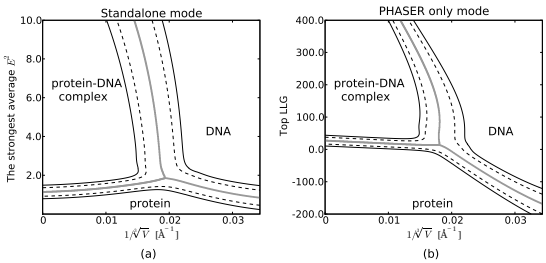

Figure 7: SVM classification boundaries for (a) standalone and (b) PHASER only mode of DIBER. The gray lines correspond to a classification at all costs. The dashed and continuous lines mark the regions of the scatter plot with correct classification probability greater than 80% and 90%, respectively.

protein-DNA crystals and 0.63 for DNA only crystals (Figs. 5b and 5c). These findings prompted us to train the classifier with the two parameters either separately or in combination, always together with the cube root of the reciprocal unit cell volume as an additional input.

3.5 Scatter of the DIBER classifier input for datasets of known content

The two-dimensional scatter-plots presented in Fig. 6 illustrate the spread of classifier input parameters for known structures (with equal representation of protein only, protein-DNA and DNA only structures). Qualitatively, DNA crystals have smaller real space and larger reciprocal space unit cells than crystals that contain protein (with or without DNA). Moreover, a large local intensity average and rotation score correlate with the presence of DNA (alone or with protein), as anticipated. However, there was no clear separation in the scatter plots between structures with continuous DNA and structures with non-continuous DNA (Fig. S1). Apparently, the bendability of DNA tends to break the phase lock of structure factor contributions from distant nucleotide pairs.

3.6 Classifier training with crystals of known content

Optimal separation lines between the three scatter plot regions for DNA, protein-DNA and protein were determined with a support vector machine. The classifier was separately trained in standalone mode (using cube root of the reciprocal unit cell size and largest local intensity average), PHASER only mode (using cube root of the reciprocal unit cell size and rotation score) and combined mode (all three parameters). The training procedure defined not only the optimal division lines between the classes, but also probabilities of the correct classification of a structure of unknown content (Fig. 7).

3.7 Benchmarking DIBER with structures of unknown content

Presented with a diffraction dataset of an unknown crystal structure, DIBER parses the decision tree of Fig. 8 to determine its output. DIBER was benchmarked with the structures from the PDB that were not used in the training phase. Again, equal numbers of structures with only protein, protein-DNA, or only DNA were used for testing. Classification at all costs led to a correct answer for 80-90% of the protein, 70%-80% of the protein-DNA and over 90% of the DNA structures (Fig. 9a and Table 1). About half of all protein and protein-DNA crystals and over 80% of the DNA crystals were located in regions of the scatter plot with greater than 80% correct classification probability (Fig. 9b). Slightly fewer structures could even be classified with greater than 90% probability (Fig. 9c). Except for DNA crystals, the PHASER dependent classification was slightly better than the "quick" standalone classification. The combined mode was best, with insignificant extra computational cost (over the PHASER requirements). Therefore, only the standalone (highest efficiency) and combined (most accurate results) modes of DIBER are available to the user.

3.8 DIBER for curated datasets

The performance figures for DIBER in Table 1 and Fig. 9 were obtained with the non-curated

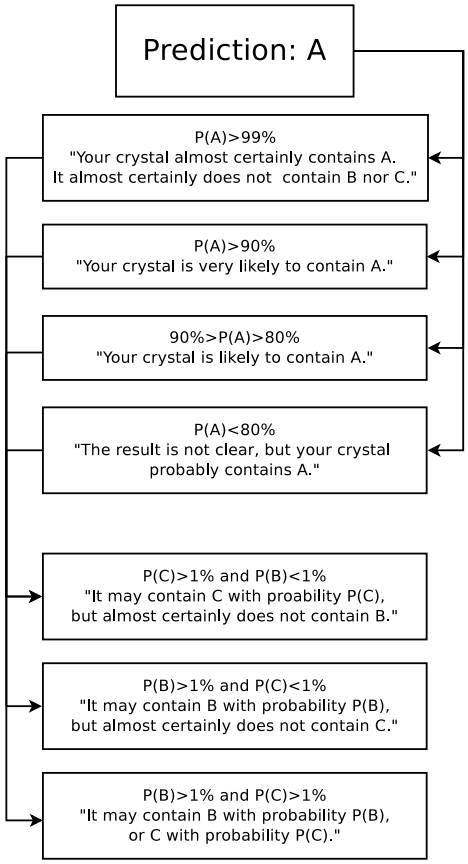

Figure 8: Decision tree of DIBER for prediction A (protein only). Analogous trees are parsed for events B (protein-DNA) and C (DNA only).

training and testing sets described in Methods. Could the performance be improved by excluding unusual data? In a first re-run of DIBER training and testing, we excluded protein-DNA and DNA only structures that did not have at least two neighbouring base pairs of double stranded B-DNA. DIBER performance improved only slightly (Figs. S2 and S3). We also took into account that non-crystallographic translational symmetry might affect the DIBER local averages and PHASER scores, because it can systematically enhance and reduce the intensities in subsets of reflections. Again, the DIBER scores improved, but again the improvement was very slight (Figs. S2 and S3). We also tested the 224 protein structures, 11 protein-DNA structures

