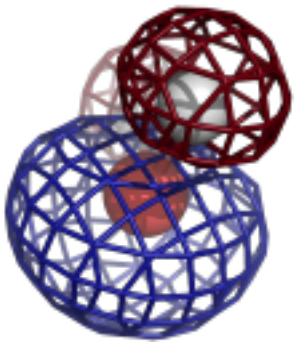
VOLUME SIX



COMPUTATIONAL CRYSTALLOGRAPHY NEWSLETTER

JULY MMXV

PHASER GUI, ALT. LOCS, BASE-PAIR STACKING

Table of Contents

Editor

Nigel W. Moriarty, NWMoriarty@LBL.Gov

Editor's Note

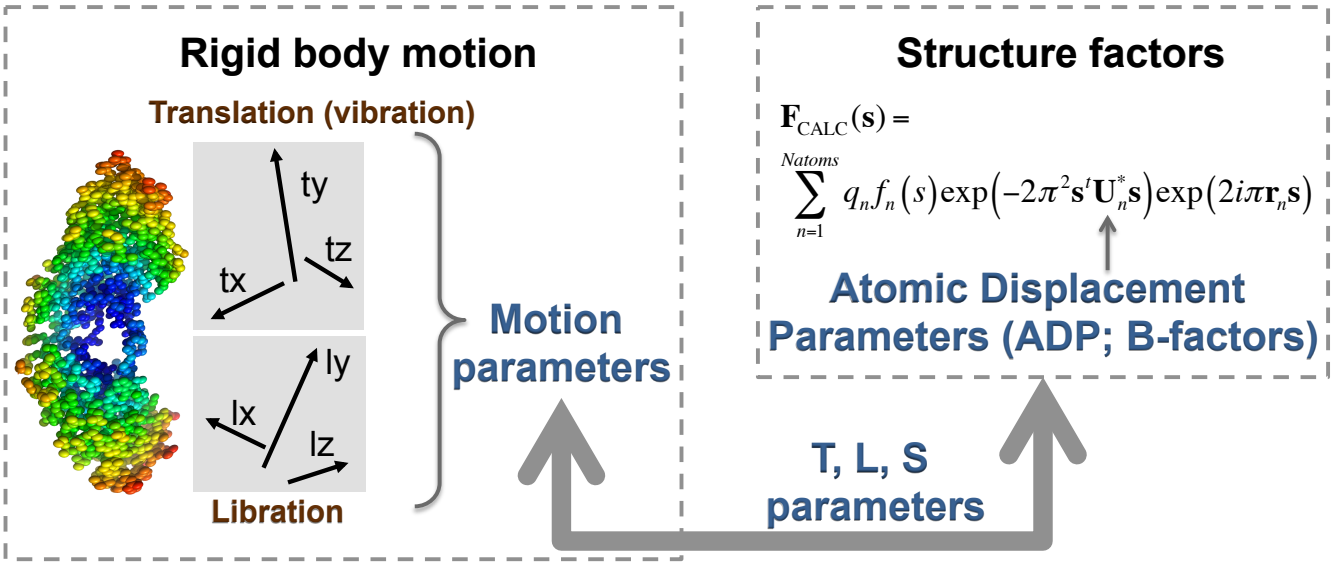
Crystallography makes use of identifiers to access a database of information. The most ubiquitous is the four-digit code used to identify an entry in the Protein Databank known as the PDB id. The first digit is numeric but the remaining three digits are alphanumeric. Alphanumeric digits are also used for all digits of the chemical component identifiers use to identify the protein residues, RNA/DNA bases, ligands and more.

The use of alphanumeric digits can lead to confusion when reading these identification codes stemming from the similarity between the numeral 0 (zero) and the letter O (15th letter of the alphabet) in most typefaces used in scientific literature. The same applies to the numeral 1 (one), the letter I (ninth letter of the alphabet) and the letter l (lowercase of the twelfth letter). This is particularly true in small fonts and sans serif typefaces.

This confusion has led this publication to implement a policy to make these codes more human readable. In all cases, the letters mentioned are to be represented in the appropriate case (o,i,L). This can be used in an otherwise uppercase or lowercase mode with the knowledge that there will be no confusion between similar digits. Examples of codes from the PDB are shown in table A.

Table A: Examples of human readable PDB codes compared with standard representations.

Standard		Human readable	
Uppercase	Lowercase	Uppercase	Lowercase
1OI0	1oi0	1oi0	1oi0
1IJJ	1ijj	1iJJ	1ijj
4OCL	4ocl	4oCL	4ocL



Rigid body motion is described by three vibration axes and three vibration amplitudes along these axes, three libration (oscillation) axes and three libration amplitudes about these axes. TLS model can be used to describe rigid-body motion in terms of individual Atomic Displacement Parameters (ADP; B-factors) thus making possible to include this information into structure factor calculation.

Phenix News

Announcement

All programs that rely on the restraints library default to using the Conformation-Dependent Library proposed by Karplus and others (Tronrud et al. 2010; Moriarty et al. 2014). This includes all types of refinements and, to be consistent, validation automatically detects the appropriate restraints library to use. The CDL is the default in Phenix version 1.10.

References

Moriarty, N.W., D.E. Tronrud, P.D. Adams, and P.A. Karplus. 2014. FEBS J. 281 (18): 4061–71.

Tronrud, D.E., D.S. Berkholz, and P.A. Karplus. 2010. Acta Cryst. D 66 (7): 834–42.

Parameter organisation

Starting with Phenix 1.10 several parameter scopes were changed. This should not affect new projects. However, when restoring jobs from previous versions, some parameters may not be restored properly. In this case one needs to select them again in the GUI. Command-line users should check the updated documentation and correct their scripts accordingly. Affected functionality includes secondary structure restraints, NCS restraints, reference model restraints and C-beta restraints.

New programs

TLS analysis and validation

`phenix.tls_analysis` is a new program for validation of refined TLS parameters based on the recent article by Urzhumtsev *at al.* (2015) entitled "From deep TLS validation to ensembles of atomic models built from elemental motions."

In a nutshell, TLS parameters are a way to pack descriptors of rigid-body motion of a group of atoms into a form suitable to calculate structure factors (see figure above). While refined elements of ***T***, ***L*** and ***S*** matrices contain information about concerted motion of corresponding atoms, the rigid-body motion descriptors themselves, such as the vibration and libration axes and amplitudes, are not readily available from TLS matrices. This information needs to be extracted from TLS matrices in order to be interpreted.

We developed a mathematical procedure that interprets TLS matrices in terms of elemental motions that these matrices fundamentally describe. Decomposing TLS matrices into underlying motions is only possible if these matrices satisfy a set of certain criteria that we formulated. To illustrate the impact of this work, we applied these criteria to all PDB structures that have TLS matrices available. To our surprise, the TLS parameters in about 80% of these structures cannot be interpreted in terms of atomic motions and therefore do not make physical sense within the TLS paradigm.

The result of this work is tri-fold. First, we developed a tool that provides a simple interpretation of TLS matrices in terms of molecular motions. Second, this tool performs a comprehensive validation of TLS matrices. Third, we suggest a reformulation of TLS refinement methods and the corresponding programs to avoid these problems in the future as well as to correct existing problems.

References

Urzhumtsev, A., P. V. Afonine, A. H. Van Benschoten, J. S. Fraser and P. D. Adams. 2015. Acta Cryst. D 71, 1668-1683.

More TLS

`phenix.tls_as_xyz` is another software outcome of the Urzhumtsev *et al.* article (2015) from the previous note. This new program provides an explicit interpretation of refined TLS parameters by decomposing them into elemental rigid-body motions and generating structural ensembles that are consistent with these motions.

Map comparisons

`phenix.map_comparison` is a program that implements some of the ideas described in Urzhumtsev *et al.* (2014). Specifically, for the two input maps it calculates Peak Correlation (CC_{peak}) and Discrepancy Function (D-function) for varying map contouring thresholds. These tools offer at least complimentary and at most better map comparison instruments compared to traditional map correlation coefficient. Also, the program reports cumulative distribution function for both maps that allows choosing meaningful map contour thresholds for comparison of two maps as described in section 2.9 and figure 16 of Afonine *et al.* (2015), the feature enhanced maps (FEM) paper.

References

Afonine, P.V., N.W. Moriarty, M. Mustyakimov, O.V. Sobolev, T.C. Terwilliger, D. Turk, A. Urzhumtsev and P.D. Adams. 2015. Acta Cryst. D 71, 646-666.

Urzhumtsev, A., P.V. Afonine, V.Y. Lunin, T.C. Terwilliger and P.D. Adams. 2014. Acta Cryst. D 70, 2593-2606.

Crystallographic meetings and workshops

The 29th European Crystallographic Meeting, August 23-28, 2015

Location: Rovinj, Croatia. Representatives from Phenix and collaborators will be presenting.

CCP4 Study Weekend, 9-10 January 2016.

Location: East Midlands Conference Centre, Nottingham. The topic of "Protein-Ligand Complexes: Understanding Biological Chemistry".

Expert advice

Fitting Tip #10 – How do your base pairs touch and twist?

Jane Richardson, *Duke University*

Nucleic-acid base pairs, either Watson-Crick or non-canonical, are hydrogen bonded and stack with one another or with other aromatics. The textbook view shows them as coplanar, but that is only approximately true. As a crystallographer, you can fit RNA and DNA more easily if you know what to expect from some of their further subtleties that are summarized and illustrated here and in the accompanying longer article (Richardson, 2015). Each of the various helical forms has its own typical pattern of base relationships,

moderately but significantly distinctive. Even stronger variation occurs for bases or base pairs in the junction, bulge, loop, and interaction regions critical to forming complex tertiary structures and to catalytic function or binding specificity.

The strongest restraint, unsurprisingly, is on planarity of the individual bases. Even modified bases (e.g., methylated, Y base, etc.) are still mostly aromatic and planar, except for a few saturated, puckered rings such as the dihydro-U that is the hallmark of the D loop in tRNAs. Next strongest is base stacking: whenever possible, bases or base pairs that are near to parallel will in fact be quite parallel and at quite ideal vdW distance across their contact, as seen for a regular A-RNA stem in figure 1A.

Next in line is the effect of the base-pair H-bonds, which generally keep quite good H-to-acceptor distances (wide "pillows" of overlap, as seen in figure 1B) but have enough flexibility in angle to accommodate the average propeller twists of -11° to -15° (left-handed) seen in A-RNA, B-DNA, and A-DNA double helices. Figure 1B looks along the base planes of the strand on the right, so the relative twist of the left-strand bases is easily visible even without the stereo. This twist is a result of the constraints of the handed backbone conformation that produces this favorable, repeating arrangement. Other helices produce different average twists: Z-DNA has near-zero propeller twist and parallel poly-A RNA duplex has about +11° (right-handed) twist; both are illustrated in the accompanying article. As seen in figure 2, the 4-fold helix of a G-quadruplex has fairly flat layers on average, with no consistent direction of the low twist.

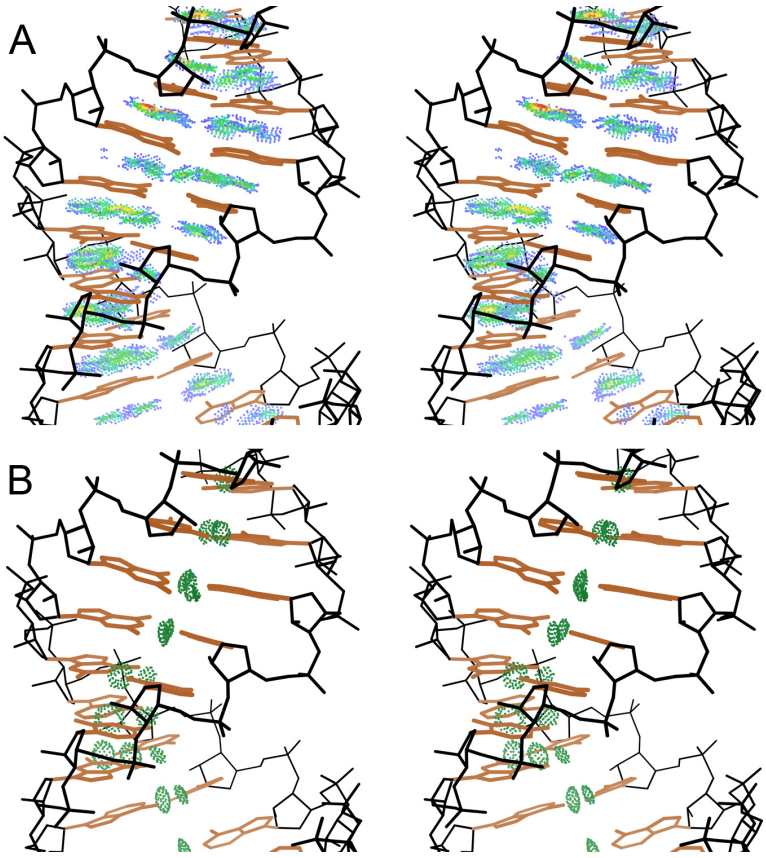


Figure 1: Base-pair stack and twist of regular A-form RNA in a lysozyme aptamer. (A) All-atom contact dots show the large, flat areas of good van der Waals stacking between base pairs, which also provides good π–π interaction. (B) Dot pillows show the base-pair H-bonds, well formed despite an average propeller twist of -14°. From 4M4o at 2.0Å resolution (Malashkevich 2013).

