

Volume 52 (2019)

Supporting information for article:

Martini bead form factors for nucleic acids and their application in the refinement of protein-nucleic acid complexes against SAXS data

Cristina Paissoni, Alexander Jussupow and Carlo Camilloni

## S1. Metainference

In order to integrate experimental data with prior information, PLUMED-ISDB makes use of metainference (Bonomi et al., 2016), a method based on Bayesian inference.

According to metainference, given a set of scattering vectors $q$ and the measured intensities $I_{q}$, considering that the global error can be modelled by a Gaussian per data point (Franke et al., 2015) and that the measured and calculated intensities are defined modulo a multiplicative constant $\lambda$, one can show that an optimal balance between the force-field energy and the experimental data can be obtained by defining the metainference energy, $E_{M I}$, as (Löhr et al., 2017):
$E_{M I}=E_{F F}+\frac{k_{B} T}{2} \sum_{r, q} \frac{\left[I_{q}-\lambda f_{q}(X)\right]^{2}}{\left(\sigma_{r, q}^{B}\right)^{2}+\left(\sigma_{r, q}^{S M}\right)^{2}}+E_{\sigma}$,
where $E_{F F}$ is the energy of the force field, $\mathrm{k}_{\mathrm{B}}$ the Boltzmann constant, $T$ the temperature, $f_{q}(\boldsymbol{X})$ the calculated intensity (forward model) for the configuration $X, \sigma_{r, q}^{B}$ is an uncertainty parameter that describes random and systematic errors, $\sigma_{r, q}^{S E M}$ is the standard error of the mean related to the conformational averaging, and $E_{\sigma}$ is an energy term that accounts for normalization of the data likelihood and error priors. $\lambda$ and $\sigma_{r, q}^{B}$ are sampled along with the MD by a Monte Carlo. The sum runs over the set of selected $q$ and, optionally, over $r$ copies of the simulation. Importantly, if conformational averaging is not considered ( $r=1$ ) then the sum runs only over $q, \sigma_{q}^{S E M}=0$, and metainference becomes equivalent to the Inferential Structure Determination approach (Rieping et al., 2005). It is worth observing that metainference can deal with both noise in the data and the approximations involved in the SAXS calculation using the framework provided by Bayesian modelling, which infers uncertainty parameters along with the model of the system. Herein, the inferred parameters $\sigma_{r, q}^{B}$ aim to include all the known sources of errors and uncertainties: these are not limited to the random errors but include also possible systematic errors and the inaccuracies of the forward model (i.e. the prediction of the observables from the 3D-structures which is often based on approximate models). In principle, explicit experimental errors can be used to set the lower limit values of $\sigma_{r, q}^{B}$, but this is generally not required.

## S2. PLUMED-ISDB implementation of hybrid all-atom/ coarse-grain SAXS calculations

We implemented in the PLUMED-ISDB module (Bonomi \& Camilloni, 2017) both the Martini form factors (which can be activated using the keyword MARTINI within the SAXS collective variable) and the atomic scattering factors, corrected by the excluded volume (which can be instead activated with the ATOMISTIC keyword). In addition, using the PARAMETERS keyword, it is possible to assign custom structure factors using a polynomial expansion to any order.

The SAXS results can be printed into an output file and, in case of a running MD simulation, can also be used in combination with metainference (or other methods) to restrain the simulation. It is worth noting that the flexibility of PLUMED allows us to adopt a multiple time-step protocol for the integration of SAXS data in simulations, i.e. applying the metainference bias only at every few time steps (Ferrarotti et al., 2015). This can be useful to further speed up the simulations and is fully justified in the case of SAXS data since the temporal fluctuations of this variable are slower than the ones in atomistic coordinates (Kimanius et al., 2015).

In the following we show an example of a PLUMED input file, to be used to activate the hybrid multi-resolution mode for SAXS-driven MD simulations. According to this approach, simulations are run with full atomistic details with the preferred MD engine, while PLUMED is exploited for the back-calculation of scattering intensities with Martini form factors and to integrate SAXS data in simulations activating metainference.

As exemplified in the box, PLUMED first computes the coordinates of the centre of mass of each Martini bead; the beads are then used by the SAXS action to calculate the Debye equation using the appropriate form factors (in this case the Martini form factors, activated with the MARTINI keyword). In the example below, SAXS intensities are evaluated for 15 scattering vectors, ranging between 0.02 and $0.29 \AA^{-1}$, metainference is activated by DOSCORE and the following keywords set the relevant parameters. Finally, the metainference energy is applied using BIASVALUE every STRIDE step, and SAXS statistics are printed to an output file.

```
MOLINFO STRUCTURE=templateAACG.pdb
WHOLEMOLECULES ENTITYO=1-11104
# Definition of Martini beads position
B1: CENTER NOPBC ATOMS=3,6,8 WEIGHTS=12,12,16
...
B1743: CENTER NOPBC ATOMS=11095,11096,11097 WEIGHTS=14,12,1
martini: GROUP ATOMS=B1,...,B1743
# Compute SAXS intensities and activate Metainference
SAXS ...
        LABEL=saxsdata
        ATOMS=martini
        NOPBC MARTINI
        QVALUE1=0.02 EXPINT1=46.946
        QVALUE15=0.29 EXPINT1=0.138
        DOSCORE NOENSEMBLE SIGMA_MEANO=0
        NOISETYPE=MGAUSS
        SCALEDATA SCALE0=1 SCALE_MIN=0.8 SCALE_MAX=1.2 DSCALE=0.01
        SIGMA0=0.5 SIGMA_MAX=0.5 SIGMA_MIN=0.005 DSIGMA=0.005
... SAXS
saxsbias: BIASVALUE ARG=(saxsdata\.score) STRIDE=10
# Compute statistics
statcg: STATS ARG=(saxsdata\.q_.*) PARARG=(saxsdata\.exp_.*)
# Optionally, other PLUMED actions and print
```