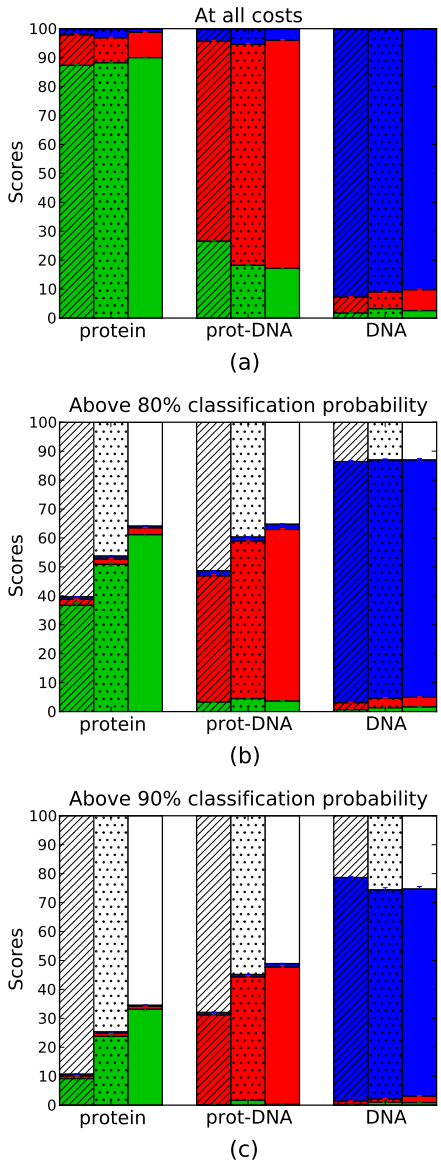

Figure 9: Benchmarking DIBER performance for the classifications (a) at all costs, (b) with greater than 80% classification probability and (c) with greater than 90% classification probability. Structures were divided into protein only, protein-DNA and DNA only. Bars indicate DIBER predictions of the crystal content (green for protein only, red for protein and DNA, blue for DNA only, and white for no prediction). Textures correspond to the standalone (left, hatched), PHASER only (middle, dotted) and combined (right, plain) modes of DIBER.

and 14 DNA structures with significant pseudoorigin peaks separately. Using standard decision criteria, DIBER classified 149 (185) protein, 10 (10) protein-DNA, and 11(11) DNA structures correctly. We attribute this result to the fact that non-translational symmetry tends to affect adjacent reflections in opposite ways, so that local averages are not much perturbed. As the overall DIBER performance did not improve significantly by excluding any unusual structures, checks for short DNA (which would have to be stated by the user) or pseudoorigin peaks (which can be done without user intervention) were not implemented in DIBER.

4 Discussion

4.1 Alternatives to DIBER

Many groups, including our own, have routinely assessed crystal content using spectroscopic and/or biochemical techniques. In addition, there are other sources of information. For small unit cells, the Matthews coefficient will sometimes be sufficient to rule out the bulkier component or a complex. As the calculation can easily be done with available software, the solvent content is not explicitly considered in DIBER. However, it is of course implicit in the unit cell size parameter for the classifier. The presence of DNA (with or without protein) tends to show up as a bump in the Wilson plot at 3.4 Å resolution. It can also manifest itself as a persistent peak that shows up in a fixed location for many higher order symmetry axes (e.g. 8-fold, 9-fold or 10-fold). If the self rotation peaks are due to DNA, much of the signal should be lost if only the low resolution data (below 3.6 Å) are taken into account. A set of two-fold axes on a great circle perpendicular to the main axis strengthens the case for DNA. In many protein-DNA complexes, one of the two-fold axes is also a local symmetry axis of a protein dimer and therefore clearly visible.

4.2 Equal a priori probabilities

The chances to get a protein-DNA co-crystal are case-dependent. Tight interaction favours complexes, loose interaction promotes the crystallization of single components. Unfortunately, good estimates for the chances to get protein-DNA co-crystals are not available. Therefore, DIBER makes the ad hoc assumption that the three possible crystallization outcomes (protein, DNA or both) are equally probable. DIBER output must be read with this assumption in mind.

4.3 Minimal information from the user

DIBER was designed to require minimal input from the user. Therefore, no attempt was made to incorporate information about packing or solvent content in the DIBER predictions. At present, we do not even request the user to state the expected length of the DNA duplex that was used in the crystallization experiments, even though the shape of the 3.4 Å peak is affected by this length. The decision was made because kinks and disordered ends of the DNA duplex are hard to predict, but can drastically affect the effective length.

4.4 Minimal input to the classifier

We also used a minimal number of parameters for the classifier. The reciprocal unit cell size was represented by a single parameter (the cube root of its volume). The overall anisotropy of the diffraction data, which was obtained as a by-product of the anisotropy correction, was not used at all. These simplifications were justified because the additional parameters mostly help to distinguish DNA only crystals from all others which can already be done based on the unit cell size alone.

4.5 Alternative measure of unit cell size

The inverse of the smallest unit cell dimension was tested as an alternative to the cube root of the reciprocal lattice volume as input for the classifier. Results were slightly inferior and therefore this option was not used (Figs. S4 and S5). Using the three unit cell constants separately improved performance only very marginally (data not shown) and was therefore not implemented.