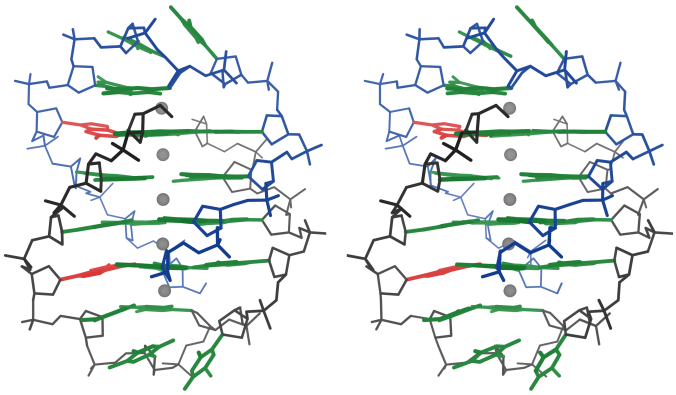


Figure 2: The 2HBN telomeric DNA G-quadruplex at 1.55Å (Gill 2006), with four layers of four H-bonded G bases and ions (Thallium here) between the layers. The two central layers are fairly flat, but in the two outer layers, the base (red) that leads into the capping loop twists about -15° relative to one neighbor base and buckles about 15° relative to the other neighbor.

In contrast to the significant but moderate twist found in repeating helical nucleic-acid forms, at the end or outside of such helices base pairs or adjacent bases can depart much more from parallel. For example, in the G-quadruplex of figure 2, the base (in red) on the strand that leaves one of the outer layers to start the loop is strongly angled. It still doubly H-bonds with both neighbor bases, but twists about -15° relative to one neighbor and buckles (bends across the line of the H-bonds) about 15° relative to the other. Along the top and bottom loops, successive bases do not stack well, but instead angle nearly 30° and touch only at one edge.

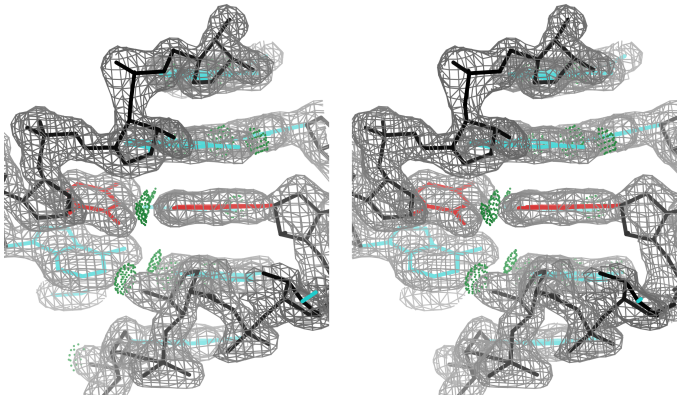


Figure 3: A 25° propeller-twist G-C base-pair (red) with its 1.6Å electron density, from the 437D viral RNA pseudo-knot (Su 1999). The stacks change direction at this base pair, next to where the pseudo-knot switches strand pairing.

Propeller twist is primarily determined by the relative stacking direction of the two paired bases, but it is also restricted by the base-pair H-bonds -- more so if there are more H-bonds to satisfy. G-C pairs, with 3 H-bonds, can twist up to about 25°, as corroborated by the clear electron density seen in figure 3 for the G-C base pair shown in red. Those two bases form their stacks in competing directions where the pseudo-knot switches strand pairing, and their resulting twist is an integral 3D feature of this structure. Base pairs with 2 H-bonds can twist up to about 40° in either direction (see examples in the accompanying article), while those with only one H-bond are not restricted at all by that factor. However *i* to *i+1* base pairs, which can form only one H-bond, happen to be closely coplanar in their most frequent contexts of either a dinucleotide platform or an S-motif.

As mentioned above, these larger departures from parallel stacking or from coplanar base pairs are needed to form the non-repetitive, distinctly recognizable local conformations that enable complex tertiary structure, binding specificity for other molecules, and RNA catalysis. It is important to fit them correctly, because most of them are functionally important. Figure 4 A and B show the location of clusters of such bases

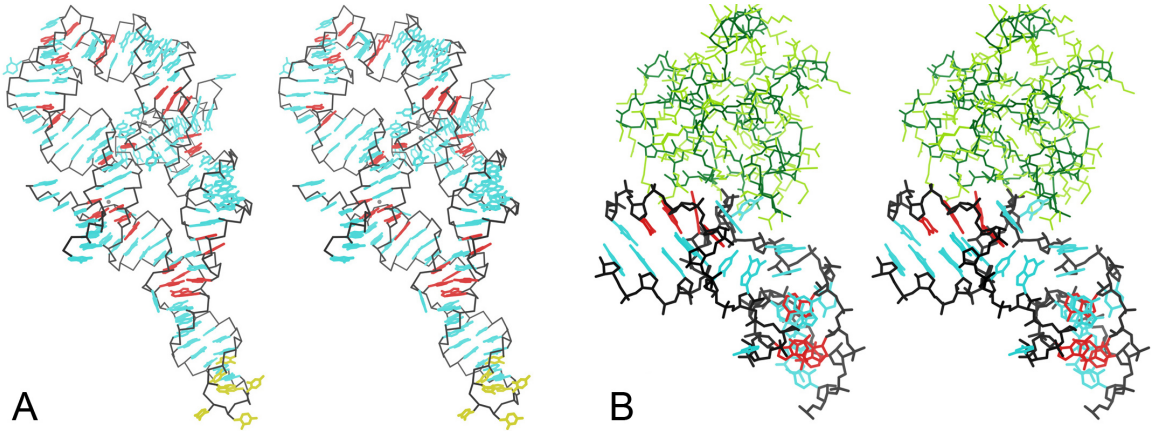


Figure 4: Location of strongly non-parallel bases (red) with high twist or buckle, or with angled, one-spot stacking. They occur at catalytic sites, tertiary contacts, helix junctions, and sites of specific binding, in: (A) the 2R8S self-cleaving P4P6 intron at 1.95Å (Ye 2008) and (B) the 1SDS kink-turn/L7Ae complex at 1.8Å (Hamma 2004). Protein is in green and poorly ordered bases in yellow.

(red) in a ribozyme and in an RNA/protein complex.

Conclusion

At high resolution, the electron density in the well-ordered regions will show unambiguously how to fit the bases, as true for the examples shown here.

At mid resolution, the expectations described here can aid in fitting, and refinement will behave better if base-pair H-bonds are restrained. In Phenix, base-pair H-bonds in RNA or DNA can be identified, classified, and outlier-filtered automatically if they are already in position (Headd, 2012), using a process similar to that for helix and sheet H-bonds in protein; however, it is also a good idea to check manually for marginal but suggestive interactions. For RNA, Phenix can diagnose the correct ribose pucker, but that pucker will be better achieved and kept if you apply the same "P perpendicular" rule (Jain, 2014) in initial fitting.

At the 3 to 4Å resolution range typical for large, important nucleic-acid molecules and complexes, even the adjacent base pairs tend to merge together in the density, and increasingly the model must be built as a unit into the continuous spiral shape of the particular duplex. Maintaining that structural integrity in refinement is improved by restraining stacking as well as H-bonding. Note that in this regime, a crystallographer can benefit greatly from the outside information of secondary and tertiary base-pair prediction if based on alignment of many related sequences, which can achieve more reliable sequence-specificity than unaided low-resolution map fitting.

References

Gill ML, Strobel SA, Loria JP (2006) Crystallization and characterization of the Thallium form of Oxytricha nova G-quadruplex, *Nucleic Acids Res* **34**: 4506-4514 (2HBN)

Jain S, Kapral G, Richardson D, Richardson J (2014) "Fitting Tips #7: Getting the pucker right in RNA structures", *Comput Crystallogr Newsletter* **5**: 4-7

Hamma T, Ferre-D'Amare A (2004) Structure of protein L7Ae bound to a K-turn derived from an archaeal box H/ACA sRNA at 1.8 Å resolution, *Structure* **12**: 893-903 (1SDS)

Headd JJ, Echols N, Afonine PV, Grosse-Kunstleve RW, Chen VB, Moriarty NW, Richardson JS, Richardson DC, Adams PD (2012) Use of knowledge-based restraints in *phenix.refine* to improve macromolecular refinement at low resolution, *Acta Crystallogr* **D68**: 381-390

Malashkevich VN, Padlan FC, Toro R, Girvin M, Almo SC (2013) Crystal structure of the aptamer minE-lysozyme complex, to be published (4M4o)

Richardson JS (2015) A context-sensitive guide to RNA & DNA base-pair & base-stack geometry, *Comp Cryst Newsletter* **6**: 47-53

Su N, Chen L, Egli M, Berger JM, Rich A (1999) Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot, *Nat Struct Biol* **6**: 285-292 (437D)

Ye JD, Tereshko V, Frederiksen JK, Koide A, Fellouse FA, Sidhu SS, Koide S, Kossiakoff AA, Piccirilli JA (2008) Synthetic antibodies for specific recognition and crystallization of structured RNA, *Proc Natl Acad Sci USA* **105**:82-87 (2R8S)

New Phaser-MR search panel in Phenix

Robert D. Oeffner, Airlie J. McCoy and Randy J. Read
Department of Haematology, University of Cambridge, Cambridge CB2 0XY, UK

The Phaser-MR graphical user interface (GUI) has undergone relatively few changes since the inception of the modern version of Phenix in 2010. Although fully functional this GUI has sometimes baffled users in the way user input is entered. In particular, users have found the "Search procedure" panel confusing, commonly leading to mistakes when entering input. With the release of Phenix 1.10 these shortcomings have been addressed. Below we discuss the new and the old GUI with a focus on these issues.

The old Phenix Phaser-MR GUI

In figure 1, the old Phaser-MR GUI is shown with the search panel active for the beta-blip tutorial included in Phenix.

This GUI embodies the notion of components defined as ensembles, each of which consists of one or more superimposed atomic coordinate sets or, alternatively, an electron density map. Components are the entities that Phaser MR attempts to place as rigid bodies during an MR search. Anything expected to behave as a rigid object can be used as a search component, which can thus range in size from a structural domain to a whole protein or even an enormous assembly such as the large subunit of the ribosome.

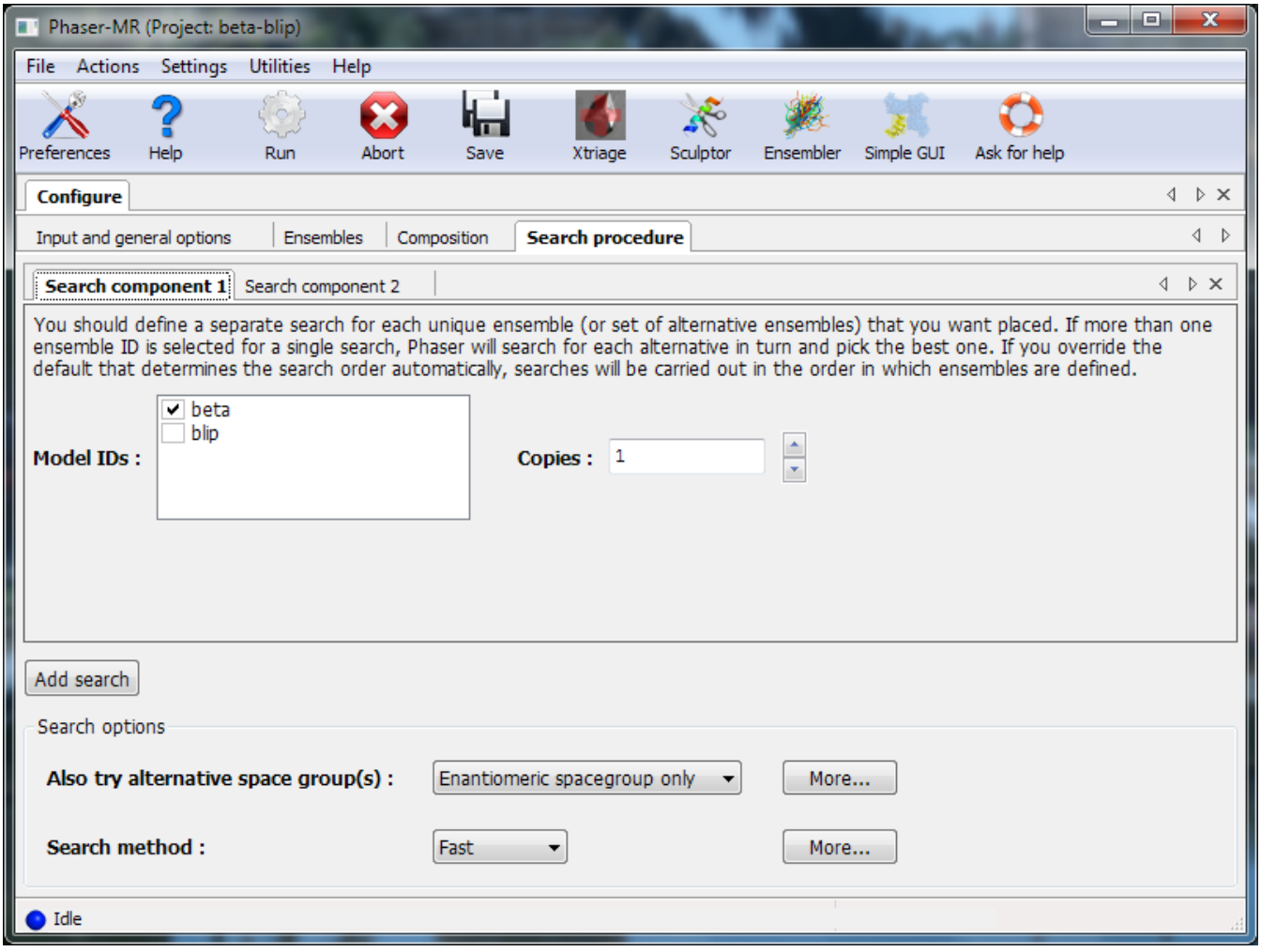


Figure 1: Old style "Search procedure" panel for Phaser MR

Although the representation of the molecular replacement search in the old GUI is technically correct, users were often confused about what to do when setting up a search to find multiple components. The problem was that there are two very different things you can do with multiple ensembles – you can search for multiple different components or you can use them as alternative models for the same component – and the distinction was not intuitively obvious in the interface. Initially there is only one "Search component" page. It features a checklist-box with the available ensembles that can be included in the definition of the particular component. To specify another "Search component" page one has to press the "Add search" button. This button is not particularly prominent. Moreover, because only one "Search component" page is visible at a time the old GUI conveys less clearly what has been specified on other "Search component" pages.

