Title

Interactive comparison and remediation of collections of macromolecular structures

Permalink
https://escholarship.org/uc/item/4wx6r55p

Journal
Protein Science, 27(1)

ISSN
0961-8368

Authors
Moriarty, Nigel W
Liebschner, Dorothee
Klei, Herbert E
et al.

Publication Date
2018

DOI
10.1002/pro.3296

Peer reviewed

TOOLS FOR PROTEIN SCIENCE

Interactive comparison and remediation of collections of macromolecular structures

Nigel W. Moriarty,[1]* Dorothee Liebschner,[1] Herbert E. Klei,[1]
Nathaniel Echols,[1] Pavel V. Afonine,[1] Jeffrey J. Headd,[1] Billy K. Poon,[1] and
Paul D. Adams[1,2]

[1]Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[2]Department of Bioengineering, University of California at Berkeley, Berkeley, CA, 94720, USA

Received 18 July 2017; Accepted 11 September 2017
DOI: 10.1002/pro.3296
Published online 13 September 2017 proteinscience.org

Abstract: Often similar structures need to be compared to reveal local differences throughout the entire model or between related copies within the model. Therefore, a program to compare multiple structures and enable correction any differences not supported by the density map was written within the *Phenix* framework (Adams et al., Acta Cryst 2010; D66:213–221). This program, called *Structure Comparison*, can also be used for structures with multiple copies of the same protein chain in the asymmetric unit, that is, as a result of non-crystallographic symmetry (NCS). *Structure Comparison* was designed to interface with *Coot*(Emsley et al., Acta Cryst 2010; D66:486–501) and *PyMOL*(DeLano, PyMOL 0.99; 2002) to facilitate comparison of large numbers of related structures. *Structure Comparison* analyzes collections of protein structures using several metrics, such as the rotamer conformation of equivalent residues, displays the results in tabular form and allows superimposed protein chains and density maps to be quickly inspected and edited (via the tools in *Coot*) for consistency, completeness and correctness.

Keywords: macromolecular crystallography; graphical user interface; validation; ligands

Introduction

It is desirable that atomic models of different crystal structures of a protein vary only in regions where genuine heterogeneity occurs (due to ligand binding, crystal packing, conformational changes, etc.), while excluding differences due to resolution, map quality or refinement methods. For models determined at low resolution, the internal consistency of related structures can be enforced during refinement by restraining the refinement model to a reference structure, which is implemented in various forms in phenix.refine[1] *Refmac*,[2,3] CNS,[4,5] and Buster-TNT.[6] Similar methods are also used to enforce non-crystallographic symmetry between multiple copies of a macromolecule in the crystallographic asymmetric unit.[7] However, none of these approaches provides a robust solution to the problems of distinguishing between real and artificial structural heterogeneity or of evaluating the likelihood that a validation outlier is an intrinsic feature of the macromolecule instead of an error.

Here, we describe the program "*Structure Comparison*" which addresses these issues by simultaneous validation and conformational analysis of multiple closely related structures. The goal is to make the entire collection of structures as correct, complete and consistent as possible. Inconsistencies should be confined to instances supported by experimental data, for example, density. We have used the *Phenix Structure*

*Correspondence to: Nigel W. Moriarty. E-mail: nwmoriarty@lbl.gov

Comparison program to analyze several sets of published structures and show that it facilitates rapid correction of spurious deviations between otherwise similar models.

Program Description

Implementation and general use

The program is written in Python with C++ extensions and is distributed as part of the *Phenix* software.[8] All user interactions are through a graphical user interface (GUI). Required inputs are a collection of two or more related protein chains. Separate files and NCS-related copies in a single file are permitted. Optionally, a sequence file is used to extract the chains of interest defaulting to the first chain found in the first model if omitted. The user may optionally supply diffraction data used for the calculation of $2mF_{\mathrm{obs}}$–DF_{model} and mF_{obs}–DF_{model} difference maps, or pre-calculated map coefficients in MTZ format.

A reference structure to use for global superposition of chains may be provided, defaulting to the first structure in the input list if omitted. If the reference structure has multiple chains, the first chain is selected for finding NCS copies in all input models, but a user provided chain ID will override this selection. Similarly, if a sequence file is provided, the index of the sequence in the file can be chosen. This is useful when a model contains more than one type of chain. The chosen sequence is then used to find all copies of that chain in all model files. The default for sequence identity between models or chains, 80%, can be adjusted, as can the number of processes used for performing calculations.

An important choice is whether to superpose the structures on a reference structure. If superposition is enabled, all similar chains (including NCS copies) and their density maps are superposed onto the reference. The results of the comparison are the same, but the visual display will show all chains superposed onto the reference structure. The superposition is performed using *phenix.superpose_pdbs* with the main chain C_α atoms selected for the matching algorithm. If map coefficients or diffraction data are provided, the resulting maps are superposed using *phenix.superpose_maps*.

