Quality Assurance for Biomolecular Simulations

Stuart E. Murdock,[a,†] Kaihsu Tai,[c] Muan Hong Ng,[b] Steven Johnston,[b] Bing Wu,[c,d] Hans Fangohr,[b] Charles A. Laughton,[e] Jonathan W. Essex[*a] and Mark S. P. Sansom[c]

25th August 2006

[a]School of Chemistry and [b]School of Engineering Sciences, University of Southampton, United Kingdom; [c]Department of Biochemistry and [d]Oxford e-Science Centre, University of Oxford, United Kingdom; and [e]School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, United Kingdom

[*]Corresponding author e-mail: j.w.essex@soton.ac.uk; phone: +44 23 80592794

[†]Current address: Schrödinger Inc., 101 SW Main Street, Suite 1300, Portland, OR 97204, USA

Contemporary structural biology has an increased emphasis on high-throughput methods. Biomolecular simulations can add value to structural biology via the provision of dynamic information. However, at present there are no agreed measures for the quality of biomolecular simulation data. In this letter, we suggest suitable measures for the quality assurance of molecular dynamics simulations of biomolecules. These measures are designed to be simple, fast, and general. Reporting of these measures in simulation papers should become an expected practice, analogous to the reporting of comparable quality measures in protein crystallography. We wish to solicit views and suggestions from the simulation community on methods to obtain reliability measures from molecular-dynamics trajectories. In a database which provides access to previously obtained simulations – for example BioSimGrid (`http://www.biosimgrid.org/`) – the user needs to be confident that the simulation trajectory is suitable for further investigation. This can be provided by the simulation quality measures which a user would examine prior to more extensive analyses.

1 Overview

For the past quarter century, biomolecular simulations have been adding value to structural biology via the provision of dynamic information.[1] As genomics move from sequencing to structural and dynamical considerations, and high-throughput technologies advance from crystallography to molecular-dynamics (MD) simulation, this process is occurring with vigor. As the bibliometric data in Fig. 1 shows, MD simulation of biopolymers is now becoming a routine technique. To help this maturation process, standardized practice should be established in the simulation community, similar to that in crystallography.[2,3] It is already regular practice to print quality measures in a formulaic table in published articles reporting crystallographic results – indeed, it is surprising if such a table is missing, and the referees would readily reject the manuscript.

We are hereby initiating a discussion on the appropriate measures of quality and convergence[4] for MD simulation trajectories of biopolymers. The process of calculating these measures are designed to be automated for large numbers of trajectories; hence the set of analyses used for this description should be general, with minimal interaction of a human curator. The scientist can then use these measures, along with sensible comparisons with known experimental data (which we recognize as essential), to decide whether a specific trajectory is suitable for further investigation. Our purpose is to solicit feedback from the simulation community with regard to the analyses we have chosen and to obtain further suggestions. We invite the community to express their views on our choices of measures.

We are motivated to do this by our work in building BioSimGrid,[5] a distributed environment for archiving and analysing biopolymer simulations. Other similar databases are emerging[6] (personal communications with Valerie Daggett and Modesto Orozco, `http://mmb.pcb.ub.es/MODEL/`); these also require some quality-assurance measures. The BioSimGrid project was implemented to satisfy the growing demand for the storage of large amounts of simulation data which is currently being produced within a number of laboratories. This environment enables the storage and analysis of large biopolymer MD trajectories, making previously logistically-difficult comparative analyses and data curation easier. For example, analyzing large numbers of trajectories distributed across many laboratories has now become as seamless through BioSimGrid, as if the data were produced in the same laboratory. Traditionally, timeseries of millions of degrees-of-freedom are collected in a biomolecular MD simulation, but only a small fraction of these are presented in the resulting paper, due to the lifetime of the project. In this scenario, other laboratories may be interested in a trajectory of a biomolecule for which MD has been done, but without a clearinghouse, access is difficult to obtain. BioSimGrid

