COMPUTATIONAL CRYSTALLOGRAPHIC NEWSLETTER

-VE IONS, MULTI-COMPONENT SF, RNA2023

Table of Contents

Editor

Nigel W. Moriarty: NWMoriarty@LBL.Gov

Phenix News

Announcements

New Phenix Release Imminent

The latest version of Phenix – 1.21 – has been released. It will be the last release using Python2. All future version will be Python3 starting with 3.7 or 3.9 depending on OS.

Several new modules have been included in the installation including the quantum chemistry code MOPAC allowing QM calculations with the latest methods with further installation.

Downloads, documentation and changes are available at phenix-online.org. Changes include

- Full support for structure determination with AlphaFold models in Phenix GUI
  - PredictAndBuild X-ray and Cryo-EM structure solution from data and sequences
  - Phenix AlphaFold server
  - Video tutorials for prediction, X-ray structure solution and Cryo-EM map interpretation
- Cryo-EM tools support ChimeraX visualization
- Cryo-EM density modification and anisotropic scaling display local resolution
- Tutorials available for automated structure determination with PredictAndBuild
- New em_placement and emplace_local tools
  - likelihood-based docking of models into cryo-EM maps
- MOPAC v22 is now distributed with Phenix
- Quantum Mechanical Restraints (QMR) to calculate ligand restraints *in situ*
  - Available in phenix.refine and separate command-line tool, *mmtbx.quantum_interface*
  - Higher level QM available via 3rd-party Orca package

- Approximately 27k of 37k restraints (QM calculated and validated) deployed in the GeoStd
- Automated tests for programs in the Phenix GUI

Crystallographic meetings and workshops

Crystallographic & Cryo-EM Structure Solution with Phenix workshop at 74th American Crystallography Association meeting July 7, 2024

Members of the Phenix team will be conducting this workshop in coordination with the ACA.

Expert Advice

Fitting Tip #24 – Negative Ions are not just charge opposites of Positive Ions

Jane Richardson and Michael Prisant, Duke University

The contrast

Positive and negative ions are more different than one might expect. Many positive ions can form metals in the solid state and negative ions do not. Here we study them as isolated, charged atoms in the context of solvated macromolecules. There, positive ions make quite short, direct interactions: ionic bonds with negatively charged atoms or polar bonds either with unprotonated partially charged atoms or with lone pairs on waters (electron donors, not H-bond donors and not drawn as lines here). In contrast, negative ions are bound by much longer, indirect interactions: multiple H-bonds from donor H atoms such as NH or water H (shown as pillows of green dots at the resulting vdW-radius overlaps).

Positive Ions

Most of us are more familiar with the properties and binding sites of positive ions because they are much more common in protein and nucleic acid structures than negative ions.

Zinc is the simplest to recognize and most reproducible. It usually has 4 strongly bound and tetrahedrally directed ligands – routinely they are unprotonated His N or Cys S (as in the Zn finger shown in figure 1), sometimes an O, but never a water. Most often Zinc plays a structural role and is fully occupied, so that its high density is obvious.

Calcium, Ca+2 in proteins (e.g. 1w0n, or similar Dy+3 of Fig. 4) almost always has 6 to 8 oxygen ligands, just arranged around it with no specific coordination geometry. It likes the negative charge of sidechain carboxyls and binds one or both O atoms.

Potassium, like many + ions, has essential biological roles and occurs in proteins that store or transport it such as ion channels, where it successively binds at rings of 4 backbone CO groups. In figure 1, a K+ sits symmetrically between two layers of O6 G atoms (red) in an RNA G-quadruplex.

Magnesium has strongly octahedral coordination, although that geometry is not directly visible at lower resolution or at the solvent surface. Most typically it has 1-2 macromolecular ligands and the rest waters, but can bind well even with all waters, as in the figure. Mg+2 is very common in RNA, since it is a major counterion that stabilizes RNA folding.

Iron, in various (+ve) oxidation states and geometries, is essential and common, as are Mn, Co, Ni and Cu to a lesser extent.

Negative Ions

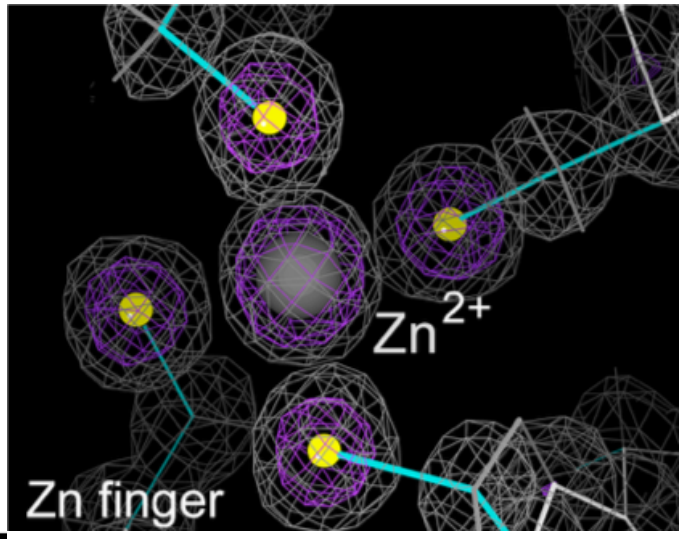
Negative ions seen in macromolecules are the halides: in order of atomic number, Fluoride F- (9), Chloride Cl (17), Bromide Br (35) and Iodide I (53). As noted, they make H-bonds rather than

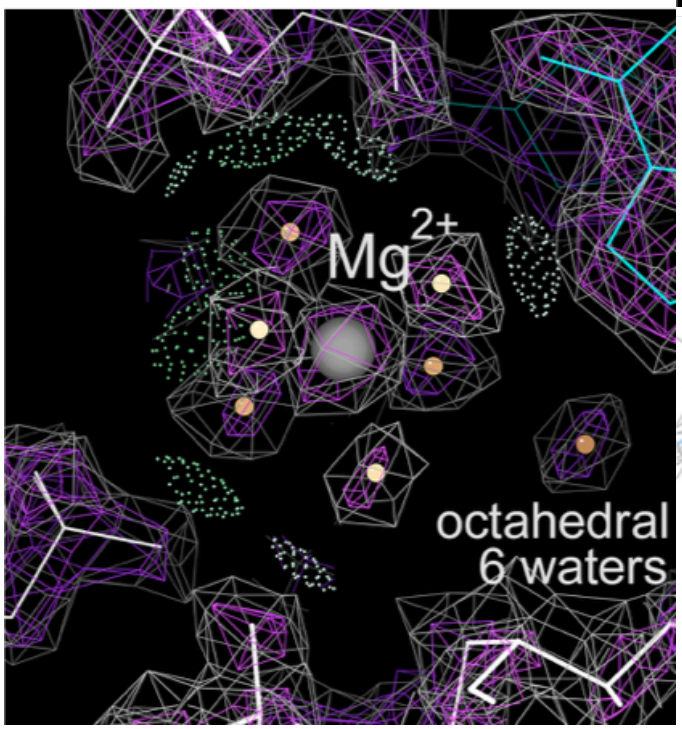
short polar bonds and prefer 6 octahedral H-bond donor ligands. But some of the octahedral directions may instead touch in good van der Walls interactions, the angles may sometimes be distorted and only some ligands will be visible when at the molecular surface or at lower resolution.

Chloride is the most common halide in the PDB. Figure 2 top is a Cl- from a lysozyme crystal soaked in NaCl. It has one backbone NH and 4 water ligands in octahedral geometry. The 6th direction is a good vdW contact (smaller, darker green dots) to the face of the peptide just below it. Fig. 2 center shows the context of that Cl-, including a Na+ and 2 Cl- at the surface each with only one Asn NH2 and one water ligand visible. Fig. 2 bottom shows a clear I- Iodide, which is very nearly indistinguishable from the Cl- in Fig. 2 top. The distance from ion to ligands ranges from 3.1 to 3.5Å in both cases, the map density at the ion is similar (12σ vs 10σ) and each is about twice the density at nearby oxygens. The Iodide may not be at full occupancy, of course.