Automatic file loading

Automatic file detection functionality is available in the *Structure Comparison* GUI. The “Add directory” option will search in the chosen directory for model and data/map file pairs. The “Add directory tree” option searches for file pairs by walking a directory tree and choosing a single pair from each directory. mmCIF format files have precedence over Protein Data Bank (PDB)[9] format files.

The program begins by extracting near-identical chains, based on comparison to the reference sequence, and performing optional superposition. Diffraction data, if supplied, are used to calculate maps. All tasks can be executed in parallel with the program typically taking only a few minutes to run on a multiprocessor computer. Depending on whether the structures are superposed or not, coordinate files for the individual chains can be saved. The results of the structure comparison are displayed in a new set of tabs within the program window, starting with a summary of the extracted chains and their basic properties. Additional tabs, described in section 2.4, include side-chain rotamer, Ramachandran, ligand, secondary structure, missing atoms, water cluster, cis-peptide, histidine protonation, and B-factors or atomic displacement parameters (ADP) analyses. Optionally, a HTML report can be created which contains tabular summaries of each of these analyses, the list of structures compared and their one-letter amino acid sequences.

Alignment and superposition of structures and maps

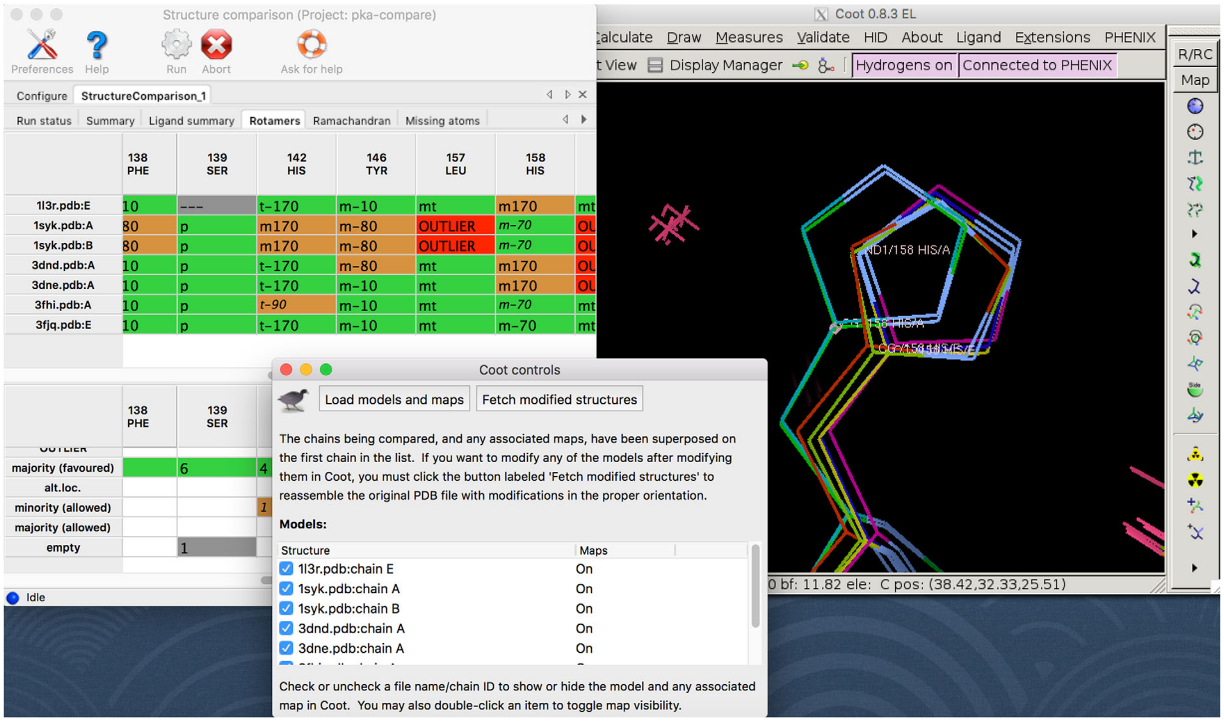
Although the program is designed to analyze structures with homogeneous residue numbering, it is also capable of handling variable insertions/deletions (and, potentially, close homologues). The default mode of operation assumes that the sequences are essentially identical, except for point mutations, and treats residue numbers as transferrable between structures. Where this is not applicable, the program may instead align the sequences of all chains under comparison to a common reference sequence using the open-source program MUSCLE.[10]

The individual chains may remain in their original positions or are superposed on to a reference model. This allows all chains to be viewed in a common frame of reference regardless of crystal form, or location in the asymmetric unit if NCS is present. Density maps, if calculated, will be simultaneously transformed to follow their associated models by interpolation onto a pseudo-P1 grid in the new orientation and saved in CCP4 map format. These routines are also available in the program *phenix.superpose_maps*.