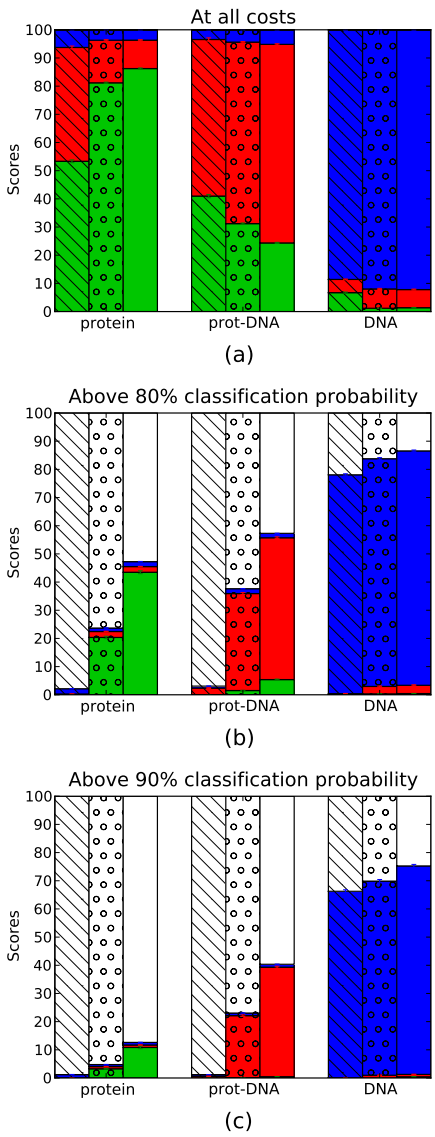

Figure 10: Quality of DIBER predictions in PHASER only mode versus the resolution range of input diffraction data. PHASER rotation scores were calculated with default parameters ($RMS = 1.5$ Å^2, $f_p = 0.5$) for the diffraction data without the 3.2-3.7 Å resolution shell (left, hatched), for all diffraction data (middle, dotted) and for the diffraction data in the 3.2-3.7 Å resolution shell (right, plain). The classification was done (a) at all costs or with a classification probability above (b) 80%, or (c) 90%. Results are colour-coded like in Fig. 9

4.6 Alternative molecular replacement scores

First, we considered to substitute MOLREP [16] for the PHASER molecular replacement scores. However, MOLREP was not written to deal with diffraction data in a thin shell only, and the scores were not useful for classification (data not shown). Next, we tested the PHASER Z-score (the number of standard deviations above the mean) as an alternative to the log likelihood gain. DIBER performance deteriorated for reasons that are currently not clear (Fig. S6). We also considered to use all diffraction data rather than the 3.4 Å shell for PHASER molecular replacement experiments. The classification performance of DIBER dropped again and was also poor for the control experiment with all data outside the 3.4 Å resolution shell (Fig. 10). We also attempted to replace the rotation score of PHASER with the corresponding translation score using either only the 3.4 Å data shell or all diffraction data. In both cases, DIBER performance was no better than with the rotation score, but calculations took much longer (data not shown).

4.7 Data outside the 3.4 Å resolution shell

It is surprising that DIBER performs best with the diffraction data in a thin shell around 3.4 Å resolution. Clearly, the information content in the rest of the data cannot be negative, so the data must be used in the wrong way. We know already that the characteristic 1.5 Å peaks of protein α-helices can be detected in favourable cases (data not shown). For high resolution data, this could be used to further confirm the distinction between protein structures (with or without DNA) and structures of DNA alone. We also know that peaks in the self-rotation function (which can be calculated without the data in the 3.4 Å shell) can be diagnostic for the presence of DNA. The information could help to distinguish structures that contain DNA (with or without protein) from structures of protein alone, which is not always possible with the current version of DIBER. In order to make good use of the low resolution data, it might already suffice to calculate separate rotation functions for different resolution ranges. The

scores could be combined into a single parameter or input to the classifier as a vector. A careful investigation of these possibilities remains for future work.

A B-DNA characteristic diffraction signals

A.1 The Fourier transforms of helix models

Scattering of B-DNA [4, 17, 7] can be understood by considering a series of models of increasing complexity. In the first step, DNA is approximated as an infinitely long helical string of radius r and pitch P along z (coordinates $x = r\cos(2\pi z/P) = r\cos(\phi)$, $y = r\sin(2\pi z/P) = r\sin(\phi)$ and $z = z$). The Fourier transform of this structure $T(R, \psi, n/P)$ is 0 except in layers at distances n/P from the origin. In the nth layer, the Fourier transform can be expressed in reciprocal, cylindrical coordinates R and ψ in terms of the nth-order Bessel function [4]

$$T(R, \psi, n/P) = J_n(2\pi Rr)\exp\left(in(\psi + \frac{1}{2}\pi)\right) \quad (1)$$

The cylinder symmetry of the modulus of this function reflects the equivalence of rotations around and translations along the helix axis. The distance of the first maximum from the origin increases with the order of the Bessel function. Therefore the well-known cross is is seen in sections that include the (reciprocal space) helix axis (Figs. 1a).

In the second step, an infinitely thin and long double helix is considered. The two strands are related to each other by a rotation around a twofold axis perpendicular to the helix axis. In the model, the infinitely long and thin helix strands have no orientation. Within this approximation, their relationship can be described by a translation along the helix axis. The axial shift $0.4 \times P$ models the non-equivalence of the major and minor grooves of DNA and translates into a phase shift $0.4 \times n \times 2\pi$ for the nth layer of the diffraction pattern. The phase shift is 0 for the 0th layer. However, it is close to odd multiples of 2π for the 1st and 4th layer, which therefore have very little intensity (Fig. 1b).