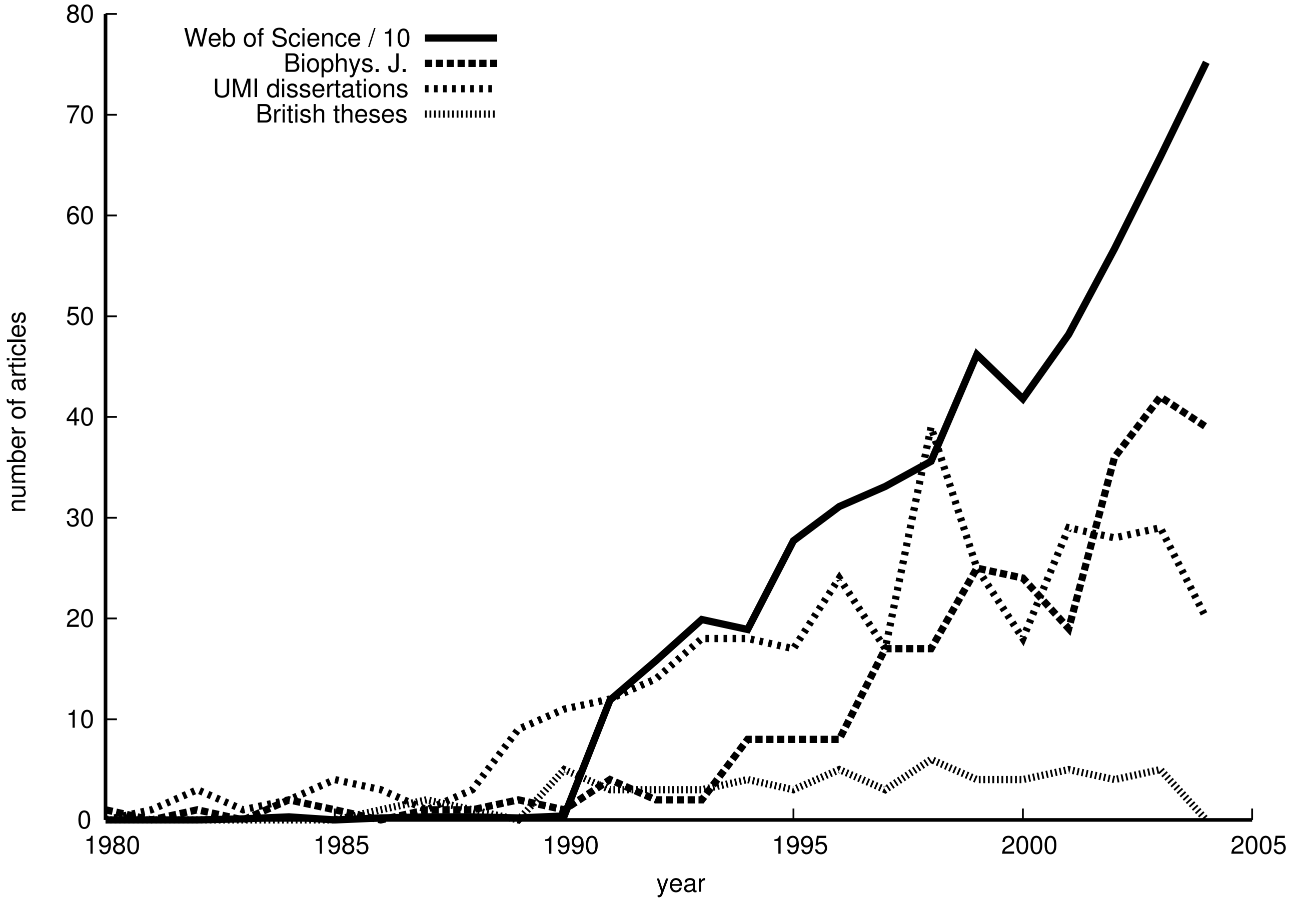

Figure 1: Bibliometrics shows that molecular-dynamics simulation has the potential to become a routine scientific technique for investigating biopolymer dynamics: search results for protein molecular dynamics simulations in Web of Science, *Biophysical Journal*, UMI ProQuest Digital Dissertations, and Index to Theses in Great Britain. Web of Science search for protein molecular dynamics simulations “`TS=((molecular SAME dynamics) AND simulation* AND protein*)`” at `http://wok.mimas.ac.uk/`. Title/abstract search in the Biophysical Journal for keywords `'molecular dynamics' and 'simulation'` at `http://www.biophysj.org/`. Abstract search “`"molecular dynamics" and "protein*"`” in the UMI ProQuest Digital Dissertations (mainly North American) at `http://wwwlib.umi.com/dissertations/`. Search “`"molecular dynamics" and "protein*"`” in the Index to Theses in Great Britain (false positives in the 1980s removed) at `http://www.theses.com/`. Data up-to-date mid-2004.