The main limitation of the superposition procedure is the transformation of molecules away from their original context, such as crystal packing and ligands, although these effects are somewhat mitigated by the transformation of density with the models. The transformation matrices are automatically saved so that the original placement in the unit cell can be recovered using the *Coot* Control Window (see below). The models may therefore be edited in *Coot*[11] to improve agreement between equivalent residues and the full structures can be afterwards regenerated.

Extraction of features and presentation of results

The program is primarily designed to display all matching residues with heterogeneous properties;



	138 PHE	139 SER	142 HIS	146 TYR	157 LEU	158 HIS	
1l3r.pdb:E	10	---	t-170	m-10	mt	m170	mt
1syk.pdb:A	80	p	m170	m-80	OUTLIER	m-70	OU
1syk.pdb:B	80	p	m170	m-80	OUTLIER	m-70	OU
3dnd.pdb:A	10	p	t-170	m-80	mt	m170	OU
3dne.pdb:A	10	p	t-170	m-10	mt	m170	OU
3fhi.pdb:A	10	p	t-90	m-10	mt	m-70	mt
3fjq.pdb:E	10	p	t-170	m-10	mt	m-70	mt

	138 PHE	139 SER	
OUTLIER			
majority (favoured)		6	4
alt.loc.			
minority (allowed)			1
majority (allowed)			
empty		1	

The chains being compared, and any associated maps, have been superposed on the first chain in the list. If you want to modify any of the models after modifying them in Coot, you must click the button labeled 'Fetch modified structures' to reassemble the original PDB file with modifications in the proper orientation.

Models:

Structure	Maps
☑ 1l3r.pdb:chain E	On
☑ 1syk.pdb:chain A	On
☑ 1syk.pdb:chain B	On
☑ 3dnd.pdb:chain A	On
☑ 3dne.pdb:chain A	On

Check or uncheck a file name/chain ID to show or hide the model and any associated map in Coot. You may also double-click an item to toggle map visibility.

Figure 1. Typical Structure Comparison session showing the results tab for the rotamer angles and a *Coot* window displaying the models selected by the *Coot* Control Window (center bottom).

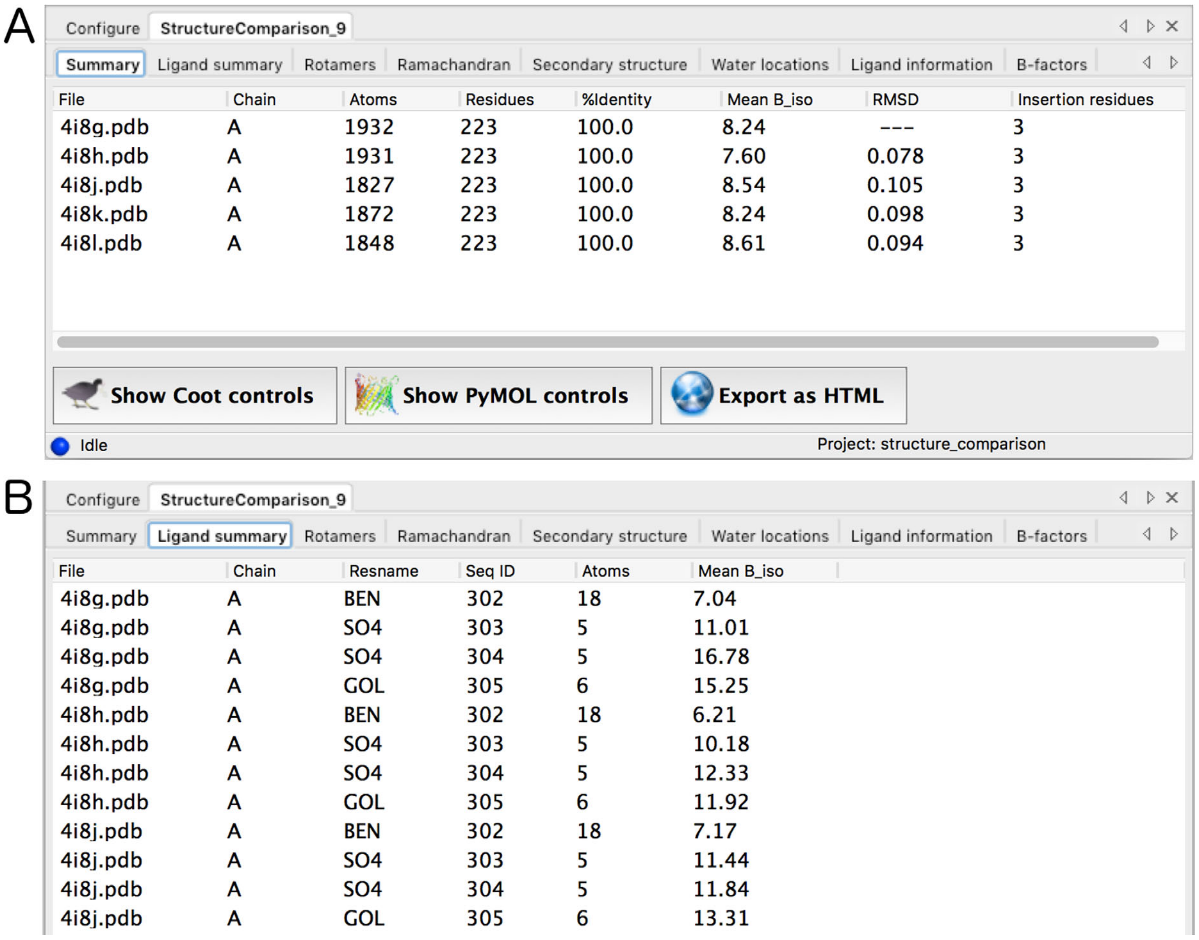
however, this may be overridden by a user-defined atom selection. Each of the reports discussed below is presented in a tab on the results page and can be viewed interactively within a *Coot* window. If no differences were found for a particular analysis, the corresponding result tab will not appear. For example, if there are no missing atoms or residues in the input models, the "missing atoms" tab will not be shown. Figure 1 shows a typical Structure Comparison session with the results tab for rotamers and a *Coot* window displaying the models selected by the *Coot* Control Window. In this example session, *Coot* is showing a portion of the models that were selected in the *Structure Comparison* results tab (note that density maps are not being displayed in this view).