In the case of the beta-blip example it is not uncommon for users to erroneously tick both the beta and the blip ensembles in the checklist-box on the "Search component" page. Although they are expecting Phaser-MR to search for each of the ensembles, they are actually telling it to use these ensembles as alternative models for the same component. As a result, Phaser-MR will do MR searches with each of the ensembles and eventually select the ensemble giving the clearest peak to represent a single component of the crystal. The unintended consequence is that Phaser-MR would consider the MR job done and not attempt to place a second component.

The correct procedure would be for the user to add another "Search component" page and tick complementary ensembles on the checklist-box on each of them so that the total of all components adds up to what the user intends to place if the MR search succeeds. In the case of the beta-blip tutorial this means ticking the blip ensemble on one "Search component" page and ticking the beta ensemble on the other "Search component" page.

The new Phenix Phaser-MR GUI

To overcome these deficiencies a new GUI has been developed, depicted in figure 2. It presents the Search procedure as a tree-list with the root branch of the tree named "Searches". Next to the "Searches" branch is an "Add Search" button that enables the user to add components to search for. These become sub-branches of the "Searches" branch. Next to each "Component" branch is a button allowing the user to add one or more ensembles to the component in question (as alternative choices of model, if more than one is specified), or remove the component and any ensembles altogether. Added ensembles feature as sub-branches on the component branches with names as defined on the "Ensembles" panel. Each of these can be expanded with the mouse to show the names of the constituent coordinate or map files of the ensembles.

To aid the eye icons have been used for all of the branches. Next to the "Searches" branch is an icon of a small protein molecule with a few domains in different colors. Each of the component branches features icons highlighting one of those domains with the remaining domains dimmed. The ensemble branches feature an icon of superposed residues as to illustrate the ensembles of sets of atom coordinates that generally comprises the search model. Expanding an ensemble branch reveals an icon of a single strand and a helix for each of the coordinate files in the ensemble. If a map file is used for an ensemble the icon depicts part of the electron density of a phenylalanine residue. All of these icons only serve as visual cues and any resemblance with targets or models used for a Phaser MR calculation is entirely coincidental.

Figure 2 demonstrates how one would specify the components in the beta-blip tutorial.

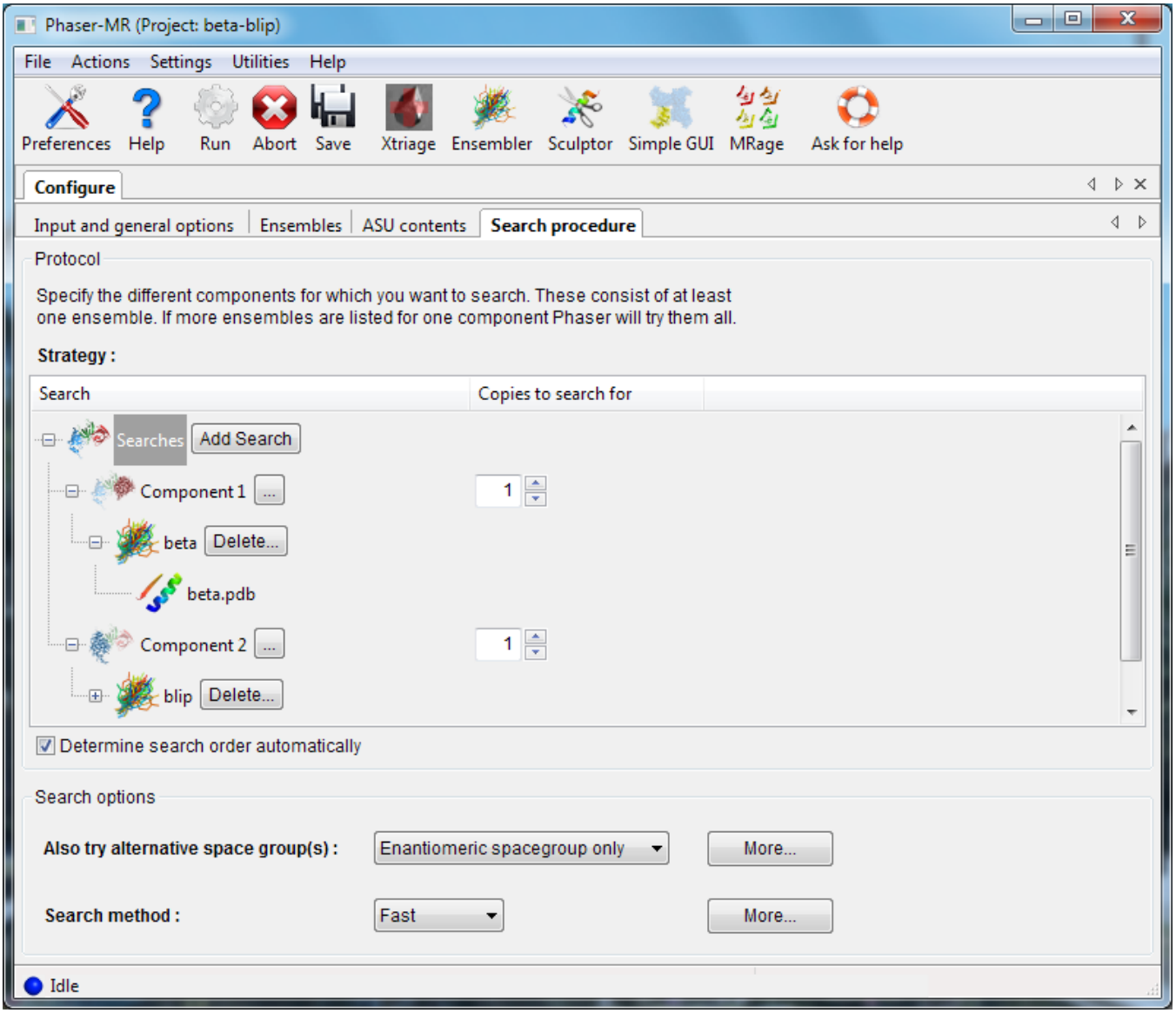


Figure 2: New "Search procedure" panel for Phaser MR

Conclusion

The new GUI addresses the issues that the old GUI suffered from: it presents what has been selected for all components without obscuring selections for one or the other. This makes it clearer to the user which different ensembles are assigned as search models to the various components.

Quantum chemical techniques for minimising ligand geometries in the active site

Nigel W. Moriarty,[a] & Janet E. Deane[b]
[a]*Lawrence Berkeley National Laboratory, Berkeley, CA 94720*
[b]*Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK*

Correspondence email: NWMoriarty@LBL.Gov

The lysosomal enzyme β-galactocerebrosidase is essential for the normal catabolism and recycling of galactosphingolipids. GALC catalyzes the removal of the terminal galactose moiety from substrates including galactosylceramide, the principal lipid component of myelin, and psychosine, a cytotoxic metabolite. Krabbe disease (also known as globoid cell leukodystrophy) is a rare, inherited autosomal recessive disorder caused by loss of GALC function, leading to devastating and ultimately fatal neurodegeneration.

Loss of GALC function can be due to mutations that result in misfolding of the enzyme such that it is degraded by cellular quality control mechanisms. Treatment options for Krabbe Disease are very limited but recent work into pharmacological chaperone therapy (PCT) has shown promise in related diseases. PCT involves the use of small molecules that bind and stabilise partially misfolded proteins to aid in their passage from their site of folding in the endoplasmic reticulum to their site of action elsewhere in the cell.

We recently identified a series of azasugar molecules that specifically bind and stabilize the GALC enzyme and may be excellent PCT candidate molecules. Six new structures were determined of these molecules bound in the GALC active site (Hill et al. 2015). The conformation details of one of these (`dgj` from PDB entry `4ufm`) was not completely clear from the density. The `dgj` molecule is shown in Figure 1 and depicts the charged nitrogen that is believed to hydrogen bond to a protein side-chain. The minimum energy conformation of the isolated ligands is the chair conformation by approximately 23 kJ/mol. However, in the active site, in order to maximise the important interactions the ligand would need to adopt the $^1S_3$ twisted-boat conformation.

The resolution of the x-ray information was 2.4Å and the data was not conclusive concerning the conformation of the ligand. The first technique employed to elucidate the conformation was the feature-enhanced map (FEM) (Afonine et al. 2015). Using both the chair and twisted boat conformations, a FEM was calculated for each. To compare the maps on an equal scale the contour level of each map was determined such that each contour level was set to enclose the same amount of electron density. This was preformed using the Phenix command-line tool, `phenix.map_comparison` (Urzhumtsev et al. 2014). The density was unchanged from the standard $2F_o$-F_c map for the chair conformation but there was a bulge in the density blob near the nitrogen that would indicate the twisted-boat was in fact the conformation in the crystal.

Wanting further proof, a second technique was applied to the problem. It appeared that the ligand was adopting a higher energy conformation in the active state compared to the isolated minima because of an energetic advantage from the surrounding amino acids.

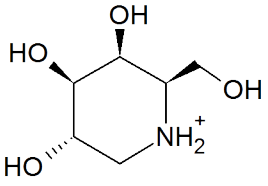


Figure 1: Representation of 1-deoxy-*galacto*-nojirimycin — `dgj` from PDB entry `4ufm`.

Quantum chemistry can provide energy differences sufficiently accurate to determine conformational energy minima of this type. Hydrogen bonding can be approximated with moderately high levels of electron correlation and basis set. In order to keep the size of the calculation tractable a reduced model of the active set was needed.

The steps involved were as follows:

1. Remove all amino acid residues that did not have an atom with 5Å of the selected ligand.
2. Remove backbone atoms expect for C_α. An exception should be made for backbone atoms and termini that interact with the ligand.
3. Add hydrogens to the resulting molecules to produce a neutral moiety except for amino acid chain termini and ligands that are charged or radial.
4. Optimise just the hydrogen positions using the quantum mechanical method of choice. It is recommended that a solvent model be included in the calculation.
5. Check that the hydrogens are still located in chemically sensible positions. Migration of hydrogens can highlight structural problems.
6. Optmise the positions of all ligand atoms and hydrogens of the protein. This allows for optimisation of the hydrogen bonding atoms.

These steps were repeated for both the chair and the twisted-boat models. The chair conformation maintained its starting conformation because of the energy barrier to convert to chair. This is approximately 45 kJ/mol in isolation. The energy difference between the two constrained minimisation gives an indication of the strain energy of the ligand in the active site.

In the case of `dgj`, all quantum mechanical methods indicated that the twisted-boat was the lower energy conformation in the presence of the amino acid side chain atoms in the active site. The best method, B3LYP/6-31G(d,p) with the PCM solvent, favoured the twisted-boat by 59 kJ/mol.

All calculations were performed using GAMESS (Schmidt et al. 2010). A python script was developed to speed the creation of suitable models of the active site and input files for the GAMESS.

References

Afonine, Pavel V, Nigel W. Moriarty, Marat Mustyakimov, Oleg V. Sobolev, Thomas C. Terwilliger, Dusan Turk, Alexandre Urzhumtsev, and Paul D. Adams. 2015. "FEM: Feature-Enhanced Map." *Acta Crystallographica Section D-Biological Crystallography* 71. doi:10.1107/S1399004714028132.

Hill, Chris H., Agnete H. Viuff, Samantha J. Spratley, Stéphane Salamone, Stig H. Christensen, Randy J. Read, Nigel W. Moriarty, Henrik H. Jensen, and Janet E. Deane. 2015. "Azasugar Inhibitors as Pharmacological Chaperones for Krabbe Disease." *Chemical Science*, March. doi:10.1039/C5SC00754B.

Schmidt, M.W., K.K. Baldridge, J.A. Boatz, S.T. Elbert, M.S. Gordon, J.H. Jensen, S. Koseki, et al. 2010. *GAMESS* (version 1 OCT 2010 (R1)). 64 bit Intel. Iowa State University.

Urzhumtsev, Alexandre, Pavel V. Afonine, Vladimir Y. Lunin, Thomas C. Terwilliger, and Paul D. Adams. 2014. "Metrics for Comparison of Crystallographic Maps." *Acta Crystallographica Section D-Biological Crystallography* 70: 2593–2606.

13 typical occupancy refinement scenarios and available options in *phenix.refine*

Pavel V. Afonine
Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Introduction

By definition, occupancy is atomic model parameter (figure 1) that describes the fraction of corresponding atoms in the crystal that occupy the position defined in the crystal structure model. If all molecules in the crystal are identical occupancies for all atoms are unit. Refining occupancy of an atom against experimental data provides an estimate of the frequency to find this atom at given position. Similarly to Atomic Displacement Parameter (ADP or B-factor), occupancy describes the disorder. Unlike ADP that describes small disorder (within harmonic approximation), occupancy describes large-scale discrete disorder.