In order to correctly associate each bead to an atom type a PDB file must be provided (templateAACG.pdb in the example above), containing both the atomistic and the coarse-grain coordinates. Attention should be given to the atom numbering, where the number of the first Martini bead should be equal to $1+$ the number of atoms in the atomistic structure, comprising ions and solvent. The renumbering can be easily achieved using the PLUMED tool pdbrenumber, where numbers greater than 100000 are written in the hybrid 36 format.

| ATOM | 1 | O5' | DA | 1 | 91.180 | 16.470 | 79.510 | 1.00 | 0.00 |
| :--- | ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| ATOM | 2 | H5T | DA | 1 | 91.480 | 15.720 | 80.040 | 1.00 | 0.00 |
| $\ldots$ |  |  |  |  |  |  |  |  |  |
| ATOM | 11103 | O1 | HIS | 256 | 57.610 | 60.880 | 50.180 | 1.00 | 0.00 |
| ATOM | 11104 | O2 | HIS | 256 | 57.730 | 62.680 | 51.440 | 1.00 | 0.00 |
| TER |  |  |  |  |  |  |  |  |  |
| ATOM | A0FMN | TE5 | DA | 1 | 91.719 | 18.257 | 78.428 | 1.00 | 0.00 |
| ATOM | A0FMO | BB3 | DA | 1 | 90.293 | 19.617 | 78.733 | 1.00 | 0.00 |
| ... |  |  |  |  |  |  |  |  |  |
| ATOM | A0GZ0 | SC2 HIS | 256 | 53.837 | 60.377 | 52.412 | 1.00 | 0.00 |  |
| ATOM | A0GZ1 | SC3 | HIS | 256 | 54.179 | 59.582 | 50.761 | 1.00 | 0.00 |

## S3. Computational details of the simulations

The two systems investigated were prepared using the amber14sb force field for protein (Maier et al., 2015) with parmbsc1 (Ivani et al., 2015) and the TIP3P water model (Jorgensen et al., 1983), solvated in a triclinic box and neutralized. After an initial energy minimization, the solute was equilibrated using the Berendsen thermostat (Berendsen et al., 1984) to obtain the desired temperature of 300 K . For each system, one 5 ns production run was performed, in which metainference on a single replica was used to introduce SAXS restraints. For the protein/DNA complex, an additional run without the inclusion of experimental information was performed as a reference. During the production runs, the md integrator was employed with a time step of 2 fs, temperature was controlled using the Bussi thermostat (Bussi et al., 2007) and bonds were constrained with the LINCS algorithm (Hess et al., 1998), using a matrix expansion on the order of 6 and 2 iterations per step. The van der Waals and short-range electrostatic interactions were truncated at 0.9 nm , whereas long-range electrostatic interactions were treated with the particle mesh Ewald method (Darden et al., 1993).

In the case of the ComE-comcde DNA-protein complex, both the metainference and the unrestrained simulations were evolved for a total of 5 ns through a series of 20 simulated annealing cycles, with a period of 250 ps each and the temperature varying between 300 and 400 K . Specifically, each cycle consisted of 100 ps at 300 K , a fast increase of the temperature from 300 to $400 \mathrm{~K}, 20 \mathrm{ps}$ at 400 K , and finally a linear cooling from 400 to 300 K in 120 ps . Only structures extracted from the intervals at 300 K in the last 10 cycles were used for analysis. In order to avoid the opening of DNA in the high temperature intervals, in both the simulations we restrained the hydrogen bonds between the first and last two couples of nucleotides adding a harmonic potential centred at 0.3 nm and with a force constant of $1000 \mathrm{~kJ} /\left(\mathrm{mol} \mathrm{nm}^{2}\right)$. Specifically, the restraints were imposed on the distances between oxygens and nitrogens involved in hydrogen bonds for the couples A1-T76, A2-T75, A39-T39, and A37-T40. In the metainference simulation, a set of 15 representative SAXS intensities at different scattering vectors, ranging between $0.02 \AA^{-1}$ and $0.3 \AA^{-1}$, were also added as restraints. These
representative intensities were extracted from the experimental data, where a 15 -point running average was performed to reduce the influence of experimental noise. Metainference was applied every 10 steps, using a single Gaussian noise per data-point and sampling a scaling factor between experimental and calculated SAXS intensities with a flat prior between 0.8 and 1.2. An initial value for this scaling factor was chosen to match the experimental and calculated intensity at the scattering vector $q=0.02 \AA^{-1}$ for the initial model.