For each type of analysis, the structure comparison results are presented and discussed using specific examples. These models were taken directly from the PDB and used without any refinement. The groupings of PDB codes are listed below. The PDB and ligand codes are written following the convention outlined in the editor's notes in the Computation Crystallography Newsletter.[12]

- Five high resolution (0.75 Å–0.87 Å) structures of bovine trypsin (BT) which resulted from a study investigating the reproducibility of protein models from different crystals obtained under similar conditions.[13]
- Two high resolution models of PfluDING at pH 4.5 (0.98 Å, PDB code 4F1U) and pH 8.5 (0.88 Å, PDB code 4F1V), obtained from two different crystals and refined independently.[14]
- A lower resolution model (PDB code: 1Z3Z, resolution 2.9 Å) and a higher resolution model (PDB code: 1ZoD, resolution 1.8 Å) of the protein dialkylglycine decarboxylase.[15]
- Four human factor Xa protein with different bound inhibitors: 3ENS (resolution: 2.3 Å, inhibitor: ENS[16]), 3HPT (resolution: 2.2 Å, inhibitor: YET[17]), 3K9X (resolution: 1.9 Å, inhibitor MBM[18]) and 3SW2 (resolution: 2.4 Å, inhibitor: Fi1[19]). The latter contains only one copy of the biological unit (two protein chains) while the other structures contain two copies.
- Several models of a human heart fatty acid-binding protein (H-FABP): two determined at high resolution (3WVM: 0.88 Å resolution, 4TJZ: 0.87 Å resolution) and one model determined via serial femtosecond crystallography (3WXQ).
- A model of cobalamin-dependent methionine synthase with two NCS related chains (3BoF) solved at 1.7 Å.[20]

A description of each of the possible results tabs follows. By default, the program runs all property analyses, but the user can limit those performed.

Summary. Each chain matching the reference occupies a row in the summary table, which lists the filename, chain ID, number of residues, and atoms, percentage identity to the reference and the mean isotropic Atomic Displacement Parameter of the chain (water molecules and ligands are not taken into account). If the chains were superposed, the



File	Chain	Atoms	Residues	%Identity	Mean B_iso	RMSD	Insertion residues
4i8g.pdb	A	1932	223	100.0	8.24	---	3
4i8h.pdb	A	1931	223	100.0	7.60	0.078	3
4i8j.pdb	A	1827	223	100.0	8.54	0.105	3
4i8k.pdb	A	1872	223	100.0	8.24	0.098	3
4i8l.pdb	A	1848	223	100.0	8.61	0.094	3

File	Chain	Resname	Seq ID	Atoms	Mean B_iso
4i8g.pdb	A	BEN	302	18	7.04
4i8g.pdb	A	SO4	303	5	11.01
4i8g.pdb	A	SO4	304	5	16.78
4i8g.pdb	A	GOL	305	6	15.25
4i8h.pdb	A	BEN	302	18	6.21
4i8h.pdb	A	SO4	303	5	10.18
4i8h.pdb	A	SO4	304	5	12.33
4i8h.pdb	A	GOL	305	6	11.92
4i8j.pdb	A	BEN	302	18	7.17
4i8j.pdb	A	SO4	303	5	11.44
4i8j.pdb	A	SO4	304	5	11.84
4i8j.pdb	A	GOL	305	6	13.31

Figure 2. Summary results tab (**A**) and ligand summary results tab (**B**) for five BT models.

root-mean-square deviation (RMSD) values of the C_α atoms are displayed as well.