aims to make the results of large-scale computer simulations of biomolecules more accessible to the biological community. As a comprehensive simulation data-management system with many analysis tools, BioSimGrid provides scientists access to trajectories stored throughout many laboratories, once public availability has been granted by the owner of the data. An early exemplar application is comparisons amongst enzymes of similar active sites.[7]

Some may hold the opinion that, based on the motivation of the scientist carrying out the work, MD simulations may be split into two categories: equilibrium and non-equilibrium. The non-equilibrium simulations are those where the scientist wished to explore unfolding pathways or large conformational changes not manifest in known crystallographic structures. The investigations of biomolecular dynamics in an equilibrium state fall into the other category. It follows that the quality-assurance measures and analyses methods for these different classes have to be quite different. We take a more agnostic approach at the original motivation that brought the trajectory into existence. There are many types of MD simulations being performed under varying degrees of non-equilibrium conditions; any boundaries or distinctions imposed *a priori* may turn out inappropriate. Another drawback of such *a priori* description is that it requires speculation about the scientist's state-of-mind: the decision reached thus may not always be accurate.

The following quality-assurance measures we introduce are not restricted to proteins, but any polymers, and may be readily applied to nucleic acids, sugars (polysaccharides), and even non-biological polymers where the monomers can be clearly identified. To keep our convergence measures general, so they may be automated, we include all atoms in each biopolymer within the trajectory, and we perform the analysis by first performing quaternion least-squares fitting[8] for the atoms with respect to the initial configuration of the stored trajectory. These measures could appropriately be split into three different classes: quality, convergence, and structural;[9,10] they are described in detail in the sections below.

2 Provenance metadata, whose reporting should be obligatory

Two particular sets of metadata that describe the provenance of the trajectory should be reported. The provenance, or ontogeny, metadata tells how and whence the trajectory came about. The first concerns the preprocessing before the initial structure for the MD simulation can be obtained; the second, the particular set-up for the simulation. As is typical for reporting of experimental procedures in scientific literature, these should be reported to the extent that an unrelated laboratory will be able to reproduce the simulation.

The first set, presentable in free-styled text, includes: the Protein Data Bank identification code for the crystallographic structure on which the simulation is based; the procedure for reconstructing residues and sidechains which were not observed in the crystallographic structure; determination of protonation states; ligand insertion; solvation (including the retention of crystallographic water molecules and addition of ions) or insertion into medium (for example, the procedure of solvation in water and that for insertion into a lipid bilayer); the equili-

Global trajectory identifier	BioSimGrid_GB-OXF_9
Trajectory name	outer-membrane phospholipase A
Trajectory type	membrane-bound protein
Method	molecular dynamics
Time-step	2 fs
Sampling frequency	10 ps
Total number of frames	566
Computational platform	commodity Intel-based personal computer
Software package	GROMACS 2.0
Ensemble	NpT (isothermal-isobaric)
Thermostat	Berendsen, 298 K
Thermostat relax time	0.1 ps
Barostat	Berendsen (anisotropic), 100 kPa
Barostat relax time	1.0 ps
Boundary condition	periodic
Unit cell	cuboid
Forcefield	GROMOS87
Solvent	water
Solvent forcefield	SPC
Electrostatics treatment	cutoff, 1.8 nm
Source Protein Data Bank identifier	1QD5
Enzyme Classification	3.1.1.32

Table 1: A formulaic table providing some of the provenance metadata for a molecular-dynamics simulation.[11]

bration protocol (including the pre-equilibration of solvents and ions).

The second set, mostly presentable in a formulaic table (Table 1), includes: MD software package (name and version); computer hardware and operating system used (general timing information if appropriate); forcefields for both solute and solvent (name and version, including any special modifications); boundary condition and shape of unit cell; electrostatics treatment; ensemble; barostat (if used: type and target pressure); thermostat (if used: type and target temperature); constraints; time-step; snapshot sampling frequency; duration of simulated trajectory; special restraints and interactive MD protocol.