In the case of the RNA-protein complex, the metainference simulation was evolved for 5 ns maintaining the temperature at the value of 300 K . Restraints in the form of harmonic upper-wall potentials were applied as described in Kooshapur et al. (2018) to maintain critical protein-RNA interface contacts, salt bridges and protein secondary structures, as found in the related crystal structure (PDB: 6DCL). 43 representative SAXS intensities were used as restraints in metainference, corresponding to scattering vectors between $0.03 \AA^{-1}$ and $0.45 \AA^{-1}$. These intensities were obtained fitting experimental data with a $16^{\text {th }}$ degree polynomial up to scattering value of $0.5 \AA^{-1}$, following the work done in Kooshapur et al. (2018). Metainference was applied every 10 steps, using a single Gaussian noise per data-point and the scaling factor was sampled from a Gaussian prior.

## S4. R-score metric to evaluate the accuracy of coarse-grained SAXS intensities

As an additional metric to assess the accuracy of Martini form factors in computing scattering intensities for nucleic acids, we adopted the $R$-scoring function:
$R=\frac{1}{N} \sum_{q=0}^{q(N)}\left[\frac{I_{A A}(q)-I_{C G}(q)}{\sigma(q)}\right]^{2}$,
where $\sigma(q)=I(q)(q+a) b$, with $a=0.15$ and $b=0.3$ (Stovgaard et al., 2010). This value aims to reproduce the usual $\chi^{2}$ metric in evaluating differences between SAXS profiles, where an empirical standard deviation is adopted since experimental errors are not available in theoretical curves. The form of $\sigma(q)$ and the values of the $a$ and $b$ parameters were chosen as in Stovgaard et al. (2010), to be stricter in the portion of the curve of major interest for structure prediction ( $q$ values lower than 0.5 $\AA^{-1}$ ). In Figure S3, the distribution of $R$-values for RNA and DNA evaluated for $q$ up to $0.5 \AA^{-1}$ is reported, along with the average $R$-value as a function of the scattering vector $q$ used as cut-off. For DNA, $99 \%$ of the structures present an $R$-value lower than 0.1 , with intensities for DNA in B-form being reproduced again slightly better than for the other forms (Figure S4). For RNA, only $68 \%$ of the structures display an $R$-value below 0.1 because of a sharp increase of $R$ for scattering vectors between 0.4 and $0.5 \AA^{-1}$. By decreasing the $q$ cut-off to $0.45 \AA^{-1}$ we found that $95 \%$ of the structures satisfies $R<0.1$ (Figure S5), further confirming this range as optimal for coarse-grain intensities calculations involving RNA molecules.

Table S1 List of the scattering types used and their atomic composition. The sum of the reduced atomic scattering factor at $q=0$ (i.e. $F_{i}(q=0)=\sum_{k \in i} f^{\prime}{ }_{k}$ ) is also reported. The cases in which this sum results in a negative value, and thus need a correction as described in the text, are highlighted in bold.