In the third step, a discrete single helix made of point scatterers of axial distance p is considered. This structure can be described as the product of two functions that describe a helical string and a set of planes of spacing p. The Fourier transform of a set of planes of distance p is a set of planes in reciprocal space $1/p$ apart. By the convolution theorem, the Fourier transform of the discrete helix is the Fourier transform of the helical string with its origin placed at each of the points $(0, 0, 0)$, $(0, 0, \pm 1/p)$, $(0, 0, \pm 2/p)$, etc [4]. In practice, the strong decrease of intensity with resolution attenuates the crosses with origins other than (0, 0, 0). Often only the halos of the $(0, 0, 0)$ peak at $(0, 0, \pm 1/p)$ are recognizable (Fig. 1c).

In the fourth step, the same transition as in the third step is made for double helices. Assuming that the point scatterers in the two strands are at the same height, the above argument can be applied again to predict a diffraction pattern like for a non-discrete double helix, but with increasingly weak halos at $(0, 0, \pm 1/p)$, $(0, 0, \pm 2/p)$. The DNA peak at 3.4 Å resolution corresponds to the halo of the origin peak at $(0, 0, \pm 1/p)$ and is due to constructive interference of all scattering points (Fig. 1d).

A.2 Transverse width of the 3.4 Å intensity peak

In order to determine the transverse width of the 3.4 Å peak, it is necessary to calculate the Fourier transform of the B-DNA double helix in the reciprocal space layer (x, y, (3.4 Å)$^{-1}$). The phases in this layer are unaffected by 3.4 Å translations along the helix axis. Therefore, the scattering contribution of all base pairs can be calculated by projecting them onto the real space xy plane, where they form a filled circle of radius $r_{bps} = 5.0$ Å. Therefore, the Fourier transform can be expressed in reciprocal space cylindrical coordinates R and ψ using the Bessel function identity [1]

$$\int x^n J_{n-1}(x)dx = x^n J_n(x) \quad (2)$$

		Crystal content		
		protein	protein-DNA	DNA
Classification results	DNA	$2.1 \pm 0.1\%$ ($1.2 \pm 0.1\%$)	$4.3 \pm 0.2\%$ ($4.0 \pm 0.1\%$)	$92.7 \pm 0.2\%$ ($90.0 \pm 0.2\%$)
	protein-DNA	$10.5 \pm 0.3\%$ ($8.8 \pm 0.2\%$)	$69.1 \pm 0.3\%$ ($78.8 \pm 0.3\%$)	$5.5 \pm 0.2\%$ ($7.2 \pm 0.2\%$)
	protein	$87.4 \pm 0.2\%$ ($90.2 \pm 0.3\%$)	$26.6 \pm 0.2\%$ ($17.2 \pm 0.2\%$)	$1.8 \pm 0.1\%$ ($2.6 \pm 0.1\%$)

Table 1: Benchmarking DIBER performance for the classifications at all costs in standalone (combined) mode.

as

$$T_{bps}(R,\psi) = \frac{1}{\pi r_{bps}^2}\int_0^{r_{bps}}\int_0^{2\pi}\exp(2\pi i R r\cos(\phi-\psi))r dr d\phi = \frac{2}{r_{bps}^2}\int_0^{r_{bps}} r J_0(2\pi R r)dr = \frac{1}{\pi r_{bps} R}J_1(2\pi r_{bps} R) \quad (3)$$

The contribution from the phosphodiester backbone of the double helix (with deoxyribose sugars) can be approximated by two non-discrete helices of radii $r_{bb} = 9.0$ Å (calculated as a weighted average of the backbone atom positions). As the 3.4 Å peak is a halo of the origin peak, it suffices to calculate the transverse width of the latter. For this purpose, the two helices can be replaced by their projections onto the xy plane. The projections coincide and form a circle of radius r_{bb}. As already implied by equation 1 for the special case $n = 0$, the Fourier transform of a circle is a Bessel function of order 0. Up to a multiplicative factor (discussed below) the Fourier transform in the $(x, y, (3.4\ \text{Å})^{-1})$ plane can therefore be written

$$T_{bb}(R,\psi) = J_0(2\pi r_{bb} R) \quad (4)$$

The scattering of the complete dsDNA can be calculated by adding the base pair and backbone atom contributions with proper weights and phases. A simple, but tedious calculation suggests a weighing of 3:1 for the contributions of bases and backbone. Interference on the axis is in antiphase, due to the position of the strongly scattering phosphorus atoms halfway between base pairs (along the helix axis).

$$|T_{dsDNA}(R,\psi)| = |T_{bps}(R,\psi) - T_{bb}(R,\psi)| = |3\frac{1}{\pi r_{bps} R}J_1(2\pi r_{bps} R) - J_0(2\pi r_{bb} R)| \quad (5)$$

Both $\frac{1}{\pi r_{bps} R}J_1(2\pi r_{bps} R)$ and $J_0(2\pi r_{bb} R)$ tend towards 1 as $R \to 0$, and both decrease with increasing R. As the $\frac{1}{\pi r_{bps} R}J_1(2\pi r_{bps} R)$ term decreases faster, the net sum describes a function with a maximum at $R \approx 0.04$ Å^{-1}.