In practice there are several scenarios when occupancy may differ from unity and therefore its refinement may be desirable.

1. Residue side chain adopts several conformations (figure 2)

This situation is recognized automatically by `phenix.refine` as long as PDB file contains both conformers labeled with unique alternative location identifiers (`altloc id`; figure 2b). Note,

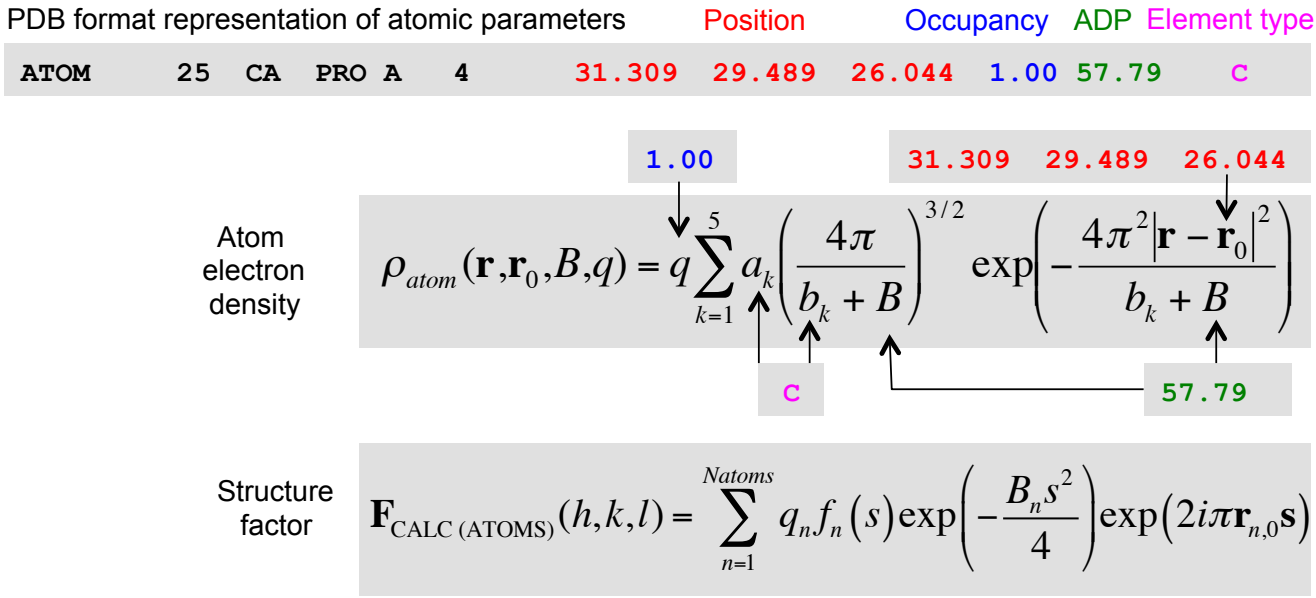


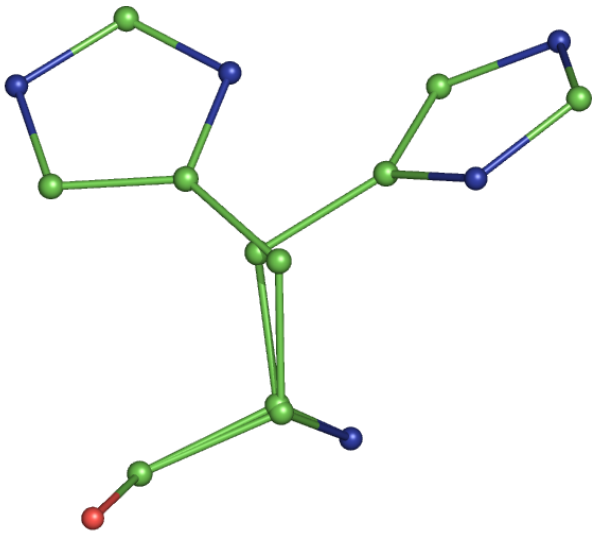
Figure 1: Illustration of atomic model parameters as represented in PDB and their relationship with atomic electron density and structure factors. Position is a triplet of Cartesian coordinates of the atom, $\mathbf{r}_0$. ADP is atomic displacement parameter, B (Å), that describes harmonic atomic vibrations around its mean position $\mathbf{r}_0$. Chemical element type defines form-factor table values to be used. Occupancy q defines probability to find the atom at its mean position $\mathbf{r}_0$.

it is essential that chain and residue `id` be identical for both rotamers. A TER record between the conformers (explicitly atom 1494 and 1495) will stop *`phenix.refine`* from grouping the occupancies of the two rotamers such that their sum adds up to 1. Initial occupancy values set in input PDB file that is subject to occupancy refinement are technically not important as *`phenix.refine`* will make sure that occupancies of all atoms within each conformer will be equal to each other and the sum of occupancies of each conformer will be 1, as shown in figure 2a. However, setting reasonable initial values may improve refinement convergence and thus provide a better refinement result.

2. Continues range of residues adopts several conformations (figure 3a)

This is very similar to the previous case (1) with the following nuance. By default *`phenix.refine`* will account for the fact that adjacent residues have alternative conformations and thus will group occupancies such that identical `altlocs` of consecutive residues will form one group. In this example, there will be two refinable

(a)



(b)

```
ATOM   1485  N   HIS A 501      -2.738  -8.187 -81.483  1.00 46.52           N
ATOM   1486  C   HIS A 501      -4.954  -8.042 -80.397  1.00 44.56           C
ATOM   1487  O   HIS A 501      -5.556  -8.218 -79.338  1.00 46.62           O
ATOM   1488  CA AHIS A 501      -3.511  -7.547 -80.420  0.48 47.00           C
ATOM   1489  CB AHIS A 501      -3.515  -6.028 -80.607  0.48 48.03           C
ATOM   1490  CG AHIS A 501      -2.151  -5.411 -80.606  0.48 46.50           C
ATOM   1491  ND1AHIS A 501      -1.359  -5.360 -79.480  0.48 47.66           N
ATOM   1492  CD2AHIS A 501      -1.447  -4.802 -81.590  0.48 49.62           C
ATOM   1493  CE1AHIS A 501      -0.221  -4.756 -79.772  0.48 50.19           C
ATOM   1494  NE2AHIS A 501      -0.249  -4.407 -81.046  0.48 51.14           N
ATOM   1495  CA BHIS A 501      -3.486  -7.611 -80.375  0.52 46.62           C
ATOM   1496  CB BHIS A 501      -3.337  -6.084 -80.311  0.52 47.45           C
ATOM   1497  CG BHIS A 501      -4.075  -5.347 -81.386  0.52 49.27           C
ATOM   1498  ND1BHIS A 501      -3.442  -4.521 -82.289  0.52 47.63           N
ATOM   1499  CD2BHIS A 501      -5.393  -5.300 -81.692  0.52 52.32           C
ATOM   1500  CE1BHIS A 501      -4.338  -4.001 -83.109  0.52 46.71           C
ATOM   1501  NE2BHIS A 501      -5.529  -4.459 -82.769  0.52 36.75           N
```

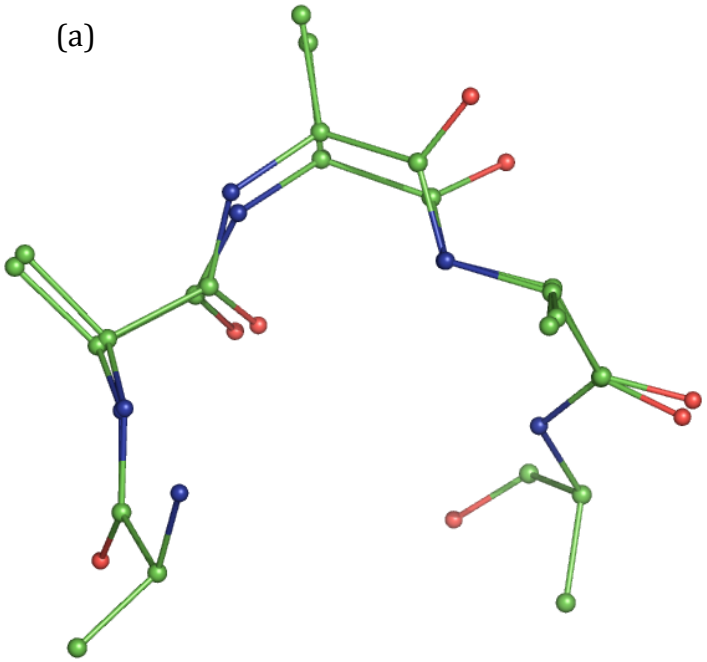
Figure 2: (a) Simple alternative location example above and (b) PDB format input for simple alternative location example below.

occupancies: one for `altloc` A of residues 2-4 and one for `altloc` B of residues 2-4 (figure 3b). As in example 1 both occupancies will add up to 1 and all atoms with the same `altloc id` will have identical occupancy value.

3. Ligand adopts several conformations.

Similarly to case 1 above, a ligand may be modeled with several instances, in this case three copies (figures 4a,b). In case only one copy of partially occupied ligand can be modeled there are two possibilities (figures 4c,d). The differences between the two are following. In

(a)



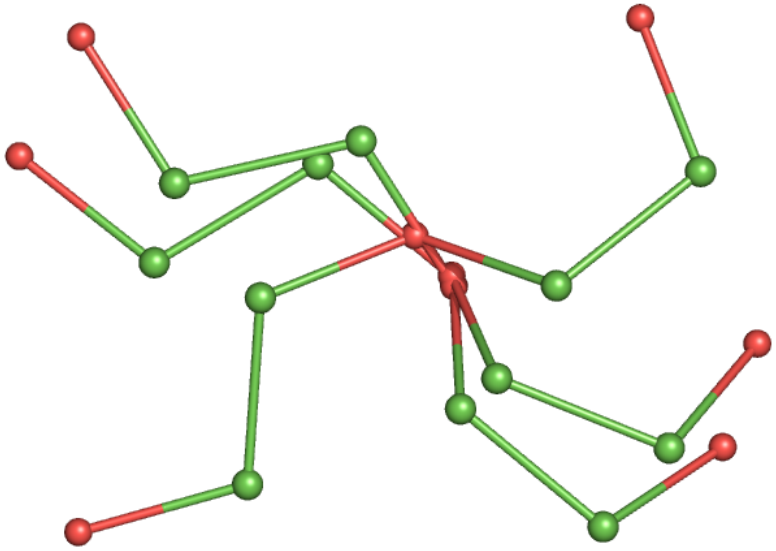
(b)

```
ATOM      1  N    ALA E   1      11.457  32.103  12.052  1.00 40.00           N
ATOM      2  CA   ALA E   1      10.493  31.763  13.123  1.00 40.00           C
ATOM      3  C    ALA E   1       9.291  31.107  12.416  1.00 40.00           C
ATOM      4  O    ALA E   1       8.325  30.669  13.094  1.00 40.00           O
ATOM      5  CB   ALA E   1      10.068  32.938  13.894  1.00 40.00           C
ATOM      6  N  AALA E   2       9.408  30.972  11.116  0.70 40.00           N
ATOM      7  CA AALA E   2       8.377  30.266  10.365  0.70 40.00           C
ATOM      8  C  AALA E   2       8.828  28.858   9.947  0.70 40.00           C
ATOM      9  O  AALA E   2       8.439  27.876  10.572  0.70 40.00           O
ATOM     10  CB AALA E   2       7.896  31.075   9.196  0.70 40.00           C
ATOM      6  N  BALA E   2       9.469  30.978  11.090  0.30 40.00           N
ATOM      7  CA BALA E   2       8.487  30.183  10.293  0.30 40.00           C
ATOM      8  C  BALA E   2       9.057  28.820   9.846  0.30 40.00           C
ATOM      9  O  BALA E   2       8.805  27.806  10.515  0.30 40.00           O
ATOM     10  CB BALA E   2       8.007  30.974   9.055  0.30 40.00           C
ATOM     11  N  AALA E   3       9.688  28.736   8.922  0.70 40.00           N
ATOM     12  CA AALA E   3       9.890  27.338   8.454  0.70 40.00           C
ATOM     13  C  AALA E   3      11.094  26.643   9.096  0.70 40.00           C
ATOM     14  O  AALA E   3      11.317  25.461   8.839  0.70 40.00           O
ATOM     15  CB AALA E   3       9.966  27.369   6.939  0.70 40.00           C
ATOM     11  N  BALA E   3       9.654  28.778   8.644  0.30 40.00           N
ATOM     12  CA BALA E   3      10.138  27.531   8.077  0.30 40.00           C
ATOM     13  C  BALA E   3      11.523  27.206   8.558  0.30 40.00           C
ATOM     14  O  BALA E   3      12.344  26.906   7.766  0.30 40.00           O
ATOM     15  CB BALA E   3      10.156  27.607   6.544  0.30 40.00           C
ATOM     16  N  AALA E   4      11.953  27.354   9.822  0.70 40.00           N
ATOM     17  CA AALA E   4      13.153  26.709  10.368  0.70 40.00           C
ATOM     18  C  AALA E   4      12.894  25.784  11.604  0.70 40.00           C
ATOM     19  O  AALA E   4      13.791  25.098  12.048  0.70 40.00           O
ATOM     20  CB AALA E   4      14.168  27.800  10.647  0.70 40.00           C
ATOM     16  N  BALA E   4      11.803  27.234   9.837  0.30 40.00           N
ATOM     17  CA BALA E   4      13.013  26.554  10.334  0.30 40.00           C
ATOM     18  C  BALA E   4      12.849  25.762  11.621  0.30 40.00           C
ATOM     19  O  BALA E   4      13.844  25.374  12.223  0.30 40.00           O
ATOM     20  CB BALA E   4      14.100  27.587  10.574  0.30 40.00           C
ATOM     21  N    ALA E   5      11.670  25.842  12.211  1.00 40.00           N
ATOM     22  CA   ALA E   5      11.279  24.881  13.256  1.00 40.00           C
ATOM     23  C    ALA E   5       9.827  24.379  13.029  1.00 40.00           C
ATOM     24  O    ALA E   5       8.913  24.936  13.515  1.00 40.00           O
ATOM     25  CB   ALA E   5      11.420  25.549  14.553  1.00 40.00           C
```

Figure 3: (a) Multi-residue alternative location example and (b) PDB format input for mutli-resiude alternative location example.

case of figure 4d a single occupancy per whole ligand will be always refined regardless of starting values constraining it to be between 0 and 1. Also, the ligand will not "see" other atoms that have `altloc` different from A. In the other case (figure 4c) two facts trigger the group occupancy refinement (one occupancy per whole ligand): occupancy is less then 1 and occupancies of all atoms are equal to each other. In this case occupancies that have zero as starting value will not be refined. If occupancies of all were not equal to each other, then occupancies $0<q<1$ will be refined individually each constrained to be in [0,1] range.