| Nucleotide | Bead | Number of atoms |  |  |  |  | $F(q=0)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | C | H | N | O | P |  |
| DA | BB1 | 0 | 0 | 0 | 4 | 1 | 32.89 |
| DA | BB2 | 2 | 3 | 0 | 1 | 0 | 3.81 |
| DA | BB3 | 3 | 4 | 0 | 0 | 0 | -1.36 |
| DA | SC1 | 1 | 0 | 1 | 0 | 0 | 6.67 |
| DA | SC2 | 1 | 1 | 1 | 0 | 0 | 5.95 |
| DA | SC3 | 1 | 2 | 2 | 0 | 0 | 11.39 |
| DA | SC4 | 2 | 1 | 1 | 0 | 0 | 6.46 |
| DA | TE3 | 3 | 5 | 0 | 1 | 0 | 2.87 |
| DA | TE5 | 2 | 4 | 0 | 2 | 0 | 8.04 |
| DC | BB1 | 0 | 0 | 0 | 4 | 1 | 32.89 |
| DC | BB2 | 2 | 3 | 0 | 1 | 0 | 3.81 |
| DC | BB3 | 3 | 4 | 0 | 0 | 0 | -1.36 |
| DC | SC1 | 1 | 1 | 1 | 0 | 0 | 5.95 |
| DC | SC2 | 1 | 0 | 1 | 1 | 0 | 11.62 |
| DC | SC3 | 2 | 3 | 1 | 0 | 0 | 5.02 |
| DC | TE3 | 3 | 5 | 0 | 1 | 0 | 2.87 |
| DC | TE5 | 2 | 4 | 0 | 2 | 0 | 8.04 |
| DG | BB1 | 0 | 0 | 0 | 4 | 1 | 32.89 |
| DG | BB2 | 2 | 3 | 0 | 1 | 0 | 3.81 |
| DG | BB3 | 3 | 4 | 0 | 0 | 0 | -1.36 |
| DG | SC1 | 1 | 0 | 1 | 0 | 0 | 6.67 |
| DG | SC2 | 1 | 2 | 2 | 0 | 0 | 11.39 |
| DG | SC3 | 1 | 1 | 1 | 1 | 0 | 10.90 |
| DG | SC4 | 2 | 1 | 1 | 0 | 0 | 6.46 |
| DG | TE3 | 3 | 5 | 0 | 1 | 0 | 2.87 |
| DG | TE5 | 2 | 4 | 0 | 2 | 0 | 8.04 |
| DT | BB1 | 0 | 0 | 0 | 4 | 1 | 32.89 |
| DT | BB2 | 2 | 3 | 0 | 1 | 0 | 3.81 |
| DT | BB3 | 3 | 4 | 0 | 0 | 0 | -1.36 |
| DT | SC1 | 1 | 1 | 1 | 0 | 0 | 5.95 |
| DT | SC2 | 1 | 1 | 1 | 1 | 0 | 10.90 |
| DT | SC3 | 3 | 3 | 0 | 1 | 0 | 4.31 |
| DT | TE3 | 3 | 5 | 0 | 1 | 0 | 2.87 |
| DT | TE5 | 2 | 4 | 0 | 2 | 0 | 8.04 |
| A | BB1 | 0 | 0 | 0 | 4 | 1 | 32.89 |
| A | BB2 | 2 | 3 | 0 | 1 | 0 | 3.81 |
| A | BB3 | 3 | 4 | 0 | 1 | 0 | 3.59 |
| A | SC1 | 1 | 0 | 1 | 0 | 0 | 6.67 |
| A | SC2 | 1 | 1 | 1 | 0 | 0 | 5.95 |
| A | SC3 | 1 | 2 | 2 | 0 | 0 | 11.39 |
| A | SC4 | 2 | 1 | 1 | 0 | 0 | 6.46 |
| A | TE3 | 3 | 5 | 0 | 2 | 0 | 7.82 |
| A | TE5 | 2 | 4 | 0 | 2 | 0 | 8.04 |
| C | BB1 | 0 | 0 | 0 | 4 | 1 | 32.89 |
| C | BB2 | 2 | 3 | 0 | 1 | 0 | 3.81 |
| C | BB3 | 3 | 4 | 0 | 1 | 0 | 3.59 |
| C | SC1 | 1 | 1 | 1 | 0 | 0 | 5.95 |
| C | SC2 | 1 | 0 | 1 | 1 | 0 | 11.62 |
| C | SC3 | 2 | 3 | 1 | 0 | 0 | 5.02 |
| C | TE3 | 3 | 5 | 0 | 2 | 0 | 7.82 |
| C | TE5 | 2 | 4 | 0 | 2 | 0 | 8.04 |
| G | BB1 | 0 | 0 | 0 | 4 | 1 | 32.89 |


| G | BB2 | 2 | 3 | 0 | 1 | 0 | 3.81 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| G | BB3 | 3 | 4 | 0 | 1 | 0 | 3.59 |
| G | SC1 | 1 | 0 | 1 | 0 | 0 | 6.67 |
| G | SC2 | 1 | 2 | 2 | 0 | 0 | 11.39 |
| G | SC3 | 1 | 1 | 1 | 1 | 0 | 10.90 |
| G | SC4 | 2 | 1 | 1 | 0 | 0 | 6.46 |
| G | TE3 | 3 | 5 | 0 | 2 | 0 | 7.82 |
| G | TE5 | 2 | 4 | 0 | 2 | 0 | 8.04 |
| U | BB1 | 0 | 0 | 0 | 4 | 1 | 32.89 |
| U | BB2 | 2 | 3 | 0 | 1 | 0 | 3.81 |
| U | BB3 | 3 | 4 | 0 | 1 | 0 | 3.59 |
| U | SC1 | 1 | 1 | 1 | 0 | 0 | 5.95 |
| U | SC2 | 1 | 1 | 1 | 1 | 0 | 10.90 |
| U | SC3 | 2 | 1 | 0 | 1 | 0 | 5.25 |
| U | TE3 | 3 | 5 | 0 | 2 | 0 | 7.82 |
| U | TE5 | 2 | 4 | 0 | 2 | 0 | 8.04 |

Table S2 PDB codes used for the calculations of the coarse-grained form factors, listed by category. The PDB codes used in the validation are highlighted with a star. The numbers in parenthesis indicate the total number of structures in the category and the number of PDB codes used for the validation, respectively.