The bottom line

The bottom line for model fitting is how to recognize ion sites, tell them from waters and





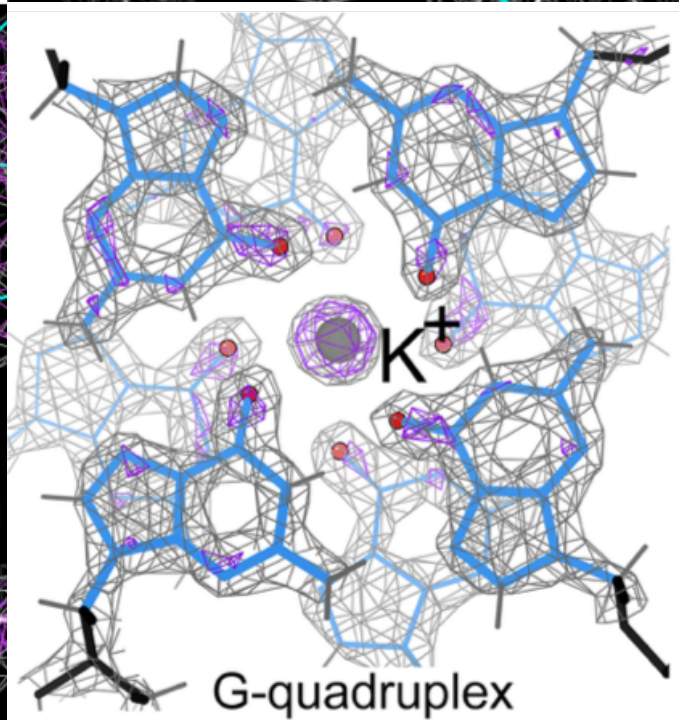


Figure 1: Positive ions, in clear high-resolution examples. (Upper) A Zn+2 ion with 4 tetrahedral Cys S ligands, in 3t7L at 1.1Å (Chaikuad 2011). (Lower left) Mg+2 ion A5103, with 6 octahedral water ligands, in the 8a3d human ribosome at 1.67Å by cryoEM (Faille 2023). (Lower right) A K+ ion in 6e8u at 1.55Å, sandwiched between layers of an RNA-aptamer G-quadruplex (Trachman 2019).

discriminate them from each other. [Also, of course, you will very often see these same atoms as part of other small-molecule ligands; that aspect is not treated here.] The above rules and examples can help you make responsible, probable assignments for individual bound ion atoms. You should always be able to tell positive and negative ions apart, since they have non-overlapping ranges of ion-to-ligand distances. Distances to positive ions vary from 1.9Å for phosphate O to Magnesium up to 2.8 Å for some ligands of Potassium. In contrast, distances to negative ions are H-bonds, with heavy-atom distances that vary from 3.1 to 3.5 Å. Only the transition-metal positive ions have tetrahedral coordination. The lighter Na+, Mg+2. K+ and Ca+2 each have somewhat different ligand and distance preferences, so that set of positive ions can often be distinguished at high to mid resolutions. The negative ions, however, look remarkably like one another and only a full-occupancy iodine is unambiguous just from the map.

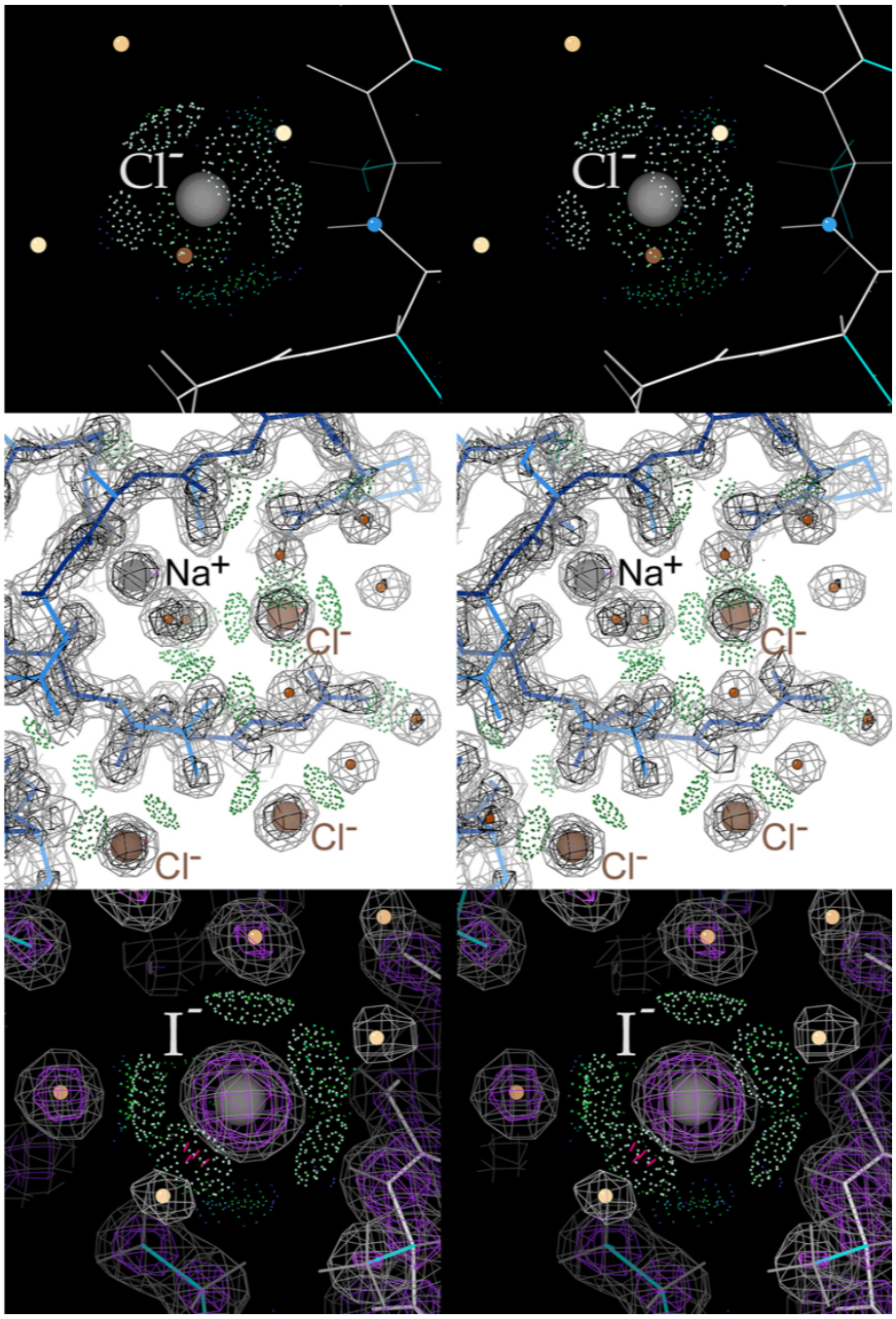


Figure 2: Negative ions in stereo at high resolution. Top: An octahedral Chloride in 7bmt (Koelmel 2012). Center: The context in 7bmt , with a Na+ and two surface Cl-. Bottom: An octahedral Iodide in 2ciw (Kuhnel 2006).

You will need more information than just the density map and coordination at resolutions lower than 2 or 2.5Å, in poorly ordered regions and always if the ion identity and presence really matter for the point of your structure. Even if you know a certain ion is essential and you put it in the crystallization medium, it might still not actually be there in the conformation you crystallized. The paper for the 4enc F- riboswitch shown in figure 3 (Ren 2012) is a great example of using many extra tests to pin down a difficult and important case.

Be sure to check the density value and B factor of the ion peak relative to its surroundings. Look at any density peak significantly higher than the density of good peaks for your macromolecular atoms and at any atoms with an unreasonably low B-factor. See whether the sequence annotation for your molecule shows any common ion binding motifs, such as a Zn finger, an EF hand, or a G-quadruplex. If your molecule is known to have a functional ion site, a distant structural or accidental second site for the same ion is not unusual (as in Fig. 3 of Fitting Tip #22). If you have more than two Cys close together, try a Zn (or perhaps Fe) site as well as disulfides (as in Fig. S4 of Lawson 2021).

Convenient other resources are now also available. The CheckMyMetal web service (Gucwa 2023) at https://cmm.minorlab.org is very useful to assess and distinguish among positive ions and it can now even model and briefly refine a potential replacement ion for you. Be aware that it treats an ion with no direct macromolecular ligands as "unattached" and that it lists but does not handle negative ions (not surprising, since they are not metals), not listing or visualizing any interactions at all for them, neither their H-bonds nor even if they are directly bonded to a positive ion. That failure to show any interactions for negative ions is also true for the NGL Viewer used at the RCSB and PDBe sites. MolProbity's KiNG viewer (Chen 2009) has the opposite problem: it explicitly shows the H-bonds to negative ions but not the ionic bonds to positive ions, as seen in the figures here.