The predictions of analytical formulas 3, 4 and 5 were tested against diffraction patterns of ideal B-DNA generated with the program 3DNA [11]. The agreement is excellent for base pairs (Fig. 2a), backbones (Fig. 2b) and for complete B-DNA (Fig. 2c). The cylinder radius 0.09 Å^{-1}, which maximizes the performance of the DIBER classifier in stand-alone mode, is only slightly smaller than the distance from the axis to the first minimum (Fig. 2c).

A.3 Longitudinal width of the 3.4 Å intensity peak

In order to determine the longitudinal width of the 3.4 Å peak, it suffices to know the Fourier transform on the helix axis in reciprocal space. Fortunately, this can be calculated by the projection

theorem as the one dimensional Fourier transform of the electron density projection on the real space helix axis. Therefore, the structure factor contributions of nucleotide pairs $m = 0, 1, \ldots, N-1$ differ only by phase factors $e^{2\pi i \zeta m p}$. These depend on the wave number ζ in the direction of the helix axis and on the axial-spacing $p = 3.4$ Å. For $\zeta = 1/p$ the phase factors are all equal to 1. For $\zeta = (1+\delta)/p$ with (dimensionless) small, but non-zero δ, complex numbers with non-trivial phase relationships are added. The combined structure factor $F(\zeta)$ (due to residues $0, 1, \ldots, N-1$) can be expressed as the product of the structure factor for a single nucleotide pair $F_s(\zeta)$ with a geometric series.

$$F(\zeta) = F_s(\zeta) \cdot \sum_{m=0}^{N-1} e^{2\pi i \zeta m p} = F_s(\zeta) \cdot \frac{1 - e^{2\pi i \zeta p N}}{1 - e^{2\pi i \zeta p}} \quad (6)$$

$$I(\zeta) = = F(\zeta)F^\star(\zeta) = |F_s^2(\zeta)| \frac{\sin^2(\pi \zeta p N)}{\sin^2(\pi \zeta p)} = \left| F_s\left(\frac{1+\delta}{p}\right)\right|^2 \frac{\sin^2(\pi N \delta)}{\sin^2(\pi \delta)} \quad (7)$$

The second term in the product is maximal for $\delta = 0$ and shrinks to 0 for $\delta = 1/N \ll 1$. In this interval, the first term is roughly constant and can be replaced by its value for $\delta = 0$. With this approximation and the well-known expansion $\sin(x) = x \cdot (1 - x^2/6 + \ldots)$ for $|x| \ll 1$, the intensity can be written as:

$$I(\zeta) = F(\zeta)F^\star(\zeta) \approx \left| F_s\left(\frac{1}{p}\right)\right|^2 \frac{\sin^2(\pi N \delta)}{\sin^2(\pi \delta)} \quad (8)$$

$$\approx N^2 \left| F_s\left(\frac{1}{p}\right)\right|^2 \left(\frac{1 - \frac{(\pi N \delta)^2}{6}}{1 - \frac{(\pi \delta)^2}{6}} \right)^2 \quad (9)$$

This is down to 50% of the value for $\delta = 0$ for:

$$\delta_l = \frac{\sqrt{6 - 3\sqrt{2}}}{\pi \sqrt{N^2 - 1/\sqrt{2}}} \approx \frac{\sqrt{6 - 3\sqrt{2}}}{\pi N} \text{ for } N \gg 1 \quad (10)$$

It translates into a half-width at half-maximum (in wave numbers):

$$HWHM_l = \frac{\delta_l}{p} \approx \frac{\sqrt{6 - 3\sqrt{2}}}{\pi p N} \approx 0.12 \text{ Å}^{-1} \frac{1}{N} \quad (11)$$

The more exact calculation agrees reasonably well with the rough estimate of the Introduction:

$$HWHM_{longitudinal} = 0.15 \text{ Å}^{-1} \frac{1}{N} \quad (12)$$

The cylinder height 0.04 Å^{-1}, which maximizes the performance of the DIBER classifier in stand alone mode, must be compared with the full-width at half-maximum, which is approximately 0.24 Å$^{-1}/N$, and with the distance between first minima, which is approximately twice larger. Apparently, the optimal height of the averaging cylinder is in between the distance between first minima and the full-width at half-maximum for most dsDNA helices.

Acknowledgements

This work was supported by a Polish Ministry of Science and Higher Education grant to MB (N N204 240834). We are grateful to Dr Honorata Czapinska for proofreading the manuscript

References

[1] G.B. Arfken and H.J. Weber. *Mathematical Methods for Physicists, Sixth Edition.* Academic Press, 2005.

[2] M. Bochtler and G. Chojnowski. The highest reflection intensity in a resolution shell. *Acta Crystallographica Section A*, 63(2):146–155, Mar 2007.

[3] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] W. Cochran, F. H. Crick, and V. Vand. The structure of synthetic polypeptides. I. The transform of atoms on a helix. *Acta Crystallographica*, 5(5):581–586, Sep 1952.

[5] Collaborative Computational Project, Number 4. The *CCP*4 suite: programs for protein crystallography. *Acta Crystallographica Section D*, 50(5):760–763, Sep 1994.

[6] K. Cowtan. The clipper c++ libraries for x-ray crystallography. *IUCr Commission on Crystallographic Computing Newsletter*, 2:4–9, 2003.