| $\begin{aligned} & \text { RNA } \\ & (77,44) \end{aligned}$ |  | 1z58, 357d, 1t0e*, 1t0d, 2oe6*, 1q93*, 1msy*, 1sdr*, 353d*, 361d*, 1dqf*, |
| :---: | :---: | :---: |
|  |  | $1 \mathrm{mme}^{*}, 2 \mathrm{a} 64^{*}, 2 \mathrm{a} 2 \mathrm{e}, 1 \mathrm{y} 0 \mathrm{q}^{*}, 1 \mathrm{x} 8 \mathrm{w}^{*}, 1 \mathrm{u} 9 \mathrm{~s}^{*}, 1 \mathrm{nbs} *, 2 \mathrm{~h} 0 \mathrm{~s}, 1 \mathrm{x} 9 \mathrm{k}^{*}, 1 \mathrm{x} 9 \mathrm{c}, 1 \mathrm{ykq}$, |
|  |  | $1 \mathrm{nuj} *, 1 \mathrm{fir}, 1 \mathrm{i9v}$, 1yfg, 2tra, 434d*, 2il9*, 1kh6, 1xjr, 1k9w*, 2b8r*, 2nok*, |
|  |  | 1csl*, 112x, 397d*, 1duq*, 1jzv, 1duh, 1d4r, 1z43*, 1kxk*, 1i9x*, 1mhk*, |
|  |  | 387d*, 1yzd, 1f1t, 1kfo, 406d, 405d*, 1f27, 1qbp, 433d*, 413d*, 157d*, |
|  |  | 205d*, 255d*, 280d*, 1kd5*, 1p79, 2ao5, 1j9h, 438d*, 2g91, 1g2j, 1sa9*, |
|  |  | 259d*, 2a0p, 2g3s*, 333d, 402d*, 409d, 472d*, 113z*, 377d |
| DNA$(175,121)$ | A-form $(45,36)$ | $118 \mathrm{~d}^{*}, 137 \mathrm{~d}^{*}, 138 \mathrm{~d}^{*}, 160 \mathrm{~d}^{*}, 1 \mathrm{~d} 78^{*}, 1 \mathrm{~d} 79^{*}, 1 \mathrm{dnz}{ }^{*}, 1 \mathrm{kgk}, 1 \mathrm{~m} 77 *, 1 \mathrm{ma} 8,1 \mathrm{mlx}$, |
|  |  | $1 \mathrm{nzg}, 1 \mathrm{jj} 4^{*}$, 1xjx*, 1z7i, 1zex*, 1zey*, 1zf1*, 1zf6*, 1zf8*, 1zf9*, 1zfa*, |
|  |  | 213d, 243d*, 260d*, 295d*, 2d94*, 317d*, 338d, 344d, 345d, 348d*, 349d*, |
|  |  | 368d*, 369d*, 370d*, 371d*, 395d*, 396d*, 399d*, 414d*, 440d*, 9dna*, |
|  |  | $1 \mathrm{vt5}$ *, 1vtb* |
|  | B-form$(72,51)$ | 122d, 123d, 158d*, 183d, 196d*, 1bd1*, 1bna*, 1cw9, 1d23*, 1d3r, 1d49*, |
|  |  | 1d56*, 1d61, 1d8g*, 1d8x*, 1dou*, 1dpn, 1edr, 1ehv*, 1en3*, 1en8*, 1en9*, |
|  |  | $1 \mathrm{ene} *$, 1enn*, $1 \mathrm{fq} 2 *, 1 \mathrm{~g} 75,1 \mathrm{i} 3 \mathrm{t}, 1 \mathrm{lkk}^{*}, 1 \mathrm{j} 81,1 \mathrm{jgr}$ *, $114 \mathrm{j}^{*}, 116 \mathrm{~b}, 1 \mathrm{m6} \mathrm{~g}^{*}, 1 \mathrm{n} 1 \mathrm{o}$, |
|  |  | 1 nvn *, 1nvy*, 1p4y*, 1p54, 1s23*, 1s2r*, 1sgs*, 1sk5*, 1ub8*, 1ve8, 1zf0*, |
|  |  | 1zf3*, 1zf4*, 1zf5*, 1zf7*, 1zfb*, 1zff*, 1zfg*, 232d*, 251d*, 2d25, 307d*, |
|  |  | 355d*, 3dnb*, 403d, 423d*, 428d*, 431d*, 436d, 454d, 455d*, 456d, 460d, |
|  |  | 463d*, 476d*, 477d*, 5dnb*, 9bna* |
|  | Z-form $(39,21)$ | 131d*, 145d, 181d* $, 1 \mathrm{~d} 40,1 \mathrm{~d} 41,1 \mathrm{~d} 48^{*}, 1 \mathrm{~d} 53^{*}, 1 \mathrm{~d} 76,1 \mathrm{da} 2,1 \mathrm{dcg}$ *, 1dj6*, |
|  |  | $1 \mathrm{dn} 4,1 \mathrm{dn} 5,1 \mathrm{dnf}, 1 \mathrm{i} 0 \mathrm{t}^{*}, 1 \mathrm{ick} *$, 1jes, 1 ljx *, 1omk, 1xa2*, 1xam*, 1zna*, 210d, |
|  |  | 211d, 242d, 292d*, 293d*, 2dc9*, 313d, 314d*, 331d*, 336d*, 351d*, 362d*, |
|  |  | 400d |
|  | Quadruplexes$(19,13)$ | 184d*, 190d*, 191d*, 1bqj*, 1cn0*, 1jpq, 111 ${ }^{*}$ *, 1mf5*, 1o0k, 1qyk*, 1qyl*, |
|  |  | 1v3n, 1v3o, 1v3p, 200d*, 241d*, 244d*, 284d*, 352d |

Table S3 List of the coefficients, to be used in a polynomial expansion of the sixth order, for each nucleotide bead.