In a large structure it is not feasible to look at all modeled water peaks to check for possible ions or other problems. However, the UnDowser function in MolProbity (Prisant 2020) and in Phenix helps by listing all clashing waters, their clash partners and the B-factor comparison for each clash. The paper shows examples of 14 different types of errors. Of the waters that clash with non-positive polar atoms, many turn out to be positive ions, but UnDowser almost never flags a negative ion since they have near-normal H-bond distances. Another

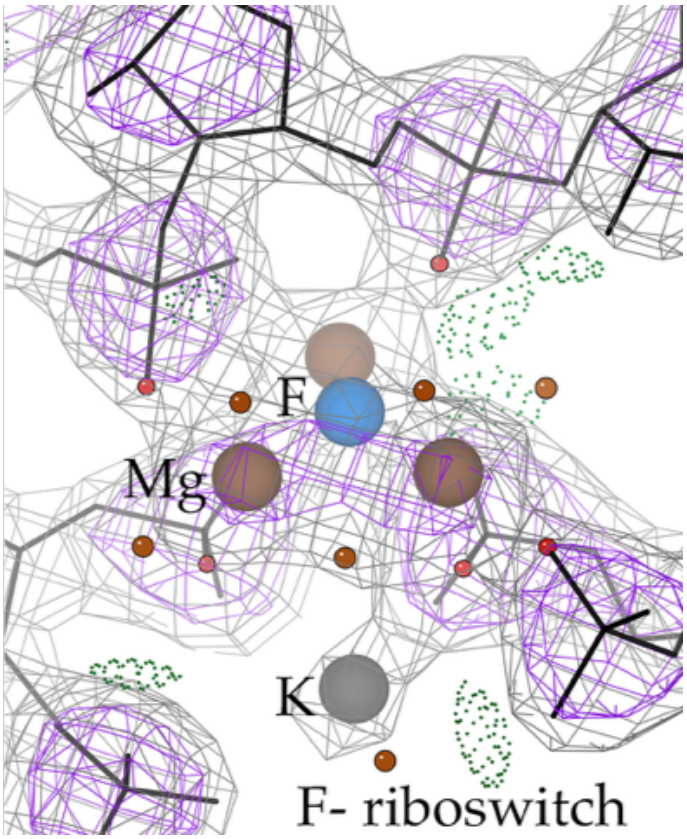


Figure 3: The Fluoride riboswitch, in the bound conformation, from 4enc at 2,27Å (Ren 2012). Contours are at 1.2 and 3s, but peaks for the waters can be seen at lower contour levels.

helpful resource is the ion identification step done in *phenix_refine* after automated water placement (Echols 2014). It is complementary to UnDowser in considering changes of ion identity as well as water-to-ion, but not trying at all to diagnose other water problems. It is less good at the lighter positive ions and is justifiably more conservative, as a fully automated procedure. However, it can fairly often find negative ions since it considers coordination geometry, checks high density values and is greatly aided by the use of anomalous data when that is available. We will hope to combine the strengths of these two methods more thoroughly in the future.

An addendum – Three Ion-related superpowers of macromolecules

I. The Fluoride riboswitch shows that, amazingly, a highly negative RNA molecule can bind a small

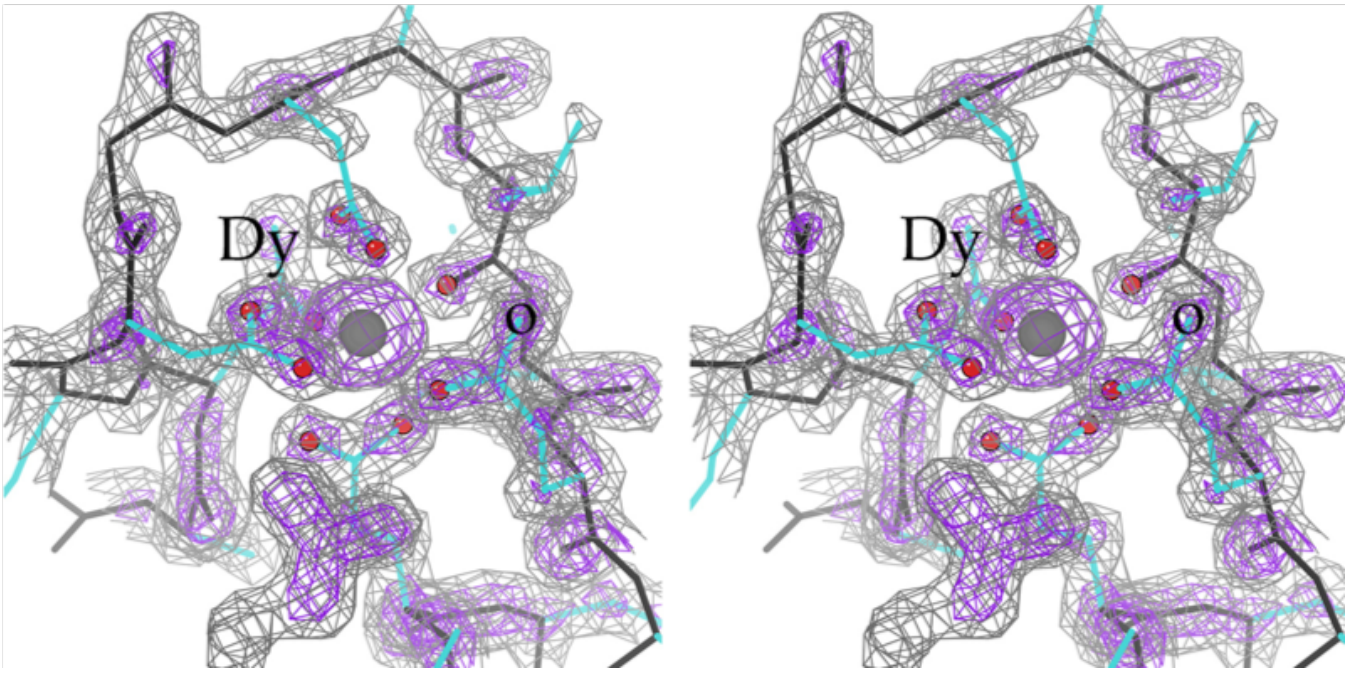


Figure 4: H. quercus lanmodulin, binding Dysprosium Dy+3 with an EF hand helix-loop-helix motif. 8fnr at 1.8Å.

negative ion strongly and selectively (Chloride is known not to bind). It does this by positioning Mg+2 ions next to pairs of phosphates that are distant in sequence (top and bottom RNA strands in Fig. 3) and presumably also distant in space for the unbound conformation of the riboswitch. When F- is present in sufficient concentration, it brings those places together, with three Mg+2 (brown) and a K+ (gray) surrounding the single F- ion (blue), as shown in Figure 3. That change puts the overall riboswitch into the conformation that controls the expression of proteins of Fluoride metabolism.

II. The 15 lanthanides, or "rare-earth", elements differ in useful electronic, magnetic, or catalytic properties, but all are +3 ions and are extremely similar in the physical and chemical properties that allow industrial separation. Current protocols use strong solvents and are very inefficient, requiring on the order of 100 passes over a column. There is a natural family of proteins called lanmodulins that bind lanthanides and it has recently been found that a lanmodulin from the *H. quercus* bacterium in oak tree buds is especially specific (Mattocks 2023). Its crystal structure (Figure 4) showed that the ions were bound at 4 EF-hand motifs with large carboxylate-rich loops. Given waste from rare-earth magnets, it can separate the Neodymium (atomic # 60) from the Dysprosium (66) with 98% purity and 99% specificity in a single column pass. This feat is accomplished because Dy+3 coordinates only 9 oxygens (red balls) while Nd+3 coordinates 10 and allows a dimer to form, further increasing the affinity difference. That extra O is the second branch of the Glu 91 carboxylate, marked with a black o. It is 3.56Å from the Dy ion, while the 9 coordinating O atoms average 2.42Å +/- 0.1 away. But with the larger-radius Nd, that COO could swing to coordinate both oxygens, which would also form a second salt-link H-bond to the Arg in the neighboring molecule (unoccupied guanidinium density below the cluster), helping the dimer form in solution as well as in the crystal.