[7] R.E. Franklin and RG Gosling. Molecular Configuration in Sodium Thymonucleate. *Nature*, 171:740–741, 1953.

[8] R. H. Hardin, N. J. A. Sloane, and W. D. Smith. *Tables of spherical codes with icosahedral symmetry. Published electronically at http://www.research.att.com/˜njas/icosahedral.codes/*, 2000.

[9] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, 9(3):90–95, 2007.

[10] A. Klug, F. H. C. Crick, and H. W. Wyckoff. Diffraction by helical structures. *Acta Crystallographica*, 11(3):199–213, Mar 1958.

[11] X. J. Lu and W. K. Olson. 3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, September 2003.

[12] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read. *Phaser* crystallographic software. *Journal of Applied Crystallography*, 40(4):658–674, Aug 2007.

[13] Randy J. Read. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallographica Section D*, 57(10):1373–1382, Oct 2001.

[14] L. C. Storoni, A. J. McCoy, and R. J. Read. Likelihood-enhanced fast rotation functions. *Acta Crystallographica Section D*, 60(3):432–438, Mar 2004.

[15] Giedre Tamulaitiene, Arturas Jakubauskas, Claus Urbanke, Robert Huber, Saulius Grazulis, and Virginijus Siksnys. The crystal structure of the rare-cutting restriction enzyme SdaI reveals unexpected domain architecture. *Structure*, 14(9):1389–1400, Sep 2006.

[16] A. Vagin and A. Teplyakov. *MOLREP*: an Automated Program for Molecular Replacement. *Journal of Applied Crystallography*, 30(6):1022–1025, Dec 1997.

[17] MHF Wilkins, WE Seeds, AR Stokes, and HR Wilson. Helical structure of crystalline deoxypentose nucleic acid. *Nature*, 172(4382):759–762, 1953.

A Supplementary material

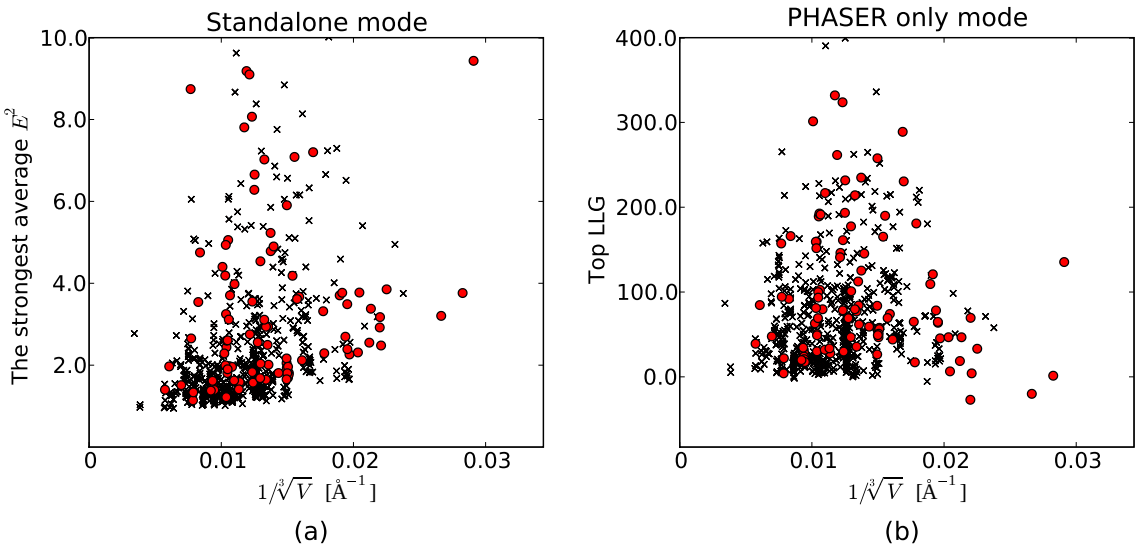

Figure S1: Comparison of characteristic signals of continuous (red circles) and discontinuous (black crosses) DNA molecules in the crystal. The (a) largest local intensity average and (b) top rotation score are plotted against the cube root of the reciprocal unit cell volume.