| Nucl | Bead | A0 | A1 | A2 | A3 | A4 | A5 | A6 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| DA | BB1 | 32.885000 | 0.081799 | -7.317359 | 2.156145 | -3.522632 | 2.306047 | -0.392701 |
| DA | BB2 | 3.806000 | -0.105977 | 9.525375 | -6.129910 | -0.540926 | 1.154291 | -0.215035 |
| DA | BB3 | -1.356000 | 0.589283 | 6.718941 | 4.140509 | -9.658599 | 4.431850 | -0.646573 |
| DA | SC1 | 6.671000 | -0.008714 | 1.632891 | -0.066377 | -1.486329 | 0.785518 | -0.120873 |
| DA | SC2 | 5.951000 | -0.026343 | 2.548643 | -0.490158 | -1.553869 | 0.866302 | -0.135462 |
| DA | SC3 | 11.394000 | 0.008595 | -0.254714 | 0.487188 | -1.745200 | 0.992462 | -0.163519 |
| DA | SC4 | 6.459000 | 0.019918 | 4.179623 | 0.974691 | -5.029504 | 2.553718 | -0.391134 |
| DA | TE3 | 2.874000 | 0.001129 | 12.511672 | -7.675480 | -2.022340 | 2.508371 | -0.494585 |
| DA | TE5 | 8.036000 | 0.004731 | 4.655544 | 0.664241 | -6.621313 | 3.961074 | -0.690758 |
| DC | BB1 | 32.885000 | 0.081899 | -7.324935 | 2.159769 | -3.526121 | 2.310586 | -0.394027 |
| DC | BB2 | 3.806000 | -0.105598 | 9.525277 | -6.121317 | -0.548994 | 1.155929 | -0.214945 |
| DC | BB3 | -1.356000 | 0.555257 | 6.803055 | 4.059247 | -9.610347 | 4.412538 | -0.643151 |
| DC | SC1 | 5.951000 | -0.028999 | 2.595878 | -0.553883 | -1.563951 | 0.889674 | -0.140625 |
| DC | SC2 | 11.621000 | 0.013581 | -0.249130 | 0.487872 | -1.528673 | 0.836949 | -0.133953 |
| DC | SC3 | 5.019000 | -0.032984 | 5.542428 | -0.960815 | -3.710516 | 2.165002 | -0.350234 |
| DC | TE3 | 2.874000 | -0.052355 | 13.092012 | -9.481282 | -0.149586 | 1.755372 | -0.393475 |
| DC | TE5 | 8.036000 | -0.005136 | 4.677057 | 0.483333 | -6.345110 | 3.833885 | -0.673678 |
| DG | BB1 | 32.885000 | 0.081829 | -7.321339 | 2.157679 | -3.523697 | 2.308396 | -0.393483 |
| DG | BB2 | 3.806000 | -0.106181 | 9.541690 | -6.151776 | -0.534624 | 1.155813 | -0.215670 |
| DG | BB3 | $-1.356000$ | 0.574891 | 6.751647 | 4.113009 | -9.633946 | 4.416754 | -0.643399 |
| DG | SC1 | 6.671000 | -0.008866 | 1.633330 | -0.068921 | -1.486835 | 0.786708 | -0.121139 |
| DG | SC2 | 11.394000 | 0.009079 | -0.224755 | 0.495351 | -1.753249 | 0.987674 | -0.161508 |
| DG | SC3 | 10.901000 | 0.022076 | 0.179322 | 0.732532 | -1.955549 | 0.983399 | -0.147636 |
| DG | SC4 | 6.459000 | 0.020184 | 4.177054 | 0.985317 | -5.043549 | 2.561237 | -0.392493 |
| DG | TE3 | 2.874000 | 0.001820 | 12.415070 | -7.473848 | -2.118647 | 2.501126 | -0.486522 |
| DG | TE5 | 8.036000 | 0.006764 | 4.659892 | 0.784825 | -6.864606 | 4.116754 | -0.722491 |
| DT | BB1 | 32.885000 | 0.082201 | -7.330068 | 2.166365 | -3.534657 | 2.314476 | -0.394454 |
| DT | BB2 | 3.806000 | -0.107230 | 9.566750 | -6.202361 | -0.495504 | 1.143006 | -0.214200 |
| DT | BB3 | -1.356000 | 0.567379 | 6.765954 | 4.089761 | -9.615125 | 4.409751 | -0.642398 |
| DT | SC1 | 5.951000 | -0.029265 | 2.596303 | -0.561522 | -1.565326 | 0.893228 | -0.141429 |
| DT | SC2 | 10.901000 | 0.021834 | 0.194630 | 0.723930 | -1.931995 | 0.968563 | -0.145126 |
| DT | SC3 | 4.314000 | -0.077456 | 12.498203 | -7.649942 | -3.003596 | 3.262633 | -0.644986 |
| DT | TE3 | 2.874000 | -0.002512 | 12.435764 | -7.553438 | -2.073635 | 2.512793 | -0.494371 |
| DT | TE5 | 8.036000 | 0.001199 | 4.917623 | 0.656370 | -7.233925 | 4.446366 | -0.794678 |
| A | BB1 | 32.885000 | 0.083391 | -7.360403 | 2.192064 | -3.565057 | 2.333236 | -0.397867 |
| A | BB2 | 3.806000 | -0.107298 | 9.589166 | -6.238736 | -0.482161 | 1.141293 | -0.213909 |
| A | BB3 | 3.594000 | 0.045373 | 9.591789 | -1.292022 | -7.108510 | 4.055712 | -0.633725 |