III. Over the years, JSR has noticed examples of well-occupied single-atom ions next to a protein or RNA but with only water ligands and no direct macromolecular interactions (Fig. 1 bottom). Why would they bind there rather than just staying in the bulk solution to interact with waters there? She now has an answer through MGP's

connections to chemical physics (Berkowitz 2021; Johnson 2003). It seems that waters form clusters around negative ions in solution, but the ion is pushed to one edge of the water cluster. The reason is that the ion wants octahedral coordination and the waters want tetrahedral coordination. A similar mismatch should also happen for positive ions with octahedral coordination. The superpower of macromolecules is to make individual water molecules happy in positions which form most of an octahedral binding site that an ion likes better than staying in solution. These all-water sites are not at all rare for RNA Mg sites –- there are three in the first 30 listed Mg in the 8a3d ribosome and two clear examples in the small 6eru aptamer.

It has long been known that proteins are tool users –- clear for the cofactors, including waters, that they co-opt to help in enzyme catalysis. The riboswitch is an effective but somewhat heavy-handed use of charge, while the lanmodulin specificity and the all-water ion sites are quite subtle uses of geometrical detail.

References

Berkowitz M (2021) Molecular simulations of aqueous electrolytes: Role of explicit inclusion of charge transfer into force fields, *J Phys Chem B* **125**: 13069-76

Chaikuad A, Williams E, Guo K, Sanvitale C ... Bullock A, SGC (2011) Crystal structure of the FYVE domain of endofin (ZFYVE16) at 1.1Å resolution, to be published [3t7L Zn2+]

Chen VB, Davis IW, Richardson DC (2009) KiNG (Kinemage, Next Generation): A versatile interactive molecular and scientific visualization program, *Protein Sci* **18**: 2403-2409

Faille A, Dent KC, Pellegrino S, Jaako P, Warren AJ (2023) The chemical landscape of the human ribosome at 1.67Å resolution, *bioRxiv*, doi: https://doi.org/10.1101/2023.02.28.530191 [8a3d Mg2+]

Gucwa M, Lenkiewicz J, Zheng H, Cymborowski M, Cooper DR, Murzyn K, Minor W (2023) CMM – An enhanced platform for interactive validation of metal binding sites, *Protein Sci* **32**: e4525

Koelmel W, Kuper J, Kisker C (2021) Cesium based phasing of macromolecules: a general ease to use approach for solving the phase problem, *Sci Rep* **11**: 17038-17038 [7bmt Na+, 3 Cl-]

Lawson CL, Kryshhtafoviych A, Adams Pd, Afonine PV, Baker ML ... Chiu W (2021) CryoEM model validation recommendations based on outcomes of the 2019 EMDataResource challenge, *Nat Meth* **18**: 156-164

Mattocks JA, Jung JJ, Lin CY, Dong Z … Cotruvo JA Jr (2023) Enhanced rare-earth separation with a metal-sensitive lanmodulin dimer, *Nature* **618**: 87-93 [8fnr Dy3+]

Prisant MG, Williams CJ, Chen VB, Richardson JS, Richardson DC (2020) New tools in MolProbity validation: CaBLAM for cryoEM backbone, UnDowser to rethink "waters" and NGL Viewer to recapture online 3D graphics, *Protein Sci* **29**: 315-329

Ren A, Rajashankar KR, Patel DJ (2012) Fluoride encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch, *Nature* **486**: 85-89 [4enc F-]

Robertson WH, Johnson M (2003) Molecular aspects of halide ion hydration: The cluster approach, *Annu Rev Phys Chem* **54**: 172-213

Trachman RJ III, Aufour A, Jeng SCY, Abdolahzadeh A ...Ferre-D-Amare AR (2019) Structure and functional reselection of the Mango-III fluorogenic RNA aptamer, *Nat Chem Biol* **16**: 472-479 [6e8u K+]

FAQ

Can I use MOPAC with the current Phenix installer?

Yes, but if you downloaded the Python2 version (the default) the environment variable $PHENIX_MOPAC needs to be set to point to the user installed copy of MOPAC. Better to install the Python3 version (at the bottom of the download page) to have a recent version of MOPAC.

Multi-Component Structure Factor Modeling in Presence of Twinning

Alexandre G. Urzhumtsev[1,2], Paul D. Adams[3,4], Pavel V. Afonine[3,#]

[1]*Centre for Integrative Biology, Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS–INSERM-UdS, 1 rue Laurent Fries, BP 10142, 67404 Illkirch, France*

[2]*Université de Lorraine, Faculté des Sciences et Technologies, BP 239, 54506 Vandoeuvre-les-Nancy, France*

[3]*Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*

[4]*Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA*

#PAfonine@lbl.gov

Introduction

In macromolecular crystals the region surrounding macromolecules is occupied by disordered solvent which contributes to the structure factors. The flat-mask bulk solvent model developed by Jiang & Brunger (1994), followed by several improvements, *e.g.*, Fokine & Urzhumtsev (2002) and Afonine *et al.* (2013), considers this region filled uniformly with the same type of solvent. The total model structure factors are then defined as a scaled sum of atomic model contribution, $\mathrm{F}_{calc}(\mathrm{s})$ and the bulk-solvent:

$$\mathrm{F}_{model}(\mathrm{s}) = k_{total}(\mathrm{s})[\mathrm{F}_{calc}(\mathrm{s}) + k_{mask}(\mathrm{s})\mathrm{F}_{mask}(\mathrm{s})] \qquad (1)$$

Here $\mathrm{F}_{mask}(\mathrm{s})$ are the Fourier coefficients calculated from the flat solvent mask and $k_{mask}(\mathrm{s})$ are corresponding resolution-dependent scale factors. The uniform character of the bulk-solvent has been challenged in the past. Indeed, bulk-solvent region may have isolated sub-regions inside macromolecules or at their interfaces that can be empty, partially occupied, or occupied by disordered chemical entities that are chemically different than the bulk-solvent itself (Liu *et al.*, 2008; Matthews & Liu, 2009; Lunin *et al.*, 2001; Sonntag *et al.*, 2011). To account for the eventually non-uniform features of the bulk-solvent, we have proposed an approach that allows multi-part solvent treatment (mosaic solvent model; manuscript in preparation) in a computationally efficient manner (Afonine *et al.*, 2023). In this approach the bulk-solvent contribution is considered as a scaled sum of contributions $\mathrm{F}_n(\mathrm{s})$ arising from N different solvent components

$$\mathrm{F}_{model}(\mathrm{s}) = k_{total}(\mathrm{s})\left[\mathrm{F}_{calc}(\mathrm{s}) + \sum_{n=1}^{N} k_n(\mathrm{s})\mathrm{F}_{mask}(\mathrm{s})\right] \qquad (2)$$

Here we describe the extension of the algorithm described in Afonine *et al.* (2023) to account for the case of twinned crystals. This new procedure essentially assembles several parts from previously published works with some specific adjustments needed to account for twinning.

Method

To simplify expressions in what follows, we introduce $\mathrm{F}_0(\mathrm{s}) = \mathrm{F}_{calc}(\mathrm{s})$ with $k_0 = 1$ and rewrite (2) as

$$\mathrm{F}_{model}(\mathrm{s}) = k_{total}(\mathrm{s}) \sum_{n=1}^{N} k_n(\mathrm{s}) \mathrm{F}_n(\mathrm{s}) \tag{3}$$

where the coefficients $k_{total}(\mathrm{s})$ and $k_n(\mathrm{s})$ are the variables to determine. In presence of twinning, intensities of the model structure factors are calculated as

$$I_{model}(\mathrm{s}) = \sum_{\mu=1}^{M} \alpha_\mu \left| \mathrm{F}_{model}(T_\mu \mathrm{s}) \right|^2 \tag{4}$$

Here T_μ is the twin operator represented by a 3x3 matrix generating a reflection of the same resolution, $|T_\mu \mathrm{s}| = |\mathrm{s}|$, and the coefficients $0 < \alpha_\mu \leq 1$ describe twinning fractions such that

$$\sum_{\mu=1}^{M} \alpha_\mu = 1 \tag{5}$$

If $M > 1$, we order them by magnitude $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_N$. In the absence of twinning, $M = 1$ and $\alpha_1 = 1$. Below, we consider respective coefficients α_μ as known (see, for example, §2.4 in Afonine *et al.*, 2013).