Calls for a standard in the output of MD packages are particularly poignant here: all the items enumerated in the second set needs to be reported in the entirety to facilitate comparison; this is not routinely done in the literature. The most convenient and sensible spot to capture these metadata is at the point of generation; that is, at the output of the MD package. However, at the moment, the defaults of each MD package (which may be different) are often not written out explicitly, and have to be speculated downstream. Capturing metadata in detail also helps in avoiding inappropriate set-up parameters from being used – for example, with the help of a validating program that alerts the scientist about inappropriate combinations of electrostatics treatment and solvent forcefield (amongst others).

3 Thermodynamic measures of quality, whose reporting should be obligatory

When MD simulations are performed, various quantities are calculated and written to output files. Some of these are well-understood, and may be used to observe what is happening in a simulation as an indication of quality. For instance, we might expect the temperature in a simulation with a thermostat applied to be fluctuating about a constant value, within a small range, over time; if this is not the case, the quality of the simulation is suspicious. We have decided that these quality-indicating measures which we consider should be: temperature, pressure, potential energy, kinetic energy, number density, volume, cell dimensions, and specific heat capacity. There are some redundancies in reporting; this is to avoid the need for the users to recalculate often-used values, but the values also need to be verified to ensure consistency. “Quality” here means “overall thermodynamic stability”, and does not necessarily guarantee to the simulation’s accuracy in reproducing physical phenomena or the “usefulness” of the trajectory: Even an ill-parameterized forcefield can yield to stable but unphysical trajectories. Further, anomalies may occur due to local unbalanced distributions of kinetic energy; a careful human curator, rather than the reported values here, will have to be relied upon to catch these.

As their behavior is well-understood and they depend on the simulation conditions, these measures may be called quality measures for the trajectory. These should be obtained from the simulation output files, though we note that current MD packages usually do not write these out for every time-step by default; this arguably should be rectified to avoid data loss. These quantities should be plotted as a time-series. The mean and standard deviation should be reported, so automatic filters can be easily applied downstream to select only the trajectories over a quality threshold.

4 Convergence and fluctuation measures, whose reporting should be obligatory

Once the quality of the simulation has been determined by examining the thermodynamic measures, it is instructive to inspect some measures that reflect the convergence and the fluctuations. As not all simulations are meant to be converging, these are not necessarily measures of the quality, but unexpected behaviors can lead to interesting investigations or possible problems. These useful measures are the root mean square deviation and fluctuation, and the radius of gyration.

The root mean square deviation (RMSD) time-series throughout a trajectory can be used to understand by how much the conformation of a biopolymer changes with respect to time. The RMSD provides a measure of conformational stability or drift. For a converging simulation, we would expect the RMSD to increase and then start to plateau. (However, RMSD plateauing does not necessarily indicate convergence.) As we wish the same analysis to be performed automatically, we want a set of basic, specific rules that can be adhered to. First, all atoms in the biopolymer molecule are used in the least-squares fitting procedure to remove the translational and rotational degrees of freedom. The RMSD is calculated with respect to the initial configuration in the trajectory and all atoms are used to calculate the RMSD. In addition, it may be useful to report RMSD for only the backbone or α-carbon atoms of proteins, or to exclude some loops with large fluctuations in RMSD calculations.

The root mean square fluctuation (RMSF), like the RMSD, is calculated for all atoms in each biopolymer. Before this calculation, all atoms included in the trajectory are used to remove the translational and rotational degrees of freedom. The RMSF gives a measure of the fluctuation of atoms around the average position and any large fluctuations should be understood in the light of the crystallographic B factors.[9,12]

The radius of gyration gives a measure of how the mass of a group of atoms is distributed around their center of mass. For converged trajectories, the radius of gyration time-series of a biopolymer should also reach a plateau.

Further convergence measures such as the cosine content[10,14] and overlap measures from block ("windowed") analyses[15,16] can be obtained from principal component analysis. For example, comparison of the first and last parts of a trajectory might help to detect inadequacies in equilibration or convergence that calls for a longer simulation time. If desirable, the obligation to report these will emerge from the simulation community.