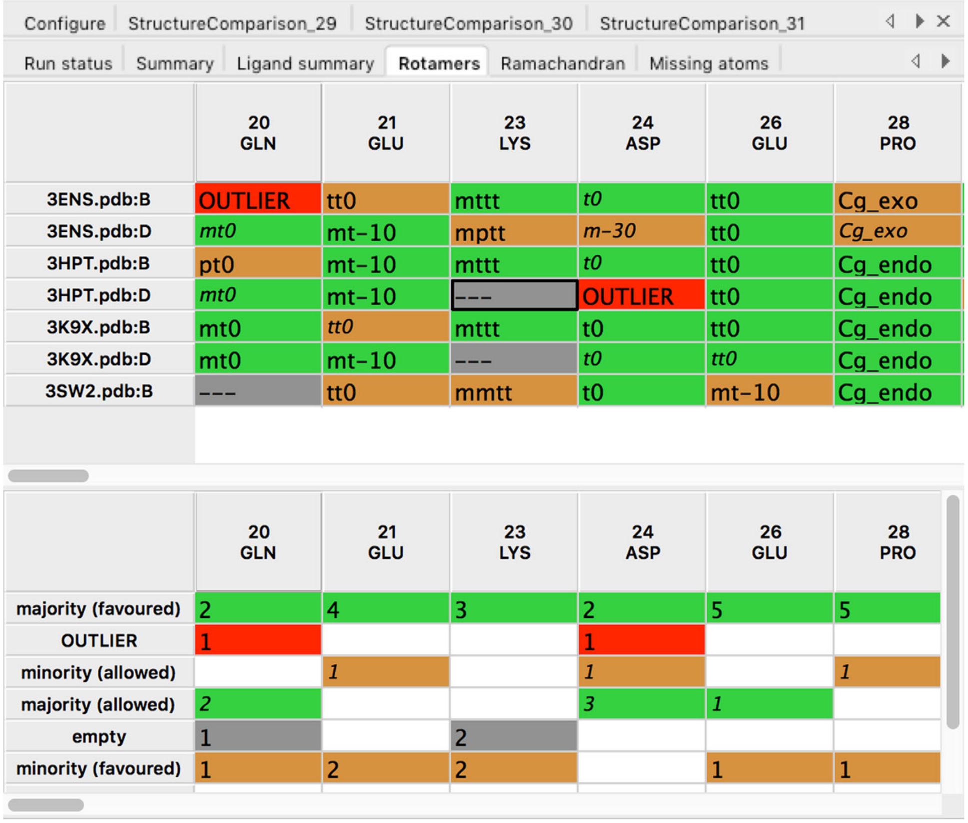
Figure 2(A) shows the summary results tab for five BT structures. The leftmost column lists the filenames of the models, followed by the chain identifier. The third and fourth columns show the number of atoms and residues, respectively in the protein chains. The number of atoms is similar for the five BT models (the only difference arises from the number of water molecules) and the number of residues is the same (223 in all cases). The fifth column indicates the sequence identity in percent, which is 100%, as no mutations or other modifications are present. The sixth column lists the mean ADPs, which are between 8.2 and 8.6 Å^2, except for model 4i8H, which has an even lower value of 7.6 Å^2. It can be noted that the diffraction data for 4i8H have also the highest resolution (0.75 Å). The seventh column summarizes the RMSD between model 4i8G and the other models (the value for 4i8G is therefore blank). The RMSD varies between 0.105 and 0.078 Å, indicating that the BT models superpose very well.

At the bottom of the results summary tab are the controls for *Coot*[11] and *PyMOL*.[21] Furthermore, the button "Export as HTML" enables saving all result tables in HTML format.

Ligand summary. If ligands are present, the ligand summary tab appears and lists basic information for each ligand, that is, chain ID, 3-letter code, residue number, number of atoms and mean isotropic ADP. Columns are sortable for ease of determining the most interesting comparisons.

Figure 2(B) shows the ligand summary tab for five models of BT, which each contain the inhibitor benzamidine (BEN), a glycerol molecule (GOL) and two sulfate ions.

Side-chain rotamers. Side-chains are compared to the Rotamer Library[22,23] using the program *phenix.rotalyze*. By default, any set of equivalent residues for which multiple rotamer classes are identified, plus any outliers, will be displayed in the results table. The color-coding of the cells allows quick identification of areas of interest. Outliers are colored red whilst cells without sufficient information, such as in the case of missing or incomplete residues, are shaded grey and otherwise left blank. If more than 50% of side-chains for a common residue share a single rotamer class, the



	20 GLN	21 GLU	23 LYS	24 ASP	26 GLU	28 PRO
3ENS.pdb:B	OUTLIER	tt0	mttt	*t0*	tt0	Cg_exo
3ENS.pdb:D	*mt0*	mt−10	mptt	*m−30*	tt0	*Cg_exo*
3HPT.pdb:B	pt0	mt−10	mttt	*t0*	tt0	Cg_endo
3HPT.pdb:D	*mt0*	mt−10	---	OUTLIER	tt0	Cg_endo
3K9X.pdb:B	mt0	*tt0*	mttt	t0	tt0	Cg_endo
3K9X.pdb:D	mt0	mt−10	---	*t0*	*tt0*	Cg_endo
3SW2.pdb:B	---	tt0	mmtt	t0	mt−10	Cg_endo