We search for the coefficients $k_{total}(\mathrm{s})$ and $k_n(\mathrm{s})$ giving the best fit of the model to experimental structure factor intensities minimizing the least-squares function

$$LS = \frac{1}{4} \sum_{\mu=1}^{M} [I_{model}(\mathrm{s}) - I_{obs}(\mathrm{s})]^2 \tag{6}$$

Coefficients $k_{total}(\mathrm{s})$ and $k_n(\mathrm{s})$ are correlated and their values are found iteratively. The initial values of $k_{total}(\mathrm{s})$ and $k_n(\mathrm{s})$ are defined using (1) as described in Afonine *et al.* (2013) with all $k_n(\mathrm{s}) = k_{mask}(\mathrm{s})$, *i.e.*, the considering then bulk solvent mask as a whole. For further iterations, when (approximate) values of coefficients $k_n, n = 0,\ldots,N$ are known, the common scale factor $k_{total}(\mathrm{s})$ is recalculated exactly as in Afonine *et al.* (2013) and we do not repeat this description here. The only difference is that $I_{model}(\mathrm{s})$ is now calculated as a sum (4) of contribution from multiple components, $M > 1$ while Afonine *et al.* (2013) supposes .

Knowing an approximate $k_{total}(\mathrm{s})$ value, individual values $k_n(\mathrm{s})$ for different components are calculated using one of four algorithms described in Afonine *et al.* (2023), with 2nd and 4th algorithms (Alg2 and Alg4, correspondingly) being preferable. However, since Alg4 operates with structure factor amplitudes, the only applicable algorithm in the case of twinning is Alg2 that we generalize below for the case of twinning.

First, we introduce real-valued coefficients

$$G_{nm}(\mathrm{s}) = G_{mn}(\mathrm{s}) = \frac{1}{2}[\tilde{\mathrm{F}}_n(\mathrm{s})\tilde{\mathrm{F}}^*_m(\mathrm{s}) + \tilde{\mathrm{F}}_m(\mathrm{s})\tilde{\mathrm{F}}^*_n(\mathrm{s})] = \tilde{F}_n(\mathrm{s})\tilde{F}_m(\mathrm{s})\,.\cos[\phi_n(\mathrm{s}) - \phi_m(\mathrm{s})] \tag{7}$$

and their twinned combinations

$$\tilde{G}_{nm}(\mathrm{s}) = \sum_{\mu=1}^{M} \alpha_\mu G_{nm}(T_\mu \mathrm{s}) \tag{8}$$

In presence of twinning, function (6) can be reduced to a polynomial of the fourth order

$$LS = \frac{1}{4}\left[\sum_{n=0}^{N}\sum_{m=0}^{N}\sum_{j=0}^{N}\sum_{l=0}^{N} k_n k_m k_j k_l \left(\sum_{\mathrm{s}} \tilde{G}_{jl}(\mathrm{s})\tilde{G}_{nm}(\mathrm{s})\right)\right] - \frac{1}{2}\left[\sum_{n=0}^{N}\sum_{m=0}^{N} k_n k_m \left(\sum_{\mathrm{s}} \tilde{G}_{nm}(\mathrm{s}) I_{obs}(\mathrm{s})\right)\right] + \frac{1}{4}\sum_{\mathrm{s}} [I_{obs}(\mathrm{s})]^2 \tag{9}$$

(for derivation, see formula (7) in Afonine *et al.*, 2023). This function can be minimized using, for example, L-BFGS (Liu & Nocedal, 1989). This procedure requires partial derivatives of (9)

generalizing expression (9) from Afonine *et al.* (2023) and naturally coinciding with it in the absence of

$$\frac{\delta LS}{\delta k_j} = \frac{1}{4}\sum_{m=0}^{N}\sum_{m=0}^{N}\sum_{l=0}^{N} k_l k_m k_n \left(\sum_{\mathrm{s}} \tilde{G}_{jl}(\mathrm{s})\tilde{G}_{nm}(\mathrm{s})\right) - \sum_{n=0}^{N} k_n \left(\sum_{\mathrm{s}} \tilde{G}_{jn}(\mathrm{s}) I_{obs}(\mathrm{s})\right) \tag{10}$$

twinning.

Acknowledgment

PVA and PDA thank the NIH (grants R01GM071939, P01GM063210 and R24GM141254) and the PHENIX Industrial Consortium for support of the PHENIX project. This work was supported in part by the US Department of Energy under Contract No. DE-AC02-05CH11231. AU acknowledge Instruct-ERIC and the French Infrastructure for Integrated Structural Biology FRISBI [ANR-10-INBS-05].

References

Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D. & Urzhumtsev, A. (2013). *Acta Cryst.* D**69**, 625–634.

Afonine, P. V., Adams, P. D. & Urzhumtsev, A. (2021).

https://www.biorxiv.org/content/10.1101/2021.12.09.471976v1

Afonine, P. V., Adams, P. D. & Urzhumtsev, A. (2023). *Acta Cryst.* A**79**, 345-352.

Fokine, A. & Urzhumtsev, A. (2002). *Acta Cryst*. D**58**, 1387–1392.

Jiang, J. S. & Brünger, A. T. (1994). *J. Mol. Biol*. **243**, 100–115.

Liu, D. C. & Nocedal, J. (1989). *Math. Program*. **45**, 503–528.

Liu, L., Quillin, M.L. & Matthews, B.W. (2008). *Proc. Natl. Acad. Sci.*

Lunin, V.Yu., Lunina, N.L., Ritter, S., Frey, I., Keul, J., Diederichs, K., Podjarny, A., Urzhumtsev, A.G. & Baumstark, M. (2001). *Acta Cryst.* D**57**, 108-121.

Matthews, B.W. & Liu, L. (2009). *Protein Sci*. **18**, 494–502.

Sonntag, Y., Musgaard, M., Olesen, C., Schiøtt, B., Møller, J.V., Nissen, P. & Thøgersen L. (2011). *Nat Commun.* **2**, 304.

Urzhumtsev, A. & Podjarny, A. D. (1995). Jnt CCP4/ESF–EACMB Newsl. Protein Crystallogr. **31**, 12–16.

The RNA2023 Residue-Filtered RNA Dataset. Negative Ions Are Not Just Charge Opposites of Positive Ions; Plus a Digression on 3 Ion-Related Superpowers of Macromolecules

Christopher J Williams and Jane S Richardson

Duke University, Durham, NC, 27710

Correspondence email: christopher.sci.williams@gmail.com

Introduction

Any learning, human or machine, requires high-quality input data. The Richardson lab uses high-quality, filtered datasets of protein or RNA residues to develop structure validations. These datasets have allowed us to set validation targets for Ramachandran and CaBLAM distribution and to define clusters of sidechain rotamers and RNA backbone conformations.

Over the years, we have created and shared several lists of high-quality protein and RNA chains for use as training sets. However, we previously had to leave the task of residue-level filtering to the individual users of these datasets due to file-size limits. Thanks to high-volume data distribution through Zenodo, the release of the Top2018 protein residue dataset (Williams, 2022) represented the Richardson lab making our residue-level filtering broadly available for proteins. Here we present the RNA2023 dataset, an equivalent, residue-filtered dataset for RNA.

Chain selection

Our selection of candidate chains was based on the 3.150 version of the http://rna.bgsu.edu/rna3dhub/nrlist list (Leontis and Zirbel, 2012). This list groups RNA structures into classes based on sequence and structure, then selects the best representative of each class based on resolution, RSR, RSCC, Rfree, clashes, and completeness. This selection process is philosophically similar to our selection process for protein chains in previous work.

We applied an additional resolution cutoff of 1.9Å or better to this list. We made a single exception to this cutoff – 6ugg, a 1.95Å tRNA structure – as we felt including a high-quality, uncomplexed tRNA representative was important.

We added two additional cryoEM ribosome structures: 8a3d, a human ribosome at 1.67Å and 8b0x, an *E.coli* ribosome at 1.55Å. 8a3d contributed three RNA chains to the candidate list: 28S rRNA, 5S rRNA, and 5.8S rRNA. 8b0x contributed five RNA chains: 16S rRNA, 23S rRNA, 5S rRNA, an mRNA, and a tRNA. These structures were solved via cryoEM, and represent the first cryo structures to be included in one of our high-quality datasets. They contribute not only a large number of residues but also, more importantly, a great diversity of local conformations.

Finally, after the residue filtering process, we removed any chains that did not have any surviving suites, i.e. the sugar-to-sugar backbone region of two sequential residues. In the protein datasets, we set a completeness cutoff of 60% after filtering. RNA chains are frequently short enough and suites are long enough that a percentage-based completeness cutoff is overly punishing. Possession of at least one complete suite – the basic unit of RNA backbone geometry (Murray, 2003) – was therefore used instead as our criterion for meaningful contribution to the dataset.