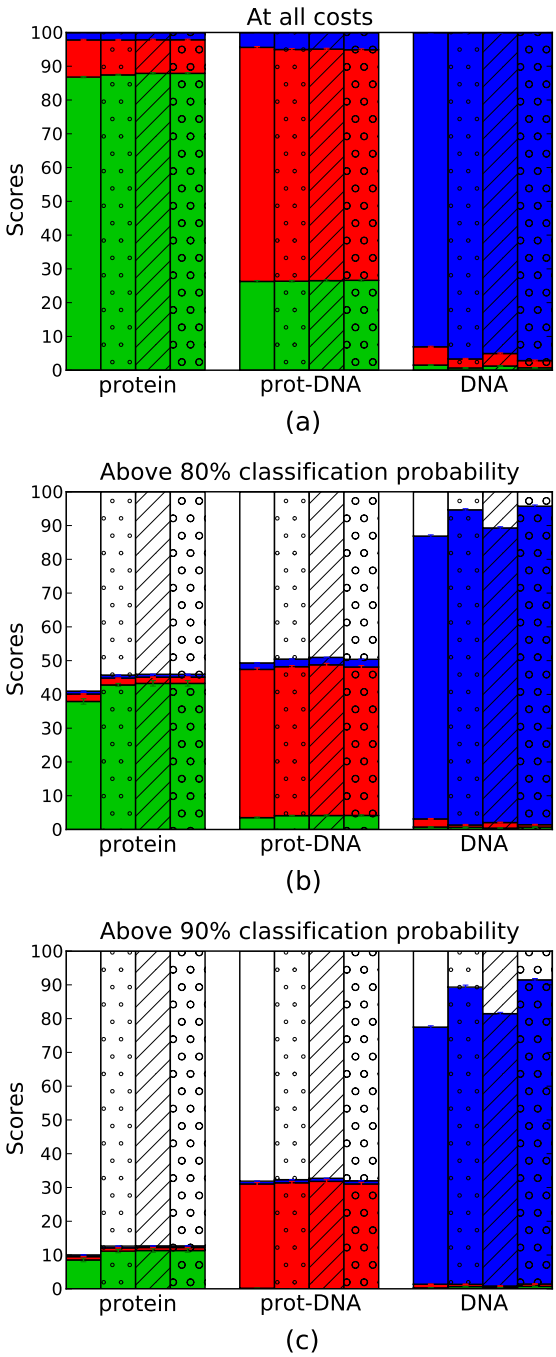

Figure S2: Standalone mode classifier performance for different training and test sets. Plain bars illustrate the results for the control set with all structures. Dotted bars describe the results after removing structures with DNA, but with less than two base pairs in B-DNA conformation. Hatched bars show the results for the set after removing structures with pseudoorigin peaks (threshold 40% of the origin peak). Bars with open circles summarize the results after removing structures with pseudoorigins or very short B-DNA fragments. Colour coding is like in Fig. 9 (green for protein only, red for protein and DNA, blue for DNA only and white for no prediction). Panels (a), (b) and (c) are for classification at all costs and with greater than 80% and 90% correct classification probability, respectively.

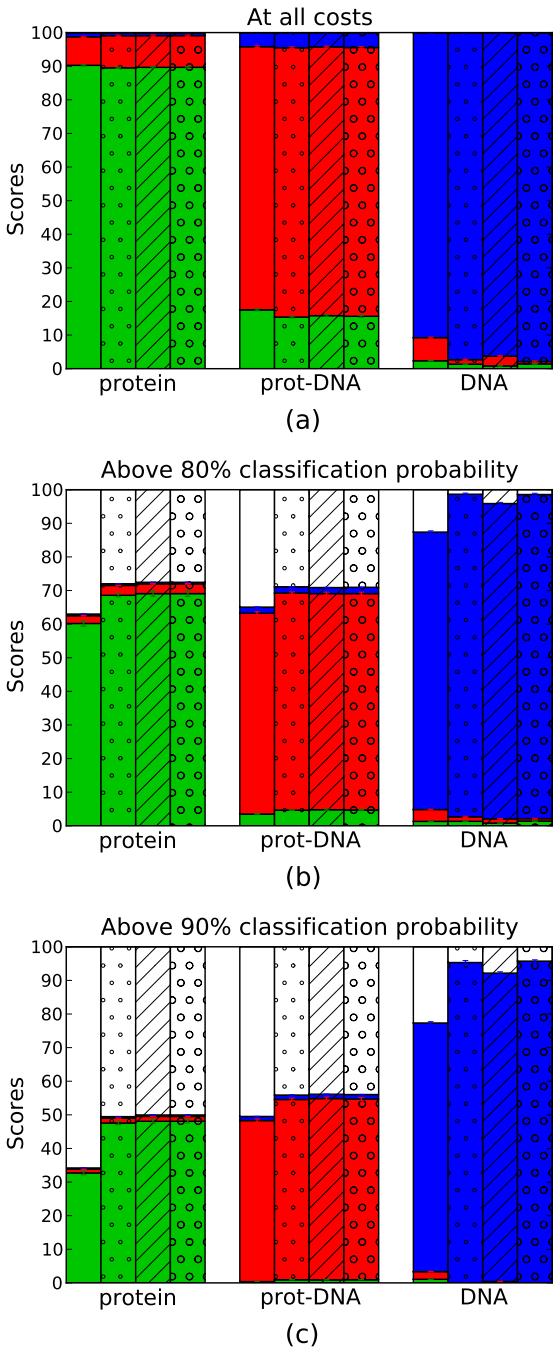

Figure S3: Combined mode classifier performance for different training and test sets. The same symbols and colours as in Fig. S2 are used.