	20 GLN	21 GLU	23 LYS	24 ASP	26 GLU	28 PRO
majority (favoured)	2	4	3	2	5	5
OUTLIER	1			1		
minority (allowed)		*1*		*1*		*1*
majority (allowed)	*2*			*3*	*1*	
empty	1		2			
minority (favoured)	1	2	2		1	1

Figure 3. Rotamer results table for the factor Xa group of proteins.

cell is highlighted in green, while minority conformations are highlighted in orange. Side-chains adopting multiple conformations are shaded blue (with white font). A similar color code is used for most other result tabs (red for outliers, blue for multiple conformations, grey for missing residues etc.).

Recently, rotamer validation was extended to include the concept of favored and allowed rotamers.[23] The cutoff values are based on the probability of the set of side-chain dihedrals in a protein model. Less than 0.3% probability is considered an outlier while a probability of greater than 2% is considered favorable. The values of the side-chain dihedral angles determine the rotamer class, with rotamer IDs taken from Lovell *et al*.[22] The allowed designation for a rotamer indicates that the side-chain is on the edge of the specific rotamer space. These are shown with the use of italic font in the appropriate grid cell.

Therefore, there are seven possible categorizations of a residue in the rotamer tab. It can be the majority consensus, either favored or allowed; in the minority group, favored or allowed; alternative location; outlier; or not included in the evaluation.

Differences in side-chain rotamers can be a sensitive indicator of functionally relevant changes in the protein structure. They can also be due to poor density in mobile portions of the protein.

Figure 3 shows the rotamer table for the factor Xa proteins. The rotamer table consists of two sections. The upper section contains the data for each chain that has non-uniform results or outliers. Because the rotamer table is the most information rich, an overview table is included in the lower section to summarize the results and to act as a legend for the cell colors. The leftmost column lists the rotamer discrepancies for residue Gln20, which is present in chain B and D of each of the models with the exception of 3SW2 that only has one copy of the two protein chains in the biological unit. The side-chain of the latter is incomplete—truncated at C_γ as shown in Figure 4 (yellow)—and the cell is therefore highlighted in grey. In the 3ENS model, chain B, Gln20 is an outlier and highlighted in red. In the overview section, there is a row that contains the number of outliers (if present). A similar scheme follows for all the other six categories available in the side-chain rotamer tables. Looking at the seven overlaying instances of Gln20 in Figure 4 shows that this outlier (red) is very similar to the majority of the other rotamers.

The majority of the Gln20 rotamers are in the *mt0* conformation and the cells are therefore highlighted green. Note that two of the instances of *mt0* are written in italics which indicates that these two residues, 3ENS:D and 3HPT:D, are in the allowed region of the *mt0* rotamer space while the other two are in the favored region. This disparity is also shown in the overview table. The instance of Gln20 in 3HPT:B is in a different rotamer state (green), which

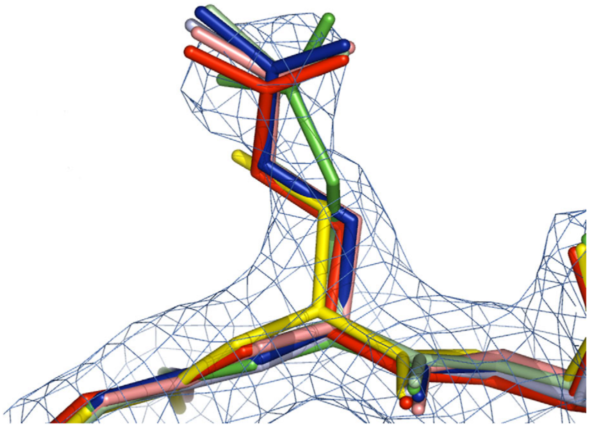


Figure 4. Gln20 in all models and chains (seven instances) of the factor Xa group of proteins as well as $2mF_{\text{obs}}–DF_{\text{model}}$ density map from 3HPT:B at 1.5σ contour level.

is supported by the density (Fig. 4). It is shown as the minority result.

Figure 5 (left) shows a close-up of the rotamer results table for Arg143, for which only one rotamer, 3K9X:D, differs from the other rotamers. Arg143 and the $mF_{\text{obs}}–DF_{\text{model}}$ and $2mF_{\text{obs}}–DF_{\text{model}}$ density maps from 3K9X:D are shown in Figure 5 (right). Clearly, the rotamer does not fit the density, as there is a negative peak at the N and C atoms of the guanidine group of the side-chain and a positive peak for the correct rotamer, *mtt180*, matching the other instances.