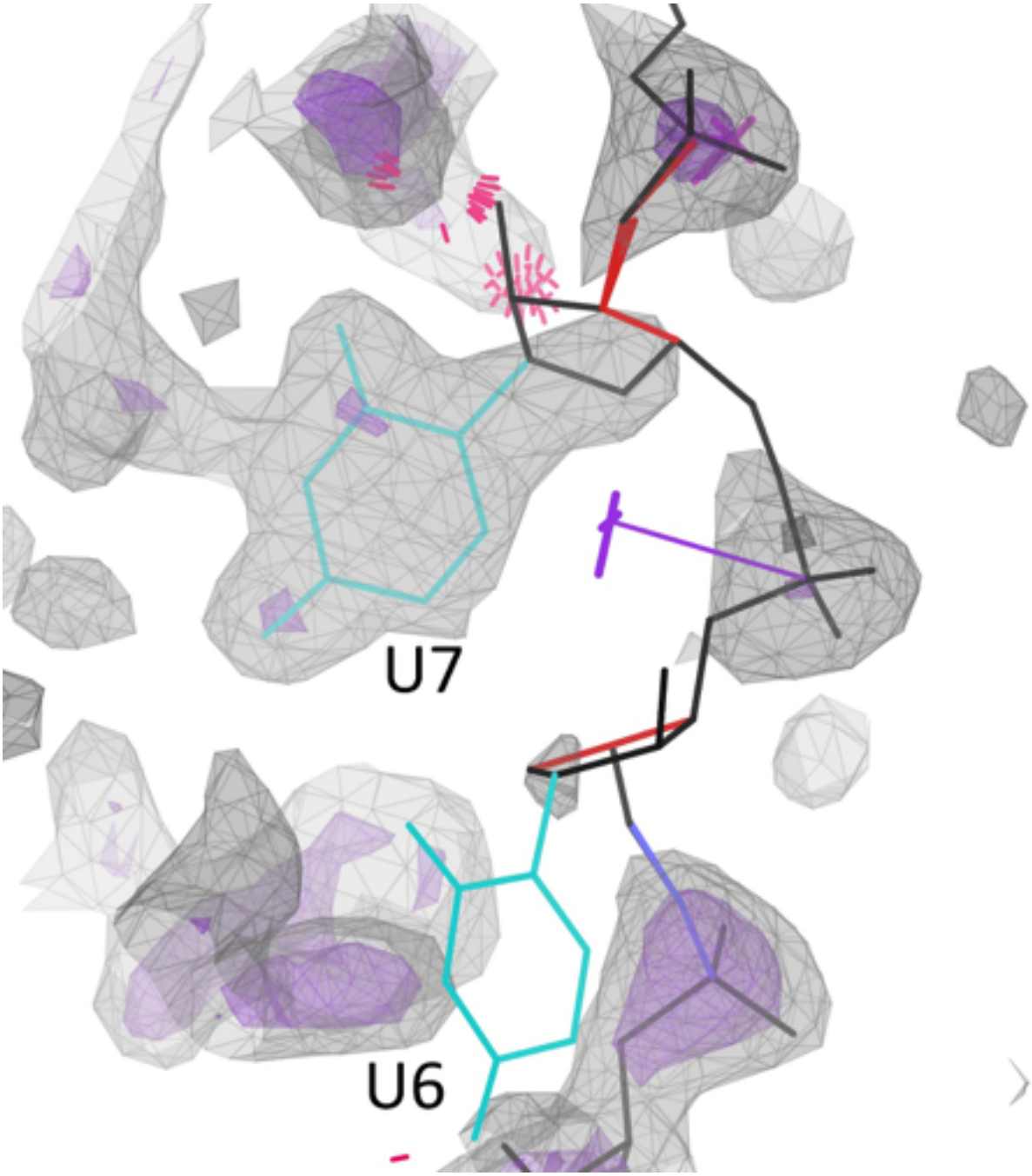


Figure 1: 3boy, chain D, residues 6-7, with electron density map at 1.2σ (gray) and 3.0σ (purple). 3boy is a 1.7Å x-ray structure of generally good quality, and its chain D, a 22-residue mRNA, is one of the candidates for this dataset. This region, however, is poorly resolved and as a result, poorly modeled. Residue 7 fails every one of our residue filtering criteria, except for occupancy. These residues should not be accepted simply because their parent structure is good overall, and their serious problems illustrate the importance of residue-level filtering.

Residue criteria

As in the protein case, model quality and confidence vary across an RNA structure. Structures of good overall quality may have regions of low quality or low confidence (Figure

1) that should not be included in a training or reference dataset.

We require that all RNA residues meet the following validation criteria:

- No steric overlaps (clashes) >= 0.5Å
- No sugar pucker outlier (Jain, 2015)
- No covalent geometry (bond or angle) outliers (>=4σ)

All RNA residues from x-ray crystallography structures must meet the following map criteria:

- 2Fo-Fc map value for P atom >= 2.4σ
- 2Fo-Fc map value, averaged for two lowest atoms, >= 1.2σ
- RSCC value, averaged for two lowest atoms, >= 0.7
- Occupancy = 1.0

All RNA residues from EM structures must meet the following map criteria:

- Residue inclusion fraction within depositor-recommended contour level >= 0.95 for 8a3d, or = 1.0 for 8b0x
- Whole-residue RSCC >= 0.7
- Occupancy = 1.0

Residues that failed any of these criteria were removed from the structure files.

We also prepared an alternative "nosuiteout" dataset that additionally removes all "!!" suite conformation outliers (Richardson, 2008).

Steric overlaps were identified using Reduce and Probe. Sugar puckers, RNA covalent geometry, and RNA suites (where applicable), were validated using `phenix.rna_validate`. We used `phenix.real_space_correlation detail=atom` to calculate 2Fo-Fc map values, RSCC, and occupancies for x-ray structures. We used `phenix.map_model_cc` to calculate RSCC scores for EM models. Residues' inclusion fractions were taken from the validation.xml files provided by the PDB.

Dataset contents

132 unique structures contributed chains to the dataset. The standard dataset contains 151 chains and 6217 complete suites. 4293 suites are the dominant 1a conformation; 1924 suites are non-1a. The nosuiteout dataset contains 149 chains and 5567 complete suites. 4127 suites are 1a; 1427 suites are non-1a, non-!!.

New residue-filtering challenges

Residue-level filtering for RNA presents new challenges relative to our previous protein work.

B-factor was frequently used in our previous datasets as a primary filtering criterion due to its ready availability in all PDB files. However, B-factor is handled differently among different refinements and carries inconsistent meaning across different resolutions. The inconsistency in B-factor became critical in the preparation of this dataset, where we did not find a B-factor cutoff that was sufficiently selective but not unreasonably punishing across the candidate chains. We therefore chose to drop B-factor as a filtering criterion and to depend instead on direct map-model metrics such as RSCC. We expect this change in filtering philosophy will persist into our future protein datasets.

This dataset is the first time we have included EM structures. Model validation criteria (clashes, sugar puckers, etc.) have consistent expectations across methods. However, cryoEM density maps differ from x-ray electron density maps in physical interaction and mathematical protocols and require new criteria, especially because their numerical density values are inherently relative even for the zero point.

Here we use the atom-fraction "residue_inclusion" value taken from the validation.xml files available on the rscbPDB. This measure is calculated for each residue and