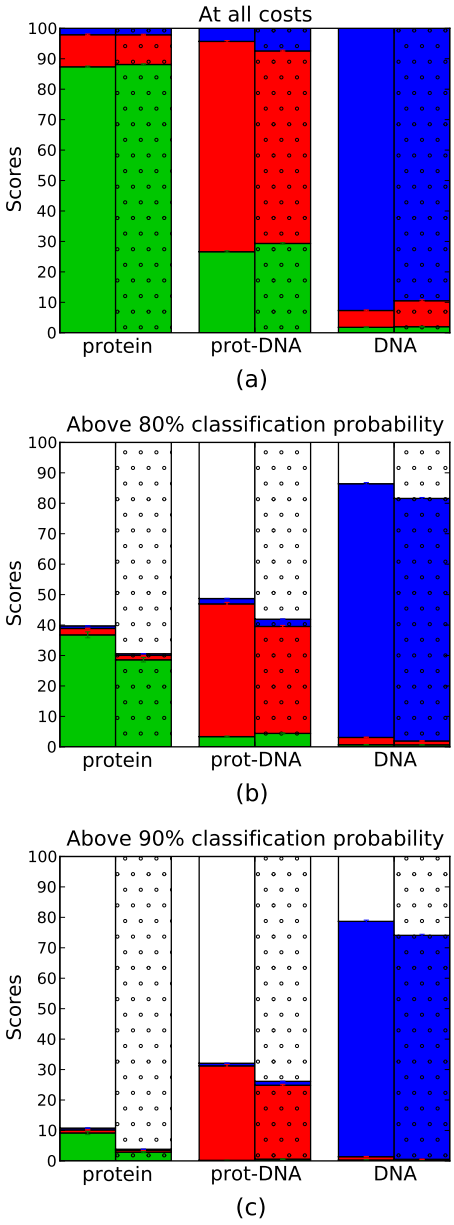

Figure S4: Standalone mode classifier performance with different measures of unit cell size. The cube root of the reciprocal unit cell volume (plain bars) is compared with the inverse of the smallest unit cell dimension (dotted bars). Colours and panels are as in Fig. S2.

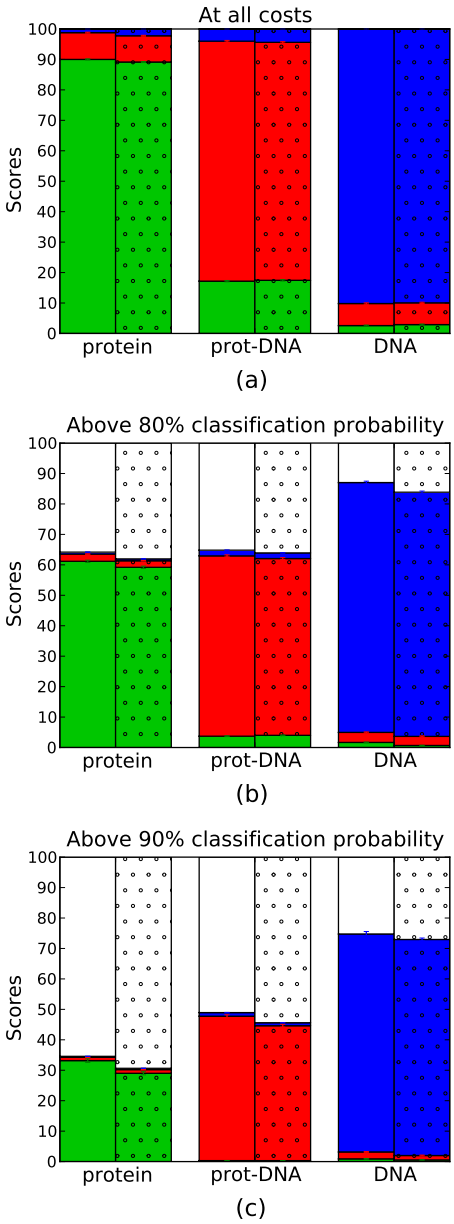

Figure S5: Combined mode classifier performance with different measures of unit cell size. The same symbols and colours as in Fig. S4 are used.

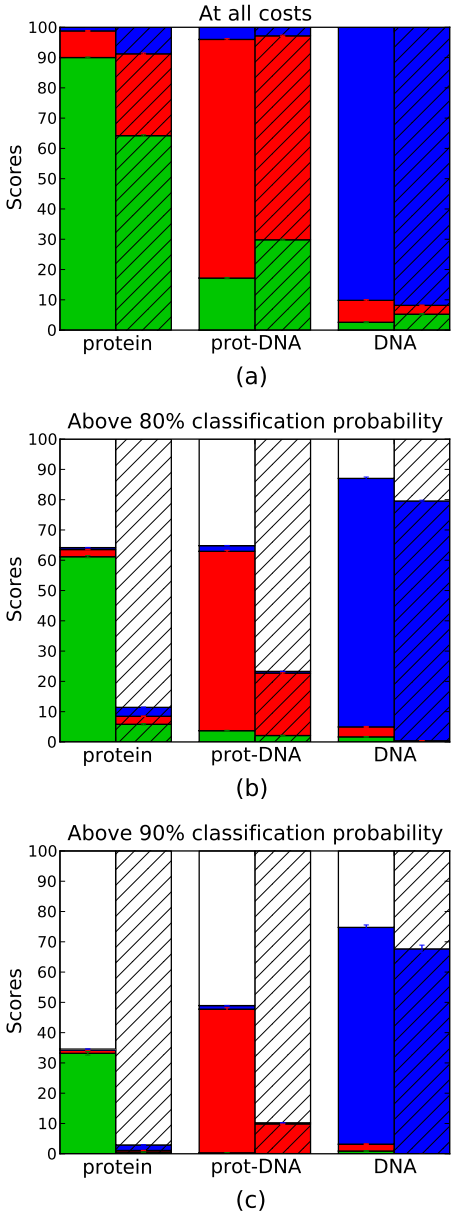

Figure S6: The rotation score (plain bars) versus Z-score (hatched bars) as an input parameter for the classifier in *Phaser* only mode. Both scores were calculated with the same *Phaser* settings and combined with the standard cube root of the reciprocal unit cell volume for classification. Colour coding is like in Fig. S2.