| A | SC1 | 6.671000 | -0.008553 | 1.632224 | -0.064662 | -1.486942 | 0.785446 | -0.120835 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| A | SC2 | 5.951000 | -0.026066 | 2.543999 | -0.484369 | -1.553574 | 0.864669 | -0.135090 |
| A | SC3 | 11.394000 | 0.008713 | -0.238913 | 0.489194 | -1.752894 | 0.992675 | -0.162913 |
| A | SC4 | 6.459000 | 0.019905 | 4.179750 | 0.976328 | -5.033298 | 2.555999 | -0.391550 |
| A | TE3 | 7.824000 | -0.048810 | 8.215579 | -0.894914 | -9.542937 | 6.331222 | -1.166729 |
| A | TE5 | 8.036000 | 0.016412 | 5.149022 | 0.834197 | -7.590683 | 4.520632 | -0.782608 |
| C | BB1 | 32.885000 | 0.083111 | -7.354321 | 2.186100 | -3.557883 | 2.329187 | -0.397200 |
| C | BB2 | 3.806000 | -0.108108 | 9.616792 | -6.287320 | -0.451266 | 1.133316 | -0.213253 |
| C | BB3 | 3.594000 | 0.044842 | 9.619198 | -1.335828 | -7.072004 | 4.039529 | -0.630982 |
| C | SC1 | 5.951000 | -0.029113 | 2.597004 | -0.555077 | -1.563446 | 0.889562 | -0.140613 |
| C | SC2 | 11.621000 | 0.013661 | -0.259592 | 0.489183 | -1.525505 | 0.836441 | -0.134073 |
| C | SC3 | 5.019000 | -0.032761 | 5.537769 | -0.951050 | -3.711308 | 2.161460 | -0.349186 |
| C | TE3 | 7.824000 | -0.058483 | 8.293199 | -1.125638 | -9.421976 | 6.354417 | -1.183569 |
| C | TE5 | 8.036000 | 0.004935 | 4.926220 | 0.648107 | -7.051000 | 4.260644 | -0.748191 |
| G | BB1 | 32.885000 | 0.083254 | -7.357360 | 2.189148 | -3.561548 | 2.331206 | -0.397523 |
| G | BB2 | 3.806000 | -0.107883 | 9.609308 | -6.274025 | -0.461927 | 1.137370 | -0.213831 |
| G | BB3 | 3.594000 | 0.045145 | 9.612347 | -1.315421 | -7.091505 | 4.047062 | -0.632010 |
| G | SC1 | 6.671000 | -0.008632 | 1.632523 | -0.065672 | -1.486805 | 0.785656 | -0.120889 |
| G | SC2 | 11.394000 | 0.009122 | -0.228690 | 0.496164 | -1.750390 | 0.986492 | -0.161416 |
| G | SC3 | 10.901000 | 0.022087 | 0.170328 | 0.732808 | -1.952920 | 0.983576 | -0.147909 |
| G | SC4 | 6.459000 | 0.020234 | 4.176650 | 0.987378 | -5.044199 | 2.561080 | -0.392438 |
| G | TE3 | 7.824000 | -0.051774 | 8.346067 | -1.029363 | -9.552119 | 6.377766 | -1.178980 |
| G | TE5 | 8.036000 | 0.005251 | 4.710706 | 0.667469 | -6.725387 | 4.036441 | -0.706057 |
| U | BB1 | 32.885000 | 0.083159 | -7.355311 | 2.187153 | -3.559038 | 2.330030 | -0.397385 |
| U | BB2 | 3.806000 | -0.107731 | 9.600999 | -6.261319 | -0.466683 | 1.136981 | -0.213516 |
| U | BB3 | 3.594000 | 0.045443 | 9.596259 | -1.292222 | -7.111432 | 4.056877 | -0.633828 |
| U | SC1 | 5.951000 | -0.029245 | 2.596687 | -0.561187 | -1.564771 | 0.892651 | -0.141308 |
| U | SC2 | 10.901000 | 0.021789 | 0.188390 | 0.722231 | -1.925816 | 0.966543 | -0.145013 |
| U | SC3 | 5.246000 | -0.045865 | 5.899781 | -1.506647 | -3.170544 | 1.937171 | -0.317010 |
| U | TE3 | 7.824000 | -0.029681 | 7.937832 | -0.330781 | -10.141202 | 6.633347 | -1.221112 |
| U | TE5 | 8.036000 | -0.009097 | 4.331935 | 0.434165 | -5.808314 | 3.524388 | -0.623824 |

Table S4 Comparison of the protein/RNA models identified via metainference simulations, using atomistic (Kooshapur et al., 2018) or Martini form factors. The agreement with SAXS data was measured with CRYSOL (Svergun et al., 1995) using the maximum order of harmonics available and 18 points for the Fibonacci grid. The contrast of the solvation shell was fixed to 0.005 ; the values for the radius of atomic group and excluded volume were the default ones computed by CRYSOL. The model quality was assessed using the Molprobity validation implemented in Phenix (Adams et al., 2010; Davis et al., 2007).