(a)



(b)

```
HETATM    1  O4 APEG M   1      -0.091   8.619  56.188  1.00 97.26           O
HETATM    2  C4 APEG M   1       0.240   9.310  57.377  1.00 93.43           C
HETATM    3  C3 APEG M   1       0.193  10.813  57.129  1.00 97.43           C
HETATM    4  O2 APEG M   1      -0.687  11.080  56.017  1.00102.78           O
HETATM    5  C2 APEG M   1      -0.510  12.375  55.423  1.00 89.67           C
HETATM    6  C1 APEG M   1      -1.088  13.457  56.368  1.00 88.16           C
HETATM    7  O1 APEG M   1      -1.445  14.612  55.612  1.00 78.51           O
HETATM    8  O4 BPEG M   1       0.133  10.917  53.508  1.00 97.26           O
HETATM    9  C4 BPEG M   1       0.605  10.017  54.492  1.00 93.43           C
HETATM   10  C3 BPEG M   1       0.387  10.610  55.879  1.00 97.43           C
HETATM   11  O2 BPEG M   1      -0.693  11.565  55.819  1.00102.78           O
HETATM   12  C2 BPEG M   1      -0.733  12.479  56.926  1.00 89.67           C
HETATM   13  C1 BPEG M   1      -1.229  11.734  58.189  1.00 88.16           C
HETATM   14  O1 BPEG M   1      -1.810  12.666  59.098  1.00 78.51           O
HETATM   15  O4 CPEG M   1       0.133   8.810  55.355  1.00 97.26           O
HETATM   16  C4 CPEG M   1       0.605   9.212  56.627  1.00 93.43           C
HETATM   17  C3 CPEG M   1       0.387  10.709  56.807  1.00 97.43           C
HETATM   18  O2 CPEG M   1      -0.693  11.135  55.949  1.00102.78           O
HETATM   19  C2 CPEG M   1      -0.733  12.550  55.711  1.00 89.67           C
HETATM   20  C1 CPEG M   1      -1.229  13.272  56.989  1.00 88.16           C
HETATM   21  O1 CPEG M   1      -1.810  14.525  56.636  1.00 78.51           O
```

(c)

```
HETATM    1  O4  PEG M   1      -0.091   8.619  56.188  0.70 97.26           O
HETATM    2  C4  PEG M   1       0.240   9.310  57.377  0.70 93.43           C
HETATM    3  C3  PEG M   1       0.193  10.813  57.129  0.70 97.43           C
HETATM    4  O2  PEG M   1      -0.687  11.080  56.017  0.70 92.78           O
HETATM    5  C2  PEG M   1      -0.510  12.375  55.423  0.70 89.67           C
HETATM    6  C1  PEG M   1      -1.088  13.457  56.368  0.70 88.16           C
HETATM    7  O1  PEG M   1      -1.445  14.612  55.612  0.70 78.51           O
```

(d)

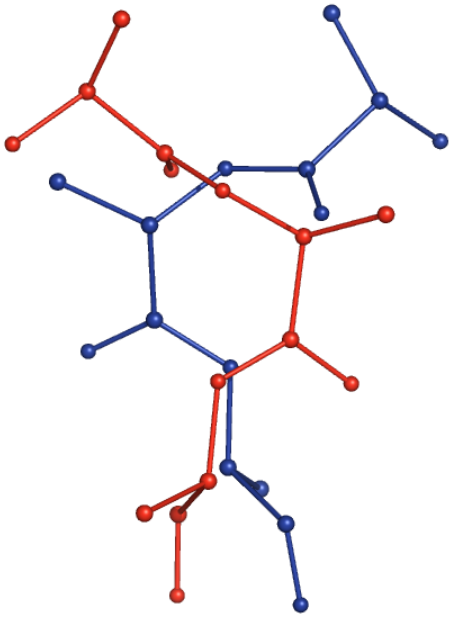
```
HETATM    1  O4 APEG M   1      -0.091   8.619  56.188  0.70 97.26           O
HETATM    2  C4 APEG M   1       0.240   9.310  57.377  0.70 93.43           C
HETATM    3  C3 APEG M   1       0.193  10.813  57.129  0.70 97.43           C
HETATM    4  O2 APEG M   1      -0.687  11.080  56.017  0.70 62.78           O
HETATM    5  C2 APEG M   1      -0.510  12.375  55.423  0.70 89.67           C
HETATM    6  C1 APEG M   1      -1.088  13.457  56.368  0.70 88.16           C
HETATM    7  O1 APEG M   1      -1.445  14.612  55.612  0.70 78.51           O
```

Figure 4: Example of ligands adopting one partial or several alternative conformation (a), and their representation in PDB file (b,c,d).

4. Overlapping chains

It may happen that stretches of two or more chains overlap in space such that when one is present the other one isn't (figure 5a). This situation is not recognized automatically by *phenix.refine* but can be dealt with successfully. This requires two actions. Firstly, overlapping atoms of both chains need to be assigned different `altloc` (figure 5b); this will ensure that these atoms will not be pushed apart by non-bonded repulsion. Secondly, one needs to compose a parameter file telling *phenix.refine* what occupancies need to be coupled (figure 5c); this will override any default behavior applicable to these groups of atoms.

(a)



(b)

```
ATOM      6  N   AALA A  12       8.148  29.454  12.242  0.70 40.00           N
ATOM      7  CA  AALA A  12       7.117  28.748  11.491  0.70 40.00           C
ATOM      8  C   AALA A  12       7.568  27.340  11.073  0.70 40.00           C
ATOM      9  O   AALA A  12       7.179  26.358  11.698  0.70 40.00           O
ATOM     10  CB  AALA A  12       6.636  29.557  10.322  0.70 40.00           C
ATOM     11  N   AALA A  13       8.428  27.218  10.048  0.70 40.00           N
ATOM     12  CA  AALA A  13       8.630  25.820   9.580  0.70 40.00           C
ATOM     13  C   AALA A  13       9.834  25.125  10.222  0.70 40.00           C
ATOM     14  O   AALA A  13      10.057  23.943   9.965  0.70 40.00           O
ATOM     15  CB  AALA A  13       8.706  25.851   8.065  0.70 40.00           C
ATOM     16  N   AALA A  14      10.693  25.836  10.948  0.70 40.00           N
ATOM     17  CA  AALA A  14      11.893  25.191  11.494  0.70 40.00           C
ATOM     18  C   AALA A  14      11.634  24.266  12.730  0.70 40.00           C
ATOM     19  O   AALA A  14      12.531  23.580  13.174  0.70 40.00           O
ATOM     20  CB  AALA A  14      12.908  26.282  11.773  0.70 40.00           C
TER
ATOM      6  N   BALA B   2       6.766  24.529   8.198  0.30 40.00           N
ATOM      7  CA  BALA B   2       6.833  26.021   8.208  0.30 40.00           C
ATOM      8  C   BALA B   2       8.019  26.552   9.041  0.30 40.00           C
ATOM      9  O   BALA B   2       9.077  26.857   8.469  0.30 40.00           O
ATOM     10  CB  BALA B   2       5.515  26.632   8.736  0.30 40.00           C
ATOM     11  N   BALA B   3       7.777  26.826  10.333  0.30 40.00           N
ATOM     12  CA  BALA B   3       8.789  27.424  11.187  0.30 40.00           C
ATOM     13  C   BALA B   3       9.696  26.384  11.781  0.30 40.00           C
ATOM     14  O   BALA B   3       9.907  26.409  12.941  0.30 40.00           O
ATOM     15  CB  BALA B   3       8.132  28.212  12.329  0.30 40.00           C
ATOM     16  N   BALA B   4      10.257  25.482  11.016  0.30 40.00           N
ATOM     17  CA  BALA B   4      11.422  24.727  11.513  0.30 40.00           C
ATOM     18  C   BALA B   4      12.556  24.525  10.521  0.30 40.00           C
ATOM     19  O   BALA B   4      13.441  23.714  10.771  0.30 40.00           O
ATOM     20  CB  BALA B   4      10.966  23.351  11.964  0.30 40.00           C
```

(c)

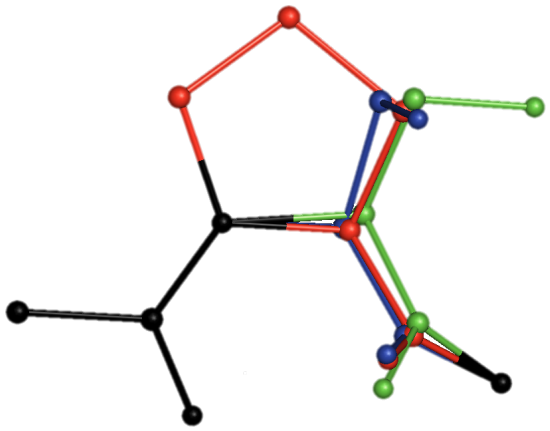
```
refinement {
  refine {
    occupancies {
      constrained_group {
        selection = chain A and resseq 12:14 and altloc A
        selection = chain B and resseq 2:4 and altloc B
      }
    }
  }
}
```

Figure 5: Different chains that overlap in space (a), PDB snippet as an example and Phenix Phil parameters showing how to handle occupancies of overlapping chains in refinement (c).

5. Site shared by several chemical moieties

PDB entry `1EJG` serves a good example of such case. Figures 6a focuses on residue number 22, which is `PRO` and two `CYS` in alternative conformations, being all together three residues sharing the same spot. Figure 6b exemplifies kinds of PDB syntax that will be automatically recognized by *`phenix.refine`*. Note, residue and chain ids are identical, while `altloc` are different, as well as residue names are allowed to be different.

(a)



(b)

```
ATOM    387  CA  THR A  21       3.664  10.576  -3.461  1.00  1.42           C
ATOM    388  C   THR A  21       4.915  11.347  -3.001  1.00  1.58           C
ATOM    389  O   THR A  21       5.844  10.729  -2.478  1.00  2.24           O
ATOM    400  N   PRO A  22       4.915  12.683  -3.102  1.00  1.83           N
ATOM    401  CA APRO A  22       6.042  13.429  -2.601  0.57  1.81           C
ATOM    402  C  APRO A  22       6.387  13.122  -1.160  0.57  1.66           C
ATOM    403  O  APRO A  22       5.480  13.006  -0.345  0.57  2.08           O
ATOM    404  CB APRO A  22       5.655  14.896  -2.744  0.57  2.86           C
ATOM    405  CG APRO A  22       4.661  14.854  -4.058  0.57  2.66           C
ATOM    406  CD APRO A  22       3.957  13.505  -3.910  0.57  2.27           C
ATOM    414  CA BSER A  22       6.034  13.399  -2.687  0.21  1.55           C
ATOM    415  C  BSER A  22       6.367  13.062  -1.223  0.21  2.92           C
ATOM    416  O  BSER A  22       5.412  13.050  -0.345  0.21  1.87           O
ATOM    417  CB BSER A  22       5.409  14.835  -2.876  0.21  2.60           C
ATOM    418  OG BSER A  22       4.760  15.243  -1.635  0.21  2.11           O
ATOM    420  CA CSER A  22       6.112  13.653  -2.656  0.22  2.31           C
ATOM    421  C  CSER A  22       6.354  13.275  -1.187  0.22  1.92           C
ATOM    422  O  CSER A  22       5.636  12.705  -0.270  0.22  1.77           O
ATOM    423  CB CSER A  22       5.605  15.097  -2.687  0.22  3.56           C
ATOM    424  OG CSER A  22       6.750  15.771  -2.280  0.22  6.38           O
ATOM    426  N   GLU A  23       7.651  13.190  -0.860  1.00  1.86           N
```

Figure 6: Example of a same space shared by different chemical moieties (a) and its representation in PDB file that is handled automatically in refinement.

6. Exchangeable H/D sites

While purely hydrogenated samples are possible, for best results neutron crystallography requires samples to be deuterated or partially deuterated. This means that some of the hydrogen (H) atoms in the structure may be partially or fully substituted with deuterium (D) atoms. Figure 7 shows syntax that is used to record this situation in a PDB file. As long as PDB file is composed as illustrated in figure 7 *`phenix.refine`* will handle this situation automatically: occupancies of H and D will be coupled such that each one is in [0,1] range and their sum is exactly 1. In addition *`phenix.refine`* will constrain coordinates and B-factors of H and D to be identical.