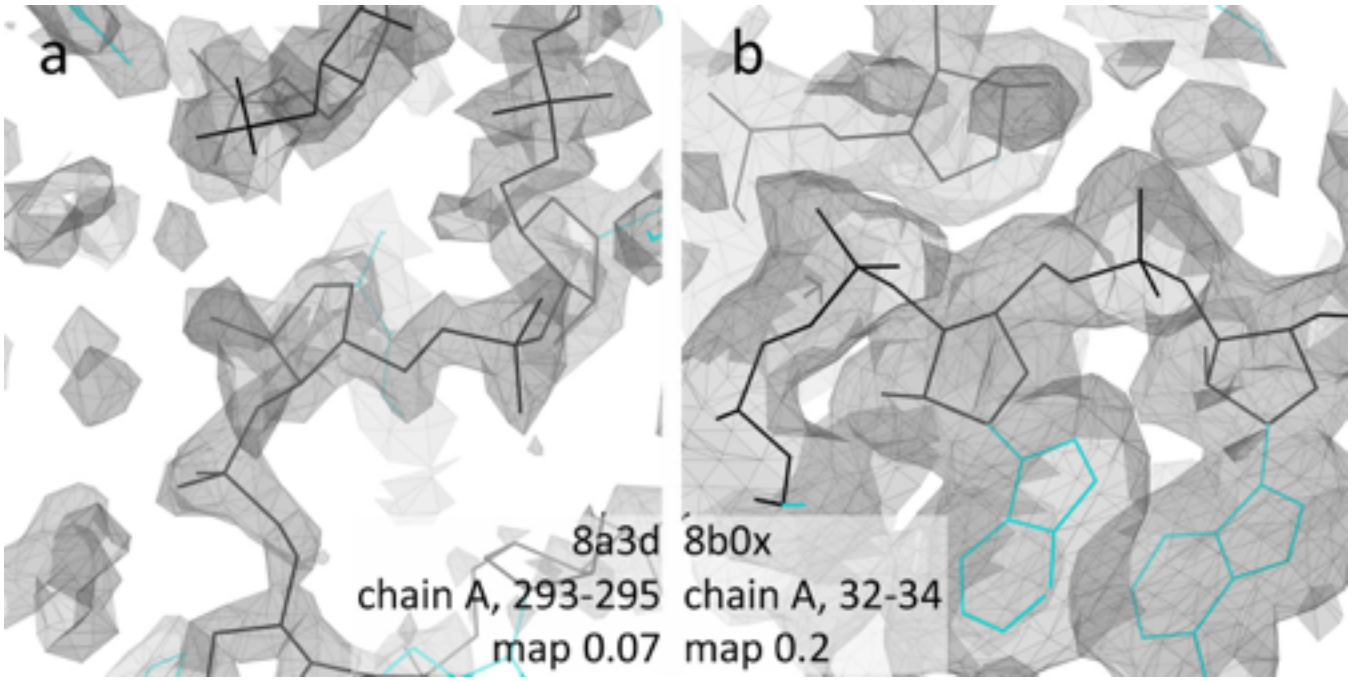


Figure 2: Good local regions with 1.0 residue inclusion fraction, from the human 8a3d and *E.coli* cryoEM ribosomes structures, shown with a single map contour at each structure's depositor-recommended contour level. A) 8a3d, centered on the ribose of 28S G294, with the highly restrictive contour at map density value 0.07. Inspection of example regions convinced us that a 0.95 residue inclusion score would keep only highly reliable residues. b) 8b0x centered on 16S A33, at the quite lenient contour at map density value 0.2. Here, we cannot be more selective than an inclusion fraction of 1.0.

indicates the fraction of that residue's non-hydrogen atoms that fall within a depositor-defined map level (contour_level_primary_map, in the same file). This is roughly equivalent to our x-ray criterion which asks whether the residue falls within the 1.2σ density envelope. (A few residues do not have residue inclusion values. All of these had some other modeling problem that would have eliminated them from the dataset. Any residue for which a residue inclusion fraction cannot be calculated can safely be assumed to fail reasonable filtering criteria.)

Using the depositor's recommendation to define the critical map level presents challenges. There are many legitimate factors a depositor may consider in setting this cutoff, which do not all correspond to our interest in determining local map quality. Figure 2 shows the differences in map character between the two ribosome structures at the depositor-recommended contour levels. Based on visual inspection of example regions near full inclusion at that contour, we chose 0.95 as the required residue inclusion fraction for 8a3d, and 1.0 as the required fraction for 8b0x.

Our use of the residue inclusion fraction represents an early and accessible measure of local map quality. Like B-factor, it requires no special knowledge or software to access, being available from PDB downloads. And like B-factor, it is handled inconsistently across different research groups and different structures. We expect our map-based filtering criteria to evolve in future datasets as we develop or discover better methods.

Dataset access

The rna2023 dataset is available on Zenodo at https://doi.org/10.5281/zenodo.8103013 This link will resolve to the latest version if updates are made.

There are two versions of filtering available. The default dataset includes residues with "!!" suite evaluations. These are conformational outliers relative to known suites. However, as the number of solved RNA structures increases, we are discovering new valid suite conformations. For most purposes, it is therefore appropriate to allow "outlier" RNA backbone conformations, if fit to map and other validation metrics support the model. For specialist purposes where it is desirable to consider only known suite conformations, we supply an alternative "nosuiteout" version of filtering where residues with "!!" conformations have been removed. Each of these filtering versions is available in PDB and mmCIF formats.

The Zenodo repository also includes a file with PDB metadata such as resolution, R values, and deposition title, a chain list file documenting the completeness statistics for member chains, and suitename tables with precomputed suitename identities for all member residues.

References

Jain, S., Richardson, D. C., & Richardson, J. S. (2015). Computational methods for RNA structure validation and improvement. In *Methods in Enzymology* (Vol. 558, pp. 181-212). Academic Press.

Leontis, N. B., & Zirbel, C. L. (2012). Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking. In RNA 3D Structure Analysis and Prediction N. Leontis & E. Westhof (Eds.), (Vol. 27, pp. 281–298). Springer Berlin Heidelberg. doi:10.1007/978-3-642-25740-7_13

Murray, L. J., Arendall III, W. B., Richardson, D. C., & Richardson, J. S. (2003). RNA backbone is rotameric. *Proceedings of the National Academy of Sciences*, *100*(24), 13904-13909.

Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., ... & Berman, H. M. (2008). RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*, *14*(3), 465-481.

Williams, C. J., Richardson, D. C., & Richardson, J. S. (2022). The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues. *Protein Science*, *31*(1), 290-300.

Histidine Protonation Dependent Library (HPDL) for Updating Restraints of the Imidazole Moiety

Nigel W. Moriarty

Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley California 94720, United States

Correspondence email: NWMoriarty@LBL.Gov

Abstract

Histidine can be protonated with a hydrogen atom on either or both of the two nitrogen atoms of the imidazole moiety. The protonation state leads to a change to the geometry of the histidine side-chain: specifically and largely the angles in the ring. Updating the restraints based on histidine protonation leads to an improvement of agreement of the refined geometry and restraints at high resolution.

Introduction

Protonation of histidine can take on three different forms – one hydrogen atom on either or both of the nitrogen atoms in the heterogeneous ring of the imidazole moiety. As (Malinska *et al.*, 2015) noted:

> This variability is of particular importance in protein structures, although it is also particularly recalcitrant to X-ray crystallographic characterization because of the limited resolution and the inability to detect H atoms that blight the method in routine applications.

This assessment led to the development of a set of ideal values for each protonation state of the histidine side-chain based on a search of the Cambridge Structural Database (Groom *et al.*, 2016). While not specifically stated or implemented, these ideal values can be used as ideal values for bond and angle restraints for specific protonation states of histidine used in a model refinement.

Incidentally, Malinska *et al.* also developed a method for predicting the protonation by relaxing the histidine side-chain restraints and using the (approximately unrestrained) refined bond and angle values in two functionals. Unfortunately, it does not work well at "routine" resolutions. In the 1Å example given, only just over 60% of the protonations could be determined.

For context, a recent investigation into the restraints for the arginine amino acid (Moriarty *et al.*, 2020) revealed that changing the ideal values and estimated standard deviation (e.s.d.) of each restraint can have a notable effect on the refinement results particularly if focused on the specific amino acid. In particular, the change in the overall angle root mean squared deviation (r.m.s.d.) values were negligible but the arginine specific r.m.s.d. improved by 0.25° at better than 2Å resolution with a much smaller improvement at lower resolutions. While muted these improvements prompted a deeper and successful investigation into the merits of the new arginine restraints revealing that the torsion restraint needed adjustment. Interestingly, this nuance was missed in the earlier arginine restraints paper (Malinska *et al.*, 2016). Another notable result is that bond restraints (and their changes) have much less influence.

Methods

In a similar fashion to the implementation of the Conformation Dependent Library (CDL, Karplus, 1996; Berkholz *et al.*, 2010, 2009; Moriarty, Adams et al., 2014; Tronrud *et al.*, 2010; Moriarty, Tronrud *et al.*, 2014; Moriarty et al., 2016), the Histidine Protonation Dependent Library (HPDL) adjusts the bond and angle restraints directly in the restraint objects in memory. This is computationally efficient and because the user just has to choose the option, there are no additional files required. The option is hpdl=True.