|  |  | Refined Models |  |
| :--- | :--- | :--- | :--- |
| Agreement | CRYSOL $\chi^{2}$ | Atomistic | Martini |
| with SAXS | CRYSOL $\chi^{2}\left(q<0.3 \AA^{-1}\right)$ | 2.05 | 2.65 |
| Model Quality | Molprobity Score | 1.46 | 1.84 |
|  | Clash-score | 0.58 | 1.15 |
|  | Ramachandran favoured | $92 \%$ | $92 \%$ |
|  | Ramachandran outliers | $2 \%$ | $2 \%$ |

Figure S1 Calculated Martini form factors for DNA and RNA nucleotides. The grey lines represent the form factors back-calculated from each nucleotide bead in the library, while the coloured lines are the averages over all the individual form factors.

## DNA



## RNA



Figure S2 Distribution of $q_{\text {threshold }}$ values for 121 DNA crystallographic structures, coloured according to DNA classification. Average $q_{\text {threshold }}$ values are $0.56,1.29,0.51$ and $0.78 \AA^{-1}$ for Aform, B-form, Z-form and quadruplex (Q) DNA, respectively.


Figure S3 Distribution of $R$-values for 44 RNA (a) and 121 DNA (c) crystallographic structures, computed over a range of scattering vector $q$ below $0.5 \AA^{-1}$. $R$-values, averaged over the whole set of structures for RNA (b) and DNA (d) and evaluated over a range of scattering vectors below a cut-off ( $q$ _cutoff), are reported as a function of the cut-off. The standard deviation is represented as a shadow.

## RNA





Figure S4 For 121 DNA crystal structures are reported: (a) the distribution of $R$-values computed over a range of scattering vector $q$ below $0.5 \AA^{-1}$, coloured according to DNA classification; (b) $R$ values, averaged over the DNA structures of X-form (with X indicating A-, B-, Z- form or quadruplex) and evaluated over a $q$ range below a cut-off, as a function of the cut-off used.

DNA


Figure S5 Distribution of $R$-values computed over a range of scattering vector $q$ below $0.45 \AA^{-1}$ for RNA (a) and DNA (b).


## References

Adams, P. D., Afonine, P. V, Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. \& Zwart, P. H. (2010). Acta Crystallogr. D. Biol. Crystallogr. 66, 213-221.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. \& Haak, J. R. (1984). J. Chem. Phys. 81, 3684-3690.

Bonomi, M. \& Camilloni, C. (2017). Bioinformatics. 33, 3999-4000.
Bonomi, M., Camilloni, C., Cavalli, A. \& Vendruscolo, M. (2016). Sci. Adv. 2, e1501177.
Bussi, G., Donadio, D. \& Parrinello, M. (2007). J. Chem. Phys. 126, 14101.
Darden, T., York, D. \& Pedersen, L. (1993). J. Chem. Phys. 98, 10089-10092.
Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall 3rd, W. B., Snoeyink, J., Richardson, J. S. \& Richardson, D. C. (2007). Nucleic Acids Res. 35, W375-W383.

Ferrarotti, M. J., Bottaro, S., Pérez-Villa, A. \& Bussi, G. (2015). J. Chem. Theory Comput. 11, 139146.

Franke, D., Jeffries, C. M. \& Svergun, D. I. (2015). Nat. Methods. 12, 419.
Hess, B., Bekker, H., Berendsen, H. J. C. \& Fraaije, J. G. E. M. (1998). J. Comput. Chem. 18, 14631472.

Ivani, I., Dans, P. D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrio, P., Goñi, R., Balaceanu, A., Portella, G., Battistini, F., Gelpí, J. L., González, C., Vendruscolo, M., Laughton, C. A., Harris, S. A., Case, D. A. \& Orozco, M. (2015). Nat. Methods. 13, 55.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. \& Klein, M. L. (1983). J. Chem. Phys. 79, 926-935.

Kimanius, D., Pettersson, I., Schluckebier, G., Lindahl, E. \& Andersson, M. (2015). J. Chem. Theory Comput. 11, 3491-3498.

Kooshapur, H., Choudhury, N. R., Simon, B., Mühlbauer, M., Jussupow, A., Fernandez, N., Jones, A. N., Dallmann, A., Gabel, F., Camilloni, C., Michlewski, G., Caceres, J. F. \& Sattler, M. (2018). Nat. Commun. 9,.

Löhr, T., Jussupow, A. \& Camilloni, C. (2017). J. Chem. Phys. 146, 165102.
Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. \& Simmerling, C. (2015). J. Chem. Theory Comput. 11, 3696-3713.

Rieping, W., Habeck, M. \& Nilges, M. (2005). Science (80-. ). 309, 303 LP-306.
Stovgaard, K., Andreetta, C., Ferkinghoff-Borg, J. \& Hamelryck, T. (2010). BMC Bioinformatics. 11, 429.

Svergun, D., Barberato, C. \& Koch, M. H. (1995). J. Appl. Crystallogr. 28, 768-773