Models 3ENS, 3HPT, and 3K9X are crystallized in space group $P2_1$ while 3SW2 is in $P2_12_12_1$. The rotamer designation for Gln26 in all chains of the former is *tt0* while it is *mt-10* for the latter. An investigation of the crystal contacts shows that Glu26 in 3SW2 is within 2.9 Å of a symmetry related residue while in the other models the distance between Gln26 and any symmetry related residue exceeds 9 Å. Furthermore, the instance of Glu26 in 3SW2 is rotated towards the symmetry copy of Asn160 in chain A to form a hydrogen bond with the side chain nitrogen atom.

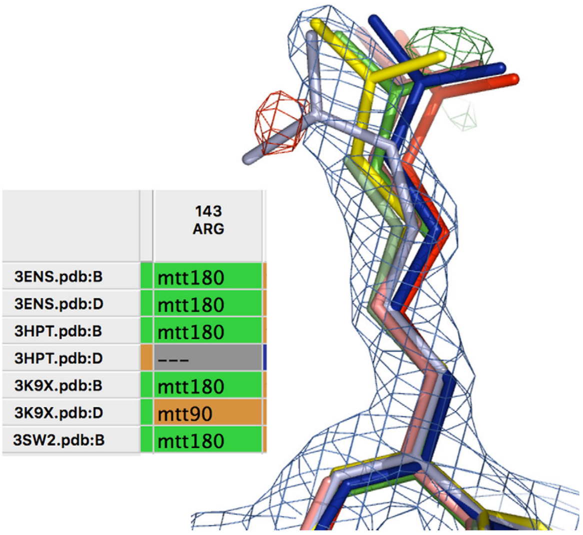
Gln30 exemplifies another interesting application of the rotamer analysis. All of the instances of residue Gln30 are in the *tp40* conformation except 3HPT:D, which is *tp-100*. These conformations are approximately 180 degrees apart and examination of the superposed side-chains shows that the conformations can be rationalized by an NQH flip[24] of the side chain oxygen and nitrogen atoms.

Ramachandran angles. Ramachandran angles are calculated using *phenix.ramalyze*, which uses the same structure database as the rotamer distributions.[25] The probability cutoffs for "favored" and "allowed" angles are identical to those used in Molprobity. By default, any set of equivalent residues for which multiple Ramachandran angles are identified, plus any outliers, will be displayed. The cells of the results table display the ϕ/ψ angles for each residue. Outliers are highlighted in red, while favored and allowed angles are shaded in green and orange, respectively. Alternative conformations are shaded in blue.

The Ramachandran angles describe the rotation of the protein chain. Differences can therefore indicate real changes in the local fold or possible fitting errors. Residues with different assignments in different models, that is, flagged as an outlier in one model while having a favored angle in another model, require special attention.

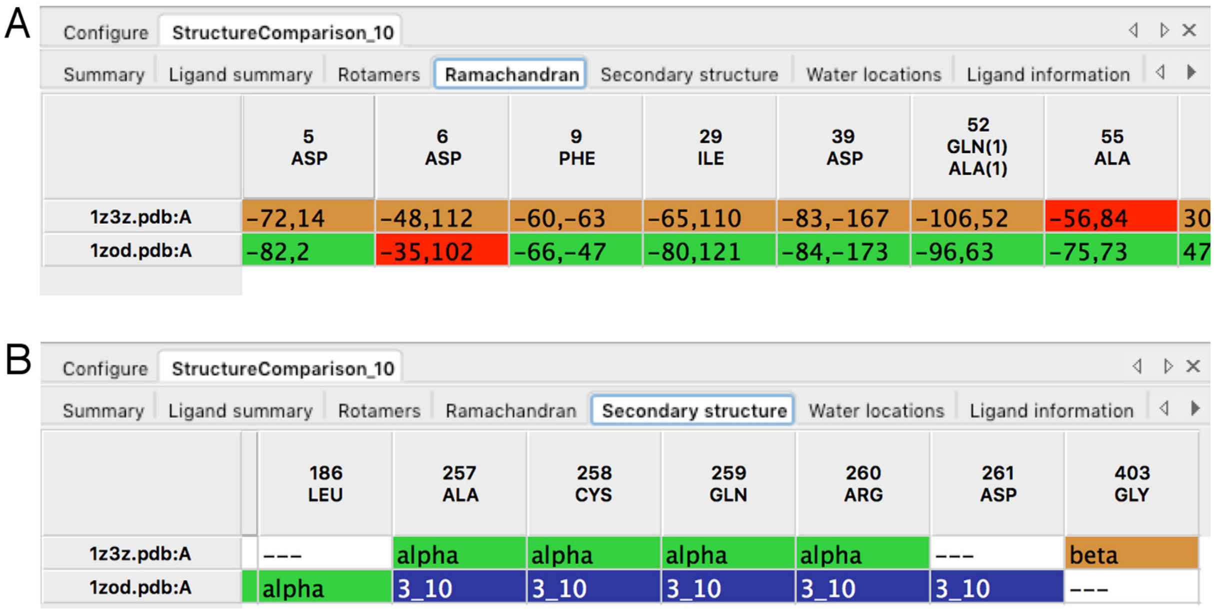
Figure 6(A) shows a part of the Ramachandran analysis results table for the two models of dialkylglycine decarboxylase. There are two outliers, Asp6 in model 1ZoD and Ala55 in model 1Z3Z. Model 1ZoD was determined at 1.8 Å resolution, which might justify the conformation of the Asp6 main-chain. However, model 1Z3Z was determined at 2.9 Å resolution, it is therefore less likely that the X-ray data definitively supports the outlier for residue Ala55. Also, the same residue adopts a favored conformation in model 1ZoD, which suggests that the Ala55 main-chain of model 1Z3Z should not be modeled as an outlier.