5 Structural measures, whose reporting is not obligatory but informative

A plot of eigenvalues from principal component analysis, ranked in decreasing magnitude, should usually shows a “scree” shape; that is, an initial sharp drop in magnitude followed by a leveling of small eigenvalues. This indicates that only a few modes of motions have large length-scales, as the magnitudes of the eigenvalues tell the contribution to the total motion from the corresponding principal component. From this plot the appropriateness of the “essential dynamics” analysis,[13] where only the motion modes with the largest length-scales receive attention, may be determined. As the motion of all the constituent atoms is to be determined the mass-weighted covariance matrix[14] is used for the analysis. A plot where such a “scree” shape can be seen is shown in Fig. 2.

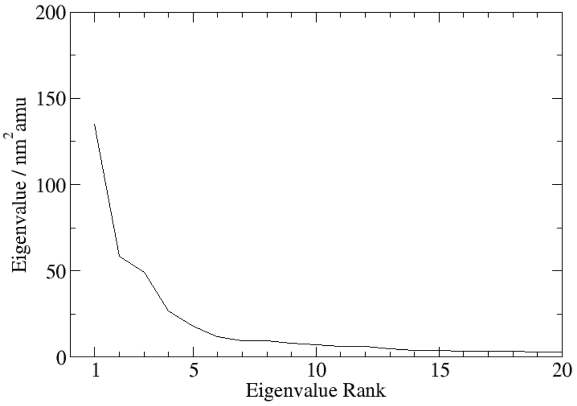

Figure 2: The 20 eigenvalues with the largest magnitude for a 9.6 ns simulation of a prion protein are plotted to obtain a scree plot to determine the most prominent modes of motion.

For each of the few principal components with large eigenvalues (as shown in the scree plot), it is recommended to plot a graph of the projection of the trajectory onto the component as a time-series; here the projection in a particular principal component shows the degree of sampling within that component. This plot reveals the global-conformational “states”, or clusters of the same projections, visited. As the trajectory becomes longer, previously unvisited states can become visited (exploration); also, previously visited states may be visited again (revisiting).

Projections of the trajectory onto a pair of principal components with large eigenvalues provide a good idea of the phase-space sampled. With an energy value plotted on the third axis, this is a way to visualize the energy landscape.[17]

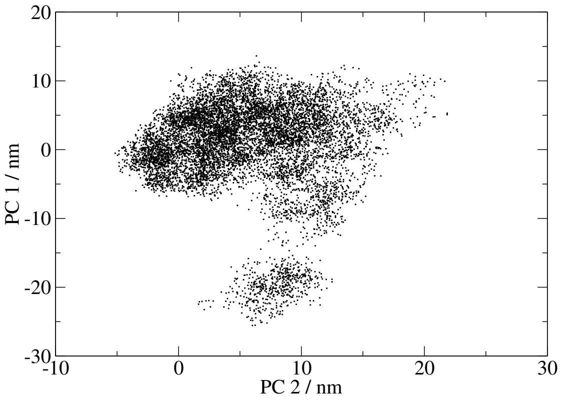

Figure 3: Projection of a trajectory for a 9.6 ns simulation of a prion protein onto the two principal components with the largest eigenvalues showing the sampling in the first (PC 1) and second (PC 2) components.

An example of the comparison of the projections onto the two principal components with the largest eigenvalues is shown in Fig. 3. In this example, two distinct regions in the phase-space have been sampled. It is possible to read from this graph whether there are revisiting events in some areas in the phase-space. This may be done by clustering or block analysis (discussed above); however, this process is highly dependent on the clustering algorithm, so we do not consider it to be obligatory. Again, if these are desirable, the obligation to report these can emerge from the simulation community.

To obtain more specific information on the structure of the biopolymers, it is advisable to perform analyses designed for the type of biopolymer in question. For example, secondary-structure determination over the trajectory can be done for proteins,[18] or the Curves analysis set for nucleic acids.[19] These may provide a more detailed picture on the quality of the simulation. In addition, quality indices used by the crystallography, nuclear magnetic resonance, and bioinformatics communities may be considered to complement the indices here.[20]

6 Conclusion

In this letter, we have suggested a scheme where a general understanding of a biosimulation trajectory may be obtained, with a set of strictly-defined analyses. This set gives the researcher a reasonable overview to decide whether to investigate further. We have tried to design the analyses to be as simple and general as possible, whilst maintaining enough complexity to understand and distinguish the simulations.