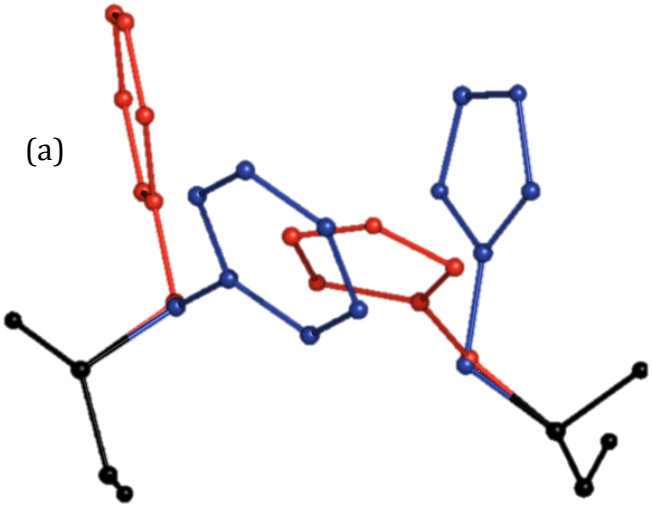
```
ATOM      1  CA  SER A  22       6.034  13.399  -2.687  1.00  1.55           C
ATOM      2  C   SER A  22       6.367  13.062  -1.223  1.00  2.92           C
ATOM      3  O   SER A  22       5.412  13.050  -0.345  1.00  1.87           O
ATOM      4  CB  SER A  22       5.409  14.835  -2.876  1.00  2.60           C
ATOM      5  OG  SER A  22       4.760  15.243  -1.635  1.00  2.11           O
ATOM      6  HB2 SER A  22       6.201  15.544  -3.119  1.00  2.60           H
ATOM      7  HB3 SER A  22       4.670  14.804  -3.676  1.00  2.60           H
ATOM      8  HG ASER A  22       4.050  14.608  -1.403  0.30  2.11           H
ATOM      9  DG BSER A  22       4.050  14.608  -1.403  0.70  2.11           D
```

Figure 7: Example of PDB representation of exchangeable H/D sites.

7. Concerted (coupled) alternative conformations

Figures 8a,b illustrate this scenario. Here two residues, PHE 73 in chain A and HIS 90 in chain B, both have two alternative conformations situated in space such that the same spot is occupied part of the time by conformation B of PHE and part of the time is occupied by conformation A of HIS. Seemingly clashing side chains of these residues are not clashing in reality because they are never present at

(a)



(b)

```
ATOM      5  N   PHE A  73      12.614  27.642  22.065  1.00 30.00           N
ATOM      6  CA  PHE A  73      12.796  27.740  23.498  1.00 30.00           C
ATOM      7  C   PHE A  73      11.723  28.575  24.143  1.00 30.00           C
ATOM      8  O   PHE A  73      11.292  29.592  23.600  1.00 30.00           O
ATOM     10  CB APHE A  73      14.228  28.251  23.830  0.70 30.00           C
ATOM     11  CG APHE A  73      15.354  27.370  23.370  0.70 30.00           C
ATOM     12  CD1APHE A  73      15.774  26.299  24.140  0.70 30.00           C
ATOM     13  CD2APHE A  73      15.993  27.613  22.165  0.70 30.00           C
ATOM     14  CE1APHE A  73      16.808  25.487  23.718  0.70 30.00           C
ATOM     15  CE2APHE A  73      17.029  26.805  21.739  0.70 30.00           C
ATOM     16  CZ APHE A  73      17.437  25.739  22.515  0.70 30.00           C
ATOM     17  CB BPHE A  73      14.173  28.323  23.818  0.30 30.00           C
ATOM     18  CG BPHE A  73      14.460  29.621  23.118  0.30 30.00           C
ATOM     19  CD1BPHE A  73      14.929  29.633  21.815  0.30 30.00           C
ATOM     20  CD2BPHE A  73      14.261  30.829  23.764  0.30 30.00           C
ATOM     21  CE1BPHE A  73      15.191  30.826  21.170  0.30 30.00           C
ATOM     22  CE2BPHE A  73      14.524  32.025  23.123  0.30 30.00           C
ATOM     23  CZ BPHE A  73      14.989  32.024  21.824  0.30 30.00           C
TER
ATOM     24  N   HIS B  90      16.239  33.606  26.584  1.00 30.00           N
ATOM     25  CA  HIS B  90      15.150  32.655  26.764  1.00 30.00           C
ATOM     26  C   HIS B  90      15.091  32.170  28.215  1.00 30.00           C
ATOM     27  O   HIS B  90      16.050  31.585  28.724  1.00 30.00           O
ATOM     28  CB BHIS B  90      15.368  31.394  25.935  0.30 30.00           C
ATOM     29  CG BHIS B  90      16.683  31.284  25.230  0.30 30.00           C
ATOM     30  ND1BHIS B  90      17.332  30.072  25.122  0.30 30.00           N
ATOM     31  CD2BHIS B  90      17.420  32.174  24.522  0.30 30.00           C
ATOM     32  CE1BHIS B  90      18.429  30.232  24.402  0.30 30.00           C
ATOM     33  NE2BHIS B  90      18.509  31.496  24.030  0.30 30.00           N
ATOM     35  CB AHIS B  90      15.468  31.394  25.935  0.70 30.00           C
ATOM     36  CG AHIS B  90      15.431  31.564  24.449  0.70 30.00           C
ATOM     37  ND1AHIS B  90      14.958  30.558  23.632  0.70 30.00           N
ATOM     38  CD2AHIS B  90      15.698  32.617  23.640  0.70 30.00           C
ATOM     39  CE1AHIS B  90      14.968  30.980  22.379  0.70 30.00           C
ATOM     40  NE2AHIS B  90      15.416  32.222  22.355  0.70 30.00           N
```

(c)

```
refinement {
  refine {
    occupancies {
      constrained_group {
        selection = chain A and resseq 73 and altloc A or \
                    chain B and resseq 90 and altloc A or
        selection = chain A and resseq 73 and altloc B or \
                    chain B and resseq 90 and altloc B or
      }
    }
  }
}
```

Figure 8: Example of concerted alternative conformations. (a) Two side chains have two alternative conformations, and two of them clash, (b) PDB snippet that exemplifies the situation, (c) Phenix Phil parameters defining how occupancies of the conformers need to be coupled in refinement.

that space simultaneously. The fact that "clashing" side chains have different `altloc id` instructs the refinement program to not use non-bonded repulsion term for involved atoms. While this may be automated in future, currently it's a requirement manually construct a file to instruct the refinement program how occupancies of these residues should be refined: `altloc` A should be the same for atoms in both residues, likewise for `altloc` B, and their sum adds up to 1. This can be done by composing a parameter file as shown in figure 8c.

A similar situation would be to have a single partially occupied water (or any ligand) instead of one of the residues. In such case the corresponding PDB file would have one ATOM entry and matching non-blank `altloc` for the water. It is essential to have matching non-blank `altloc` for the water even though it has just one entry (one atom) in the file because this will instruct the program to disable repulsions between this water and corresponding conformer of the residue side chain.

8. Occupancy refinement and special positions

If a single atom such as water or metal ion sits exactly on or near to a special position, for example, two-fold axis, there are two possible scenarios. One is occupancy of the atom in input is equal to 1. In this case the symmetry factor ½ is applied to the occupancy internally and the atom always stays with occupancy 1 in the file. Alternatively, if occupancy is set to 0.5 before the refinement (or, more generally, is not equal to 1) then symmetry factor is not applied internally and also the occupancy will be refined individually following *`phenix.refine`* convention to refine occupancies of atoms that are different from 0 and 1.

9. Occupancy refinement and heavy atoms

Atoms heavier than typical protein atoms (C,O,N) are more electron rich and therefore more pronounced in Fourier maps. Likewise, errors in parameters of such atoms are more visible on residual Fourier maps. It is not uncommon to observe residual map features around heavy atoms. One of several possible reasons for this is that the ion may not be fully occupied and therefore modeling it with full occupancy may not be adequate. Refining occupancies of heavy atoms in such case may be desirable.

10. Radiation damage

Radiation damage may result in structure changes such as loss of hydroxyl groups from tyrosine residues or decarboxylation of aspartates and glutamates. This effect may not be identical for all unit cells of the crystal meaning that occupation of corresponding groups may vary from unit cell to unit cell. If residual map suggests this situation then refining occupancies of these groups may be desirable.

11. Partial Selenium incorporation in SE-MET

This is typically observed as negative residual density around Selenium atoms. Enabling individual occupancy refinement of Se atoms to accommodate the fact that they may not fully substitute Sulfur is desirable in this case.

12. Single partially occupied instance of residue or ligand

If residue or ligand has multiple conformations but only one can be identified and modeled, it is desirable to refine one occupancy factor for involved atoms. This can be done in two ways. One is to assign a non-blank `altloc id` to these atoms. Another possibility is to set occupancy value to these atoms that is greater than 0 and less then 1. It is important that this partial occupancy is

(a)
```
refinement {
  refine {
    occupancies {
      constrained_group {
        selection = chain A and resseq 73 and altloc A or \
                    chain B and resseq 90 and altloc A
      }
      constrained_group {
        selection = chain A and resseq 73 and altloc B
      }
      constrained_group {
        selection = chain B and resseq 90 and altloc B
      }
    }
  }
}
```

(b)
```
refinement {
  refine {
    occupancies {
      remove_selection = (chain A and resseq 73) or (chain B and resseq 90)
  }
}
```

(c)
```
refinement {
  refine {
    occupancies {
      individual = element Zn or water
  }
}
```

Figure 9: Examples of overriding default refinement behavior. See text for details.

equal for all involved atoms; different occupancy values will trigger individual occupancy refinement.

13. Overriding default behavior

Any user-provided selections for occupancy refinement will override the default behavior of the refinement package. For example, considering the scenario shown in figure 8: it is possible to request the program to refine a single occupancy for all atoms with `altloc` A in both residues, and one occupancy per each B conformer, resulting in this case in three refinable occupancy factors. Parameter file shown in figure 9a is what's needed for this situation.

In some cases it may be desirable to keep partial occupancy value during refinement (that is not refine it as *`phenix.refine`* would do otherwise by default). This can be done by excluding selected atoms from default behavior. Using case 7 as an example, it is possible to keep occupancies of involved residues at 0.3 and 0.7 by using this syntax in figure 9b.

Finally, it is possible to enable individual occupancy refinement for atoms that otherwise by default are not subject to occupancy refinement. Example in figure 9b enables individual occupancy refinement for Zinc ion and all water molecules; refined occupancy values will range between 0 and 1.

Summary

`phenix.refine` automatically considers occupancy refinement for atoms that have occupancy greater than zero and not equal to one, as well as if atoms have non-blank `altloc` identifier. Refinement scenarios may vary

broadly and most typical cases are described above. User has a full control over occupancy refinement, such as a possibility to undo or override default occupancy refinement behavior for any selected atoms, request individual occupancy refinement for selected atoms or add constrained occupancy groups.

A context-sensitive guide to RNA & DNA base-pair & base-stack geometry

Jane S. Richardson
Department of Biochemistry, Duke University, Durham NC 27710, USA
Correspondence email: Jane.Richardson@duke.edu

Introduction

Increasingly large and biologically important nucleic-acid/protein complexes are now being solved, typically at resolutions of 3-4Å and even lower, where individual bases are seldom clearly visible. There is thus an increasing need for reliable rules of thumb to guide their modeling. This article combines prior textbook-level knowledge (e.g., Saenger 1984; Creighton 2011) with a survey of the low-B-factor parts of recent high-resolution RNA and DNA structures (≤ about 2Å) with deposited structure-factor data, for which the base positions and orientations can be confirmed as unambiguous to very high accuracy. The goal here is to define expectations for planarity and twist values of base stacking and pairing relationships for various RNA and DNA double-helix types, including common local irregularities and accommodations to complex tertiary-structures or functional sites. These expectations can then guide manual modeling at poorer resolution, and especially can form the basis for setting the parameters of automated restraints.

Base-stacking and its variation

Base stacking is a very strong effect, optimizing both van der Waals contact and pi-pi interaction. It occurs with amino acids or small molecules as well as with other bases. If backbone conformation permits, stacked bases are closely parallel at optimal 3.5Å separation, but usually sheared from complete overlap and often better stacked between successive base-pairs than between successive bases on a given strand. As typical, the green and blue dots of favorable all-atom vdW contacts (Word 1999) are extensive for the lower stack in figure 1 (2R8S; Ye 2008). When good stacking is not possible, bases still maintain vdW contact at an edge, as between the upper two bases in fig. 1. Stacking may be opened up at helix junctions, to compromise between interactions on opposite ring faces, or to allow for an inserted base (figure 2), but is essentially always near optimal within regular duplexes.

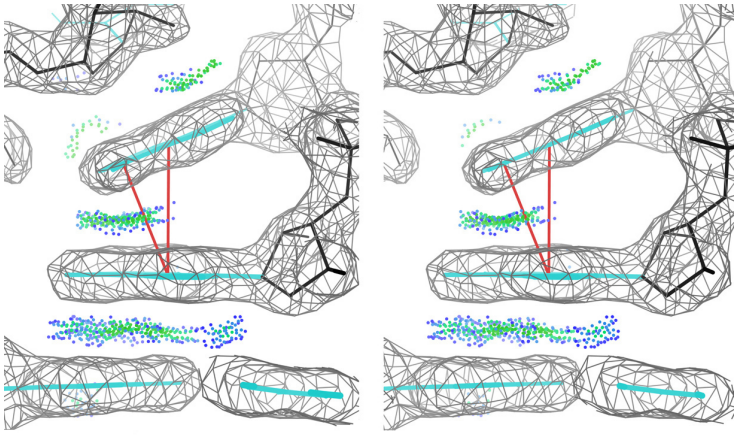


Figure 1: Optimal vs compromised base stacking. All-atom contact dots show a wide area of optimal, parallel stacking at 3.5Å separation for the lower pair of bases, but the upper pair are tilted 22° (normal in red) by their backbone conformations and touch only at one edge. 2R8S Fab/P4P6 ribozyme complex (1.95 Å), around residue 183.