Interestingly, even though Malinska et al. developed new ideal values for different protonation states of histidine, the update angle restraints for the hydrogen atoms were absent from the paper. Even though the refinement program used in the study – REFMAC (Murshudov *et al.*, 2011) – routinely does not write hydrogen atoms in the final result model, it can use hydrogen atoms internally using the restraints in the Monomer Library (Vagin *et al.*, 2004). If this was the case, the hydrogen atoms would have been refined incorrectly and there would have been scant evidence of the error. The simplest solution for the current work was to bisect the external angle of each nitrogen atom such that the hydrogen atom is restrained to the plane of the ring. As a guide, the internal angle of a protonated nitrogen atom is larger than the un-protonated nitrogen atom.

To test the HPDL restraints, models from the PDB with two different sets of restraints. The first set ('standard') uses the restraints for histidine from the Monomer Library, which is the standard restraints library for the refinement of macromolecules in Phenix (Liebschner *et al.*, 2019). The second set of refinements used the HPDL restraints.

All refinements were performed using phenix.refine (Afonine *et al.*, 2012). Coordinate and experimental data files were obtained from the PDB that met the following criteria: resolution better than 3.05Å, data completeness >90%, data are not twinned, Rwork<30%, Rfree<35% and Rfree-Rwork>1.5%. For entries with resolutions of better than 1.05Å, the Rfree-Rwork criterion was changed to >0.5%. By using these criteria, we excluded suspicious entries and low-resolution data, allowing automatic refinement strategies with default options. Hydrogen atoms were added to the models using Phenix ReadySet!. Ligand restraints were generated by Phenix eLBOW (Moriarty *et al.*, 2009). Each model was then subjected to ten macrocyles of refinement using the default strategy in phenix.refine for the refinement of coordinates, atomic displacement parameters (ADP) and occupancies. Nondefault refinement options included optimization of the weight between the experimental data and the geometry restraints. In addition, anisotropic ADPs were used for non-H protein atoms at resolutions better than 1.55Å and for water oxygen atoms at resolutions better than 1.25Å. The quality of the resulting models was assessed numerically using MolProbity (Williams *et al.*, 2018) in Phenix. To filter out problematic structures, refined models with a clashscore of greater than 12 were not included in the analysis. The results were grouped into resolution bins of width 0.1Å. Resolution bins with less than 10 refined structures were not taken into account. This led to a total of 40,694 protein structures refined with conventional and modified arginine restraints.

Results

The quintessential comparison of the two sets of restraints is shown in Figure 1. The overall r.m.s.d. values for the overall protein models are identical except for the 0.8Å bin which differs by 0.04° in 25 models. Figure 2 has a zoomed in depiction of Figure 1. For the histidine specific comparison, the 0.8Å bin is improved by 0.11°. The r.m.s.d. values differ by about 0.05° for the two next two lower resolution bins. Neither of the comparisons are significantly different. The bond r.m.s.d. Value differences are even less significant.

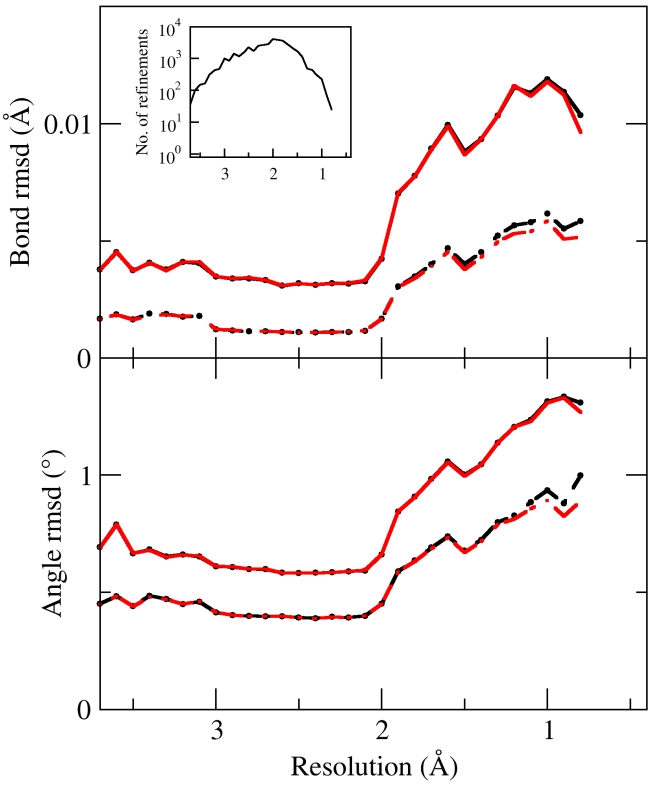


Figure 1: Comparison of bond and angle r.m.s.d. values for whole protein (solid lines) and histidine (dashed lines) for two sets of refinements using the standard restraints (black lines) and HPDL restraints (red lines) in 0.1Å bins. Insert shows the number of refinements in each resolution bin.

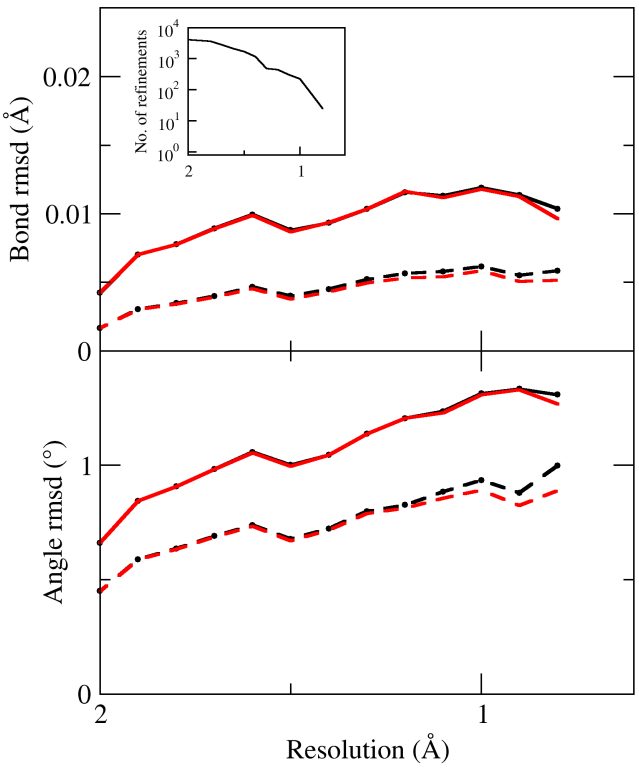


Figure 2: Same as Fig1 but zoomed into the high-resolution range.

Conclusions

The histidine protonation dependent library (HPDL) improves the r.m.s.d. values by an insignificant amount.

References

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). Acta Crystallogr. Sect. D-Biol. Crystallogr. 68, 352–367.

Berkholz, D. S., Krenesky, P. B., Davidson, J. R. & Karplus, P. A. (2010). Nucleic Acids Research 38, D320–D325.

Berkholz, D. S., Shapovalov, M. V., Dunbrack, Jr., R. L. & Karplus, P. A. (2009). Structure 17, 1316–1325.

Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). Acta Cryst B, Acta Cryst Sect B, Acta Crystallogr B, Acta Crystallogr Sect B, Acta Crystallogr B Struct Crystallogr Cryst Chem, Acta Crystallogr Sect B Struct Crystallogr Cryst Chem 72, 171–179.

Karplus, P. A. (1996). Protein Science 5, 1406–1420.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). Acta Cryst D 75, 861–877.

Malinska, M., Dauter, M. & Dauter, Z. (2016). Protein Sci 25, 1753–1756.

Malinska, M., Dauter, M., Kowiel, M., Jaskolski, M. & Dauter, Z. (2015). Acta Crystallogr. D Biol. Crystallogr. 71, 1444–1454.

Moriarty, N. W., Adams, P. D. & Karplus, P. A. (2014). Computational Crystallography Newsletter 5, 42–49.

Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. (2009). Acta Crystallogr. Sect. D-Biol. Crystallogr. 65, 1074–1080.

Moriarty, N. W., Liebschner, D., Tronrud, D. E. & Adams, P. D. (2020). Acta Cryst D 76, 1159–1166.

Moriarty, N. W., Tronrud, D. E., Adams, P. D. & Karplus, P. A. (2014). FEBS Journal 281, 4061–4071.

Moriarty, N. W., Tronrud, D. E., Adams, P. D. & Karplus, P. A. (2016). Acta Crystallographica Section D-Biological Crystallography 72, 176–179.

Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). Acta Crystallogr. Sect. D-Biol. Crystallogr. 67, 355–367.

Tronrud, D. E., Berkholz, D. S. & Karplus, P. A. (2010). Acta Crystallographica Section D-Biological Crystallography