We have implemented these measures in the BioSimGrid toolkit. This toolkit has been developed to enable users to perform predefined analyses easily within the analysis environment. When the archiving of a trajectory is complete, these analyses will automatically be performed on the recently deposited data via the standard tools included in the toolkit.

We consider the set of analyses suggested here to be adequate, and hope that the simulation community will adopt it as a basis for development and discussion. Suggestions may be sent to the editors of this journal, or to the authors.

7 Acknowledgements

We thank all our colleagues for discussion, especially Oliver Beckstein, Richard Boardman, Leo Caves, Simon J. Cox, David Moss, Adrian Mulholland, and Paul Jeffreys. We also thank the anonymous referee for helpful comments. This project is funded by the Biotechnology and Biological Sciences Research Council, United Kingdom.

References

[1] Karplus, M.; McCammon, J. A. *Nature Struct. Biol.* **2002,** *9*, 646–652.

[2] Brown, I. D.; McMahon, B. CIF: the computer language of crystallography. Acta Cryst. 2002, B58, 317–324.

[3] Vriend, G. WHAT IF: A molecular modeling and drug design program. J. Mol. Graph. 1990, 8, 52–56.

[4] van Gunsteren, W. F.; Mark, A. E. Validation of molecular dynamics simulation. J. Chem. Phys. 1998, 108, 6109–6116.

[5] Tai, K.; Murdock, S.; Wu, B.; Ng, M. H.; Johnston, S.; Fangohr H.; Cox, S. J.; Jeffreys, P.; Essex, J. W.; Sansom, M. S. P. BioSimGrid: towards a worldwide repository for biomolecular simulations. Org. Biomol. Chem. 2004, 2:3219–3221.

[6] J. Wozniak, P. Brenner, D. Thain, A. Striegel, J. Izaguirre, Generosity and Gluttony in GEMS: Grid Enabled Molecular Simulations. To appear in Proceedings of 14th IEEE International Symposium on High-Performance Distributed Computing 2005.

[7] Tai, K.; Baaden, M.; Murdock, S.; Wu, B.; Ng, M. H.; Johnston, S.; Cox, K.; Essex, J. W.; Sansom, M. S. P. Active-site dynamics of hydrolases: comparison of simulations using the BioSimGrid database. Manuscript in preparation.

[8] Mackay, A. L. Quaternion transformation of molecular orientation. Acta Cryst. 1984, A40, 165–166.

[9] Brooks III, C. L.; Karplus, M.; Pettitt, B. M. Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics, Vol. LXXI of Wiley Series on Advances in Chemical Physics. Wiley-Interscience: New York, 1988.

[10] Hess, B. Similarities between principal components of protein dynamics and random diffusion. Phys. Rev. E, 2000, 62, 8438–8448.

[11] Baaden, M.; Meier, C.; Sansom, M. S. P. A Molecular Dynamics Investigation of Mono and Dimeric States of the Outer Membrane Enzyme OMPLA. J. Mol. Biol. 2003, 331, 177–189.

[12] McCammon, J. A.; Harvey, S. Dynamics of Proteins and Nucleic Acids. Cambridge University Press: Cambridge, UK., 1987.

[13] Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. Proteins: Struct. Funct. Genet. 1993, 17, 412–425.

[14] Hess, B. Convergence of sampling in protein simulations. Phys. Rev. E, 2002, 65, 031910.

[15] Faraldo-Gomez, J. D.; Forrest, L. R.; Baaden, M.; Bond, P. J.; Domene, C.; Patargias, G.; Cuthbertson, J.; Sansom, M. S. P. Conformational sampling and dynamics of membrane proteins from 10-nanosecond computer simulations. Proteins. 2004, 57, 783–791.

[16] Flyvbjerg, H.; Petersen, H. G. Error estimates on averages of correlated data. J. Chem. Phys. 1989, 91, 461–466.

[17] Becker, O. M. Principal coordinate maps of molecular potential energy surfaces. J. Comp. Chem. 1998, 19, 1255-1267.

[18] Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983, 22, 2577–2637.

[19] Stofer, E; Lavery, R. Measuring the geometry of DNA grooves. Biopolymers. 1994, 34, 337–346.

[20] Kleywegt, G. J. Validation of protein crystal structures. Acta Cryst. D. 2000, 56, 249–265.