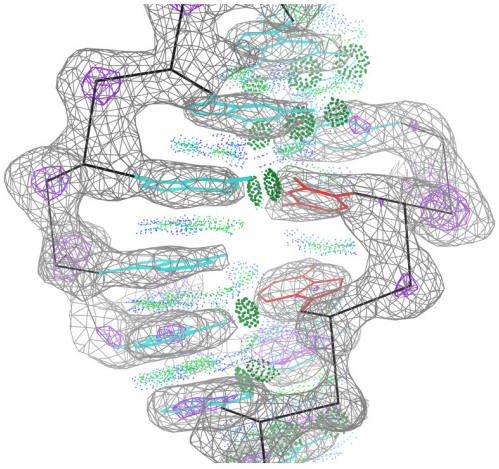


Figure 2: Stack-opening by 44° (red bases), to accommodate an inserted base in the left strand. U205-A206 in 2R8S.

Base-pair propeller-twist

The individual RNA and DNA bases are strongly planar, aromatic systems, but base-pairs are only indirectly pulled toward co-planarity, by similarity of stacking directions and by the base-pair hydrogen bonds. Their relative orientation also responds to limitations from the local backbone conformation and the demands of forming tertiary structure.

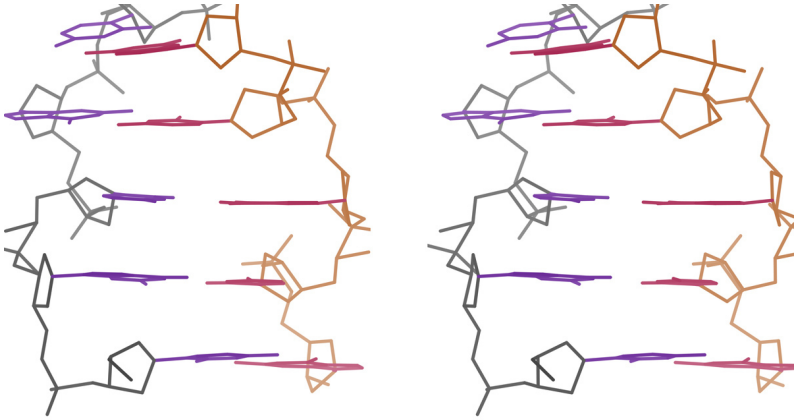


Figure 3: Stereo of co-planar base-pairs with essentially zero propeller-twist in a Z-form DNA 12-mer. 4oCB at 0.75Å.

In this context, the two most revealing of the classic parameters that measure deviation from base-pair co-planarity (Olson 2001) are the "propeller-twist" torsion around a line joining the two bases, and the "buckle" angle of their bend across the line of base-pair H-bonds. Individual nucleotides and the helices they form are both handed, so it should not be a surprise that propeller-twist is typically not 0°. The only place in this dataset that has multiple, truly flat base pairs with near-zero twist and buckle is in the rare, left-handed Z-form DNA duplexes (figure 3). Perhaps the two very different, alternating conformations that make up Z-DNA have opposite twist preferences that cancel each other out.

Each type of double helix shows a typical average for base-pair propeller-twist, from -15° (left-handed) to +11° (right-handed), not well correlated to duplex handedness. Figure 4 shows stereos of: a) regular B-form DNA, here at -15° left-handed twist (1K61; Aishima 2002), while short B-DNAs average about -12°; b) all-GC-pair A-form RNA with -12° left-handed average twist (4MS9; Sheng 2014); c) mixed AT/GC A-form duplex within a complex RNA, with -14° average twist (2R8S); and d) a parallel poly-A duplex with +11° right-handed average twist (4JRD; Safaee 2013). The twist of individual pairs varies about ±5° from average (modulated by GC vs AT, helix bending, and specific local interactions), and thus can occasionally be quite near zero.

Base-pair buckle

Base-pairs can buckle, as well as twist, away from co-planarity. Buckle is a bend between the two base planes, folded along the line of the base-pair H-bonds. There is very little buckle in regular double-helix types for either DNA or RNA, as can be seen in figure 4. However, figure 5 shows the 33° buckle of a specific non-canonical A-G base-pair in the kink-turn binding site of 1SDS (Hamma 2004).

Large deviations from base-pair co-planarity in complex RNA structures

Quite substantial deviations from co-planar base-pairs (either in twist or buckle) are sometimes needed to make stable, unique RNA tertiary structure and to support function for ribozymes and for RNA or DNA aptamers. These occur at junctions between helices, at characteristic bends or kinks, at active sites, and at specific binding sites for other molecules. High propeller-twist can occur in Watson-Crick or non-canonical base-pairs, and can be either right- or left-handed. Figures 6a and 6b show two individual base-pairs with a high propeller-twist of about +40°, one example in end view (3G9Y; Loughlin 2009) and one in side view (2R8S).

The maximal tolerance for base-pair propeller-twist differs as a function of how many hydrogen bonds they share. 40° is high but feasible for 2-H-bond pairs, as seen in

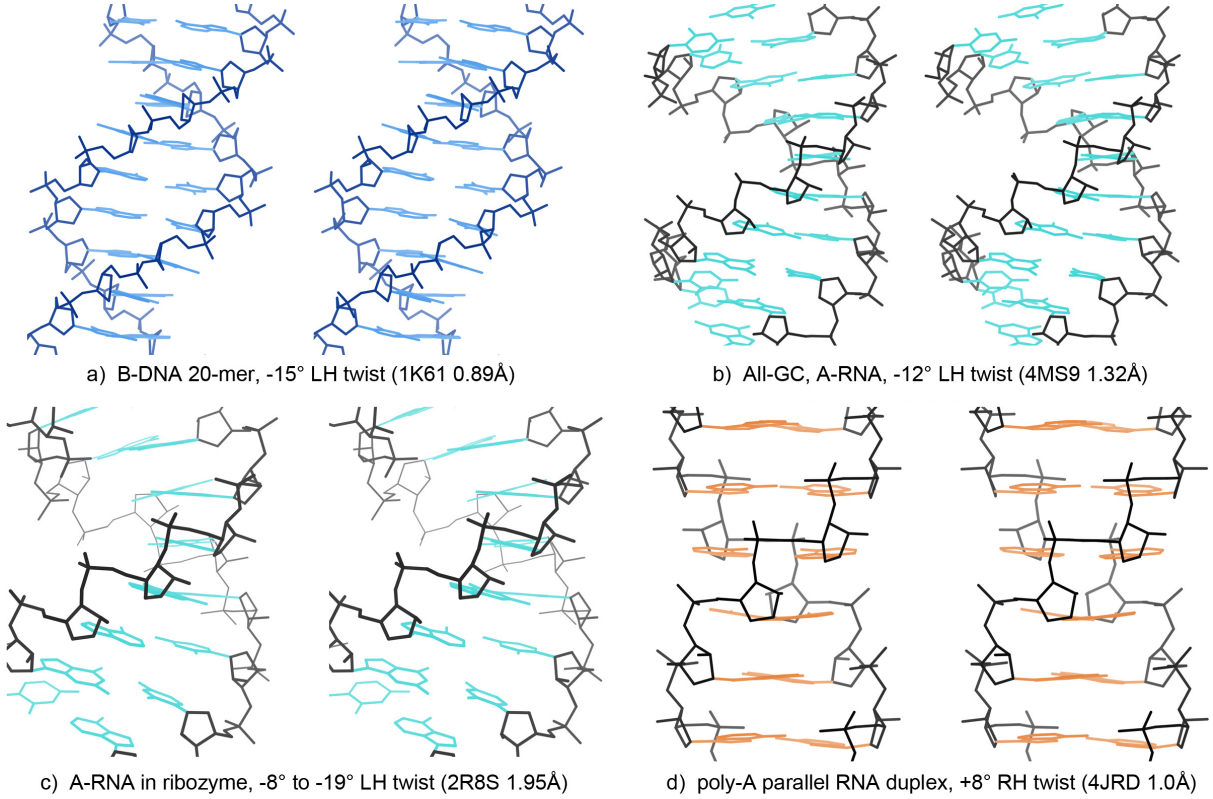


Figure 4: Basepair propeller-twist typical of different double-helix forms, best visible along basepairs where stands cross.

figure 6. When only a single H-bond joins the pair, the twist is presumably free to follow whatever else is needed, as in figure 7a. Note, though, that a pair between two bases adjacent in sequence can only make one (non-canonical) H-bond but is usually close to co-planar, as occurs in S-motifs or dinucleotide platforms. For 3-H-bond pairs the maximum twist is lower but often significant, as for the -24° propeller-twist of the GC pair in figure 7b.

Context consequences, and conservation of non-coplanar base pairs

Twist and buckle of base pairs can best be understood as a somewhat flexible tie through the H-bonds, dominated by the directionality of each stack that contains one base of the pair. Within a helix, the two stacks are close in direction, but have a relative twist characteristic of the helix type. In a helix junction, those two base stacks almost never come from the same direction. Figure 8 shows two cases of cross-stack relationships in helix junctions of an intron (2R8S) and an aptamer (3DD2; Long 2008), with one stack in orange and the other in blue, and joined at the junction by one or more base pairs (purple H-bond dots) twisted + or - by the torsion angle between the stack directions. Such places are among the most

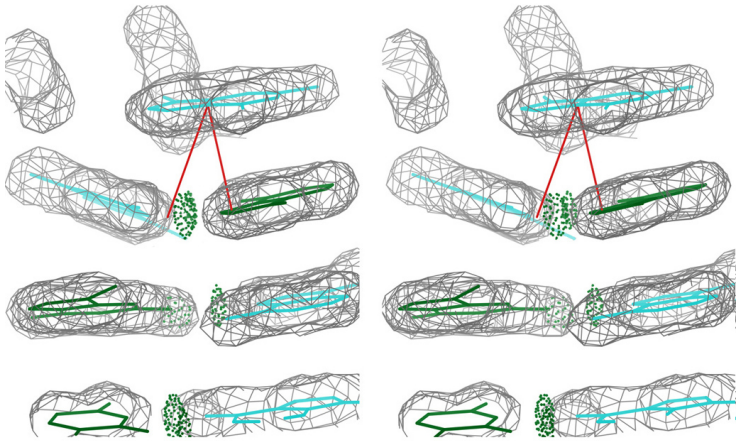
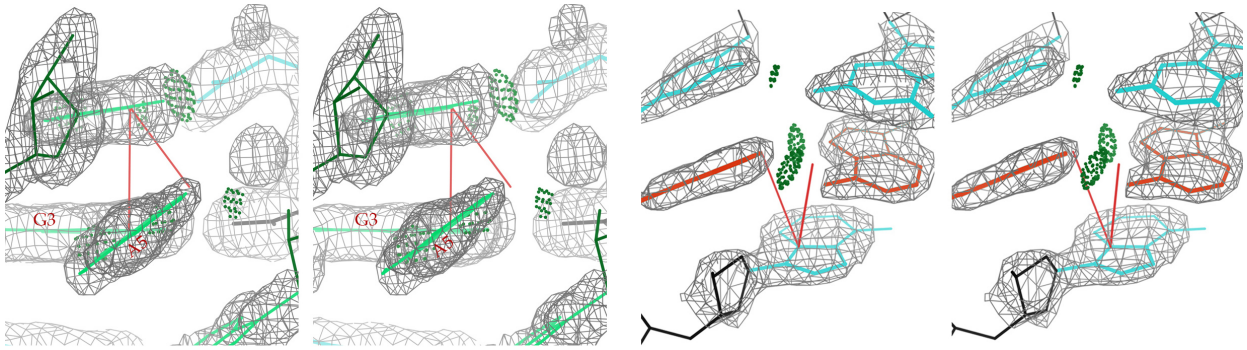


Figure 5: 33° buckle of base-pair in kink turn. Base normal in red (1SDS 1.8Å).



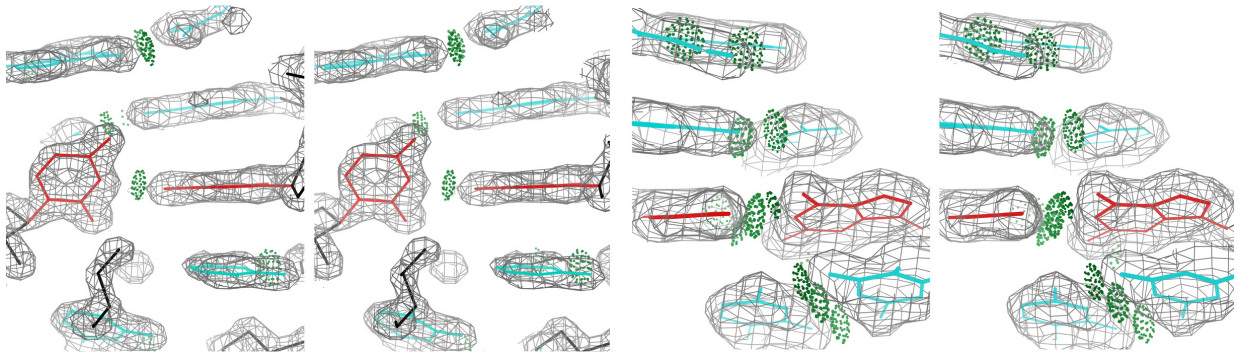
a) +37° twisted G-A base pair in splice-site/ZnF complex (3G9Y 1.4Å)

b) +43° twisted A-A base pair at helix junction in ribozyme (2R8S)

Figure 6: Stereos of high propeller-twist RNA base-pairs with 1 good H- buckle of base-pair in kink turn. Base normal in red (1SDS 1.8Å).

functionally important for large RNAs and their complexes, and sometimes also for DNAs.

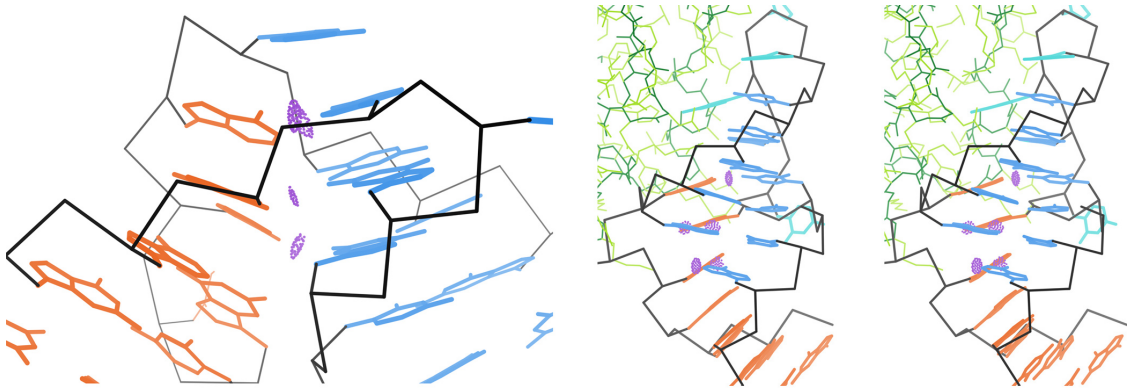
Complex regions of nucleic acid tertiary structure usually involve strongly non-coplanar base pairs and are the glue for forming specific large molecules and the arrangements that enable catalysis and binding. The strongly non-coplanar base pairs (or poorly stacked successive bases) are highlighted in red in figure 9 for overview examples of a riboswitch (4FEJ; Stoddard 2013) and a DNA aptamer (3ZH2; Cheung



a) Free twist (52°) for a 1-H-bond base pair (4FEJ riboswitch 1.5Å)

b) Lower maximum twist (-24°) for GC 3-H-bond base pair (2R8S)

Figure 7: Differing maximum propeller-twist as a function of number of base-pair H-bonds.



a) + twist helix-junction base pairs, P4P6 intron (2R8S)

b) - twist junction base pairs, thrombin RNA aptamer (3DD2)

Figure 8: Twisted cross-stack pairing across helix junctions. Stacks orange & blue, cross-stack H-bonds purple, protein green.

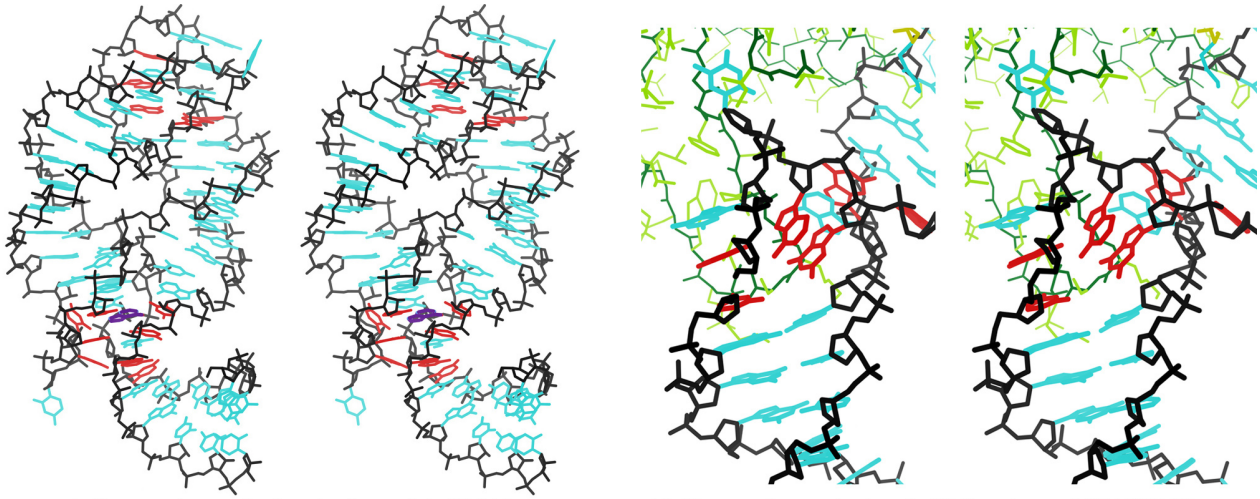


Figure 9: Context of strongly non-coplanar base-pair or non-stacked base clusters (red); effector purple, protein green.

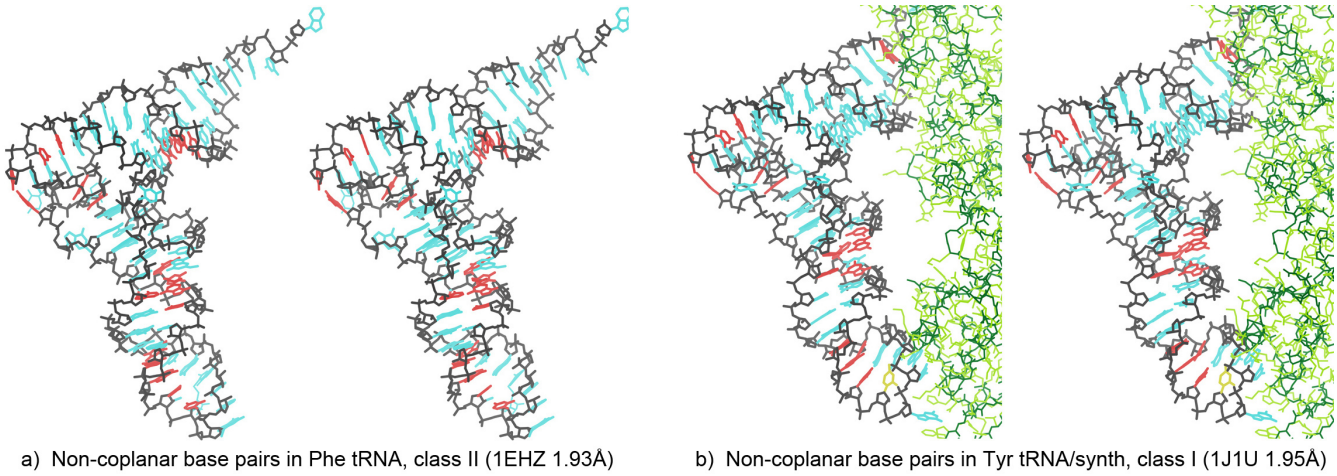


Figure 10: Conservation of 4 of the 5 clusters of non-coplanar base-pairs (red), between tRNA of the two distinct synthetase classes.

2013), to show their concentration at the functional binding sites.

Such clusters of highly non-coplanar or non-parallel bases are evolutionarily conserved, more so than the more obvious irregularities of flipped-out bases. As an example, figure 10 compares the clusters in tRNAs that are charged by tRNA synthetases from the two very different classes, Class I for the Tyr tRNA (1J1U; Kobayashi 2003) and Class II for the classic Phe tRNA (1EHZ; Shi 2000).

Conclusions

Very high-resolution structures show accurately the relationships of base-pair geometry and their response to overall context in the molecule, and those rules of thumb can be used at lower resolution to inform model-building and refinement protocols and parameters.

Base stacking is a strong interaction; when backbone conformation allows, stacking is nearly always parallel and at the 3.5Å separation optimal for van der Waals interaction. In contrast, base pairs are seldom completely coplanar, with each type of double-helix having its own characteristic value of base-pair propeller-twist, from -15° to +11°. In tertiary-structure interactions such as helix junctions, or in molecular

binding sites, base-pair twist or buckle can be quite large — it is controlled by the rather flexible base-pair H-bonds and the dominant effect of directionality of the stacking for each of the two bases. Those strongly non-coplanar base pairs tend to occur in places critical to forming the overall molecular structure or to providing its biological functionality.

Notes on measures and figures

The structures were displayed and examined in KiNG (Chen 2009). $2mF_{obs}$-dF_{calc} electron density maps (contours at 1.2σ) were downloaded from the Electron Density Server (Kleywegt 2004; eds.bmc.uu.se). All-atom contacts for hydrogen bonds (pillows of dark green dots) and van der Waals interactions (blue and green dots) were calculated on the MolProbity website (Chen 2010; molprobity.biochem.duke.edu). Base normals were constructed in Mage (Richardson 2001) and twist torsions and buckle angles were measured from them; propeller-twists were measured as dihedral angles around an axis between N1/9 atoms.

References

Aishima J, Gitti RK, Noah JE, Gan HH, Schlick T, Wolberger C (2002) A Hoogsteen base pair embedded in undistorted B-DNA, Nucleic Acids Res 30: 5244-5252 (1K61)

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

Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography, Acta Crystallogr D66: 12-21

Cheung YW, Kwok J, Law AWL, Watt RM, Kotaka M, Tanner JA (2013) Structural basis for discriminatory recognition of Plasmodium lactate dehydrogenase by a DNA aptamer, Proc Natl Acad Sci USA 110: 15967 (3ZH2)

Creighton T (2011) The Biophysical Chemistry of Nucleic Acids, Helvetian Press

Hamma T, Ferre-D'Amare A (2004) Structure of protein L7Ae bound to a K-turn derived from an archaeal box H/ACA sRNA at 1.8 A resolution, Structure 12: 893-903 (1SDS)

Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA (2004) The Uppsala Electron-Density Server, Acta Crystallogr D60: 2240-2249

Kobayashi T, Nureki O, Ishitani R, Yaremchuk A, Tukalo M, Cusack S, Sakamoto K, Yokoyama S (2003) Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion, Nature Struct Biol 10: 425-432 (1J1U)

Long, S.B., Long, M.B., White, R.R., Sullenger, B.A (2008) Crystal structure of an RNA aptamer bound to thrombin, RNA 14: 2504-2512 (3DD2)

Loughlin FE, Mansfield RE, Vaz PM, McGrath AP, Setiyaputra S, Gamsjaeger R, Chen ES, Morris BJ, Guss JM, Mackay JP (2009) The zinc fingers of the SR-like protein ZRANB2 are single-stranded RNA-binding domains that recognize 5' splice site-like sequences, Proc Natl Acad Sci USA 106: 5581-5586 (3G9Y)

Luo Z, Dauter M, Dauter Z (2014) Phosphates in the Z-DNA dodecamer are flexible, but their P-SAD signal is sufficient for structure solution, Acta Crystallogr D70: 1790-1800 (4OCB)

Olson WK, Bansal M, Burley SK, Dickerson RE, Gerstein M, Harvey SC, Heinemann U, Lu X-J, Neidle S, Shakked Z, Sklenar H, Suzuki M, Tung C-S, Westhof E, Wolberger C, Berman HM (2001) A standard reference frame for the description of nucleic-acid base-pair geometry, J Mol Biol 313:: 229-237

Richardson DC, Richardson JS (2001) MAGE, PROBE, and Kinemages, Chap. 25.2.8 in International Tables for Crystallography Vol. F (Eds. M.G. Rossmann and E. Arnold), pp 727-730. Kluwer Academic Publishers, The Netherlands, Dordrecht

Saenger W (1984) Principles of Nucleic Acid Structure, Springer-Verlag

Safaee N, Noronha AM, Rodionov D, Kozlov G, Wilds CJ, Sheldrick GM, Gehring KB (2013) Crystal Structure of the Parallel Double-Stranded Helix of Poly(A) RNA, Angew Chem Int Ed Engl DOI: 10.1002/anie.201303461 (4JRD)

Sheng J, Li L, Engelhart AE, Gan J, Wang J, Szostak JW (2014) Structural insights into the effects of 2'-5' linkages on the RNA duplex, Proc Natl Acad Sci USA 111: 3050-3055 (4MS9, the native 10-mer)

Shi H, Moore PB (2000) Crystal structure of yeast phenylalanine tRNA at 1.93 A resolution: a classic structure revisited, RNA 6: 1091-1105 (1EHZ)

Stoddard CD, Widmann J, Trausch JJ, Marcano-Velazquez JG, Knight R, Batey RT (2013) Nucleotides adjacent to the ligand-binding pocket are linked to activity tuning in the purine riboswitch, J Mol Biol 425: 1596-1611 (4FEJ)

Word JM, Lovell SC, LaBean TH, Zalis ME, Presley BK, Richardson JS, Richardson DC (1999) "Visualizing and Quantitating Molecular Goodness-of-Fit: Small-probe Contact Dots with Explicit Hydrogen Atoms", *J Mol Biol* **285**: 1711-1733

Ye JD, Tereshko V, Frederiksen JK, Koide A, Fellouse FA, Sidhu SS, Koide S, Kossiakoff AA, Piccirilli JA (2008) Synthetic antibodies for specific recognition and crystallization of structured RNA, Proc Natl Acad Sci USA 105:82-87 (2R8S)