Secondary structure. Secondary structure annotation can be performed with different procedures, including CaBLAM,[26] which uses pseudo-torsion and other angles along the protein main chain to classify secondary structure, or ksDSSP (the default), an open-source implementation of the original DSSP



	143 ARG
3ENS.pdb:B	mtt180
3ENS.pdb:D	mtt180
3HPT.pdb:B	mtt180
3HPT.pdb:D	---
3K9X.pdb:B	mtt180
3K9X.pdb:D	mtt90
3SW2.pdb:B	mtt180

Figure 5. Left: Close-up of the rotamer results table for Arg143 in the factor Xa group of proteins. Right: Arg143 in all models and chains of the factor Xa group of proteins as well as $mF_{\text{obs}}–DF_{\text{model}}$ (positive: green, negative: red) and $2mF_{\text{obs}}–DF_{\text{model}}$ density maps from 3K9X:D at 1.5σ and ±3σ contour level, respectively. It can be noted that the side-chain is incomplete for 3HPT:D (lightgreen).



	5
	ASP
		6
		ASP
			9
			PHE
				29
				ILE
					39
					ASP
						52
						GLN(1)
						ALA(1)
							55
							ALA	

1z3z.pdb:A	-72,14	-48,112	-60,-63	-65,110	-83,-167	-106,52	-56,84	30
1zod.pdb:A	-82,2	-35,102	-66,-47	-80,121	-84,-173	-96,63	-75,73	47

	186 LEU	257 ALA	258 CYS	259 GLN	260 ARG	261 ASP	403 GLY
1z3z.pdb:A	---	alpha	alpha	alpha	alpha	---	beta
1zod.pdb:A	alpha	3_10	3_10	3_10	3_10	3_10	---

Figure 6. (**A**) Results table from the Ramachandran angle analysis and (**B**) Secondary structure results tab for models of 1ZoD and 1Z3Z of dialkylglycine decarboxylase.

algorithm.[27] Any set of matching residues with more than one type of secondary structure assignment, for example, a residue labeled as helix in one chain and without secondary structure assignment in another, will be reported in the results table. Color-coding is based on the secondary structure annotation: α-helices are green, 3_{10}-helices are blue, β-sheets are orange, non-secondary structure residues are white and missing residues are grey.

Differences in secondary structure annotation can highlight structural changes between models, such as those induced by ligand binding.[28,29] However, they can also occur if there are relatively small conformational differences between models, such as at the extremities of helices and strands or in short loops.[30]

Figure 6(B) shows an example for the secondary structure results tab after comparing a lower resolution model and a higher resolution model of the protein dialkylglycine decarboxylase. Different secondary structure annotations are found for 12 residues (His16, Thr172, Tyr173, Arg174, Tyr185, Leu186, Ala257, Cys258, Gln259, Arg260, Asp261, and Gly403). Five of them are at helix termini (His16, Tyr185 and Leu186, Asp261) or at the extremity of a beta-sheet (Gly403). Residues 172–174 in 1ZoD were apparently mis-assigned as 3_{10}-helix by the annotation algorithm. Visual inspection shows that the main chain adopts a loop in this region. Finally, residues 257–260 adopt an imperfect helical conformation, which leads to assignment of type alpha or 3_{10} in 1Z3Z and 1ZoD, respectively. The different assignments are most likely due to small conformational differences between 1Z3Z and 1ZoD in this area. This shows the necessity to visually inspect the results from secondary structure differences to identify areas of genuine variations.

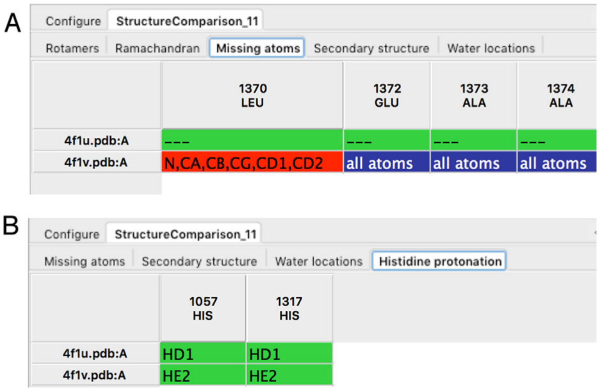
Omega angles. *Cis*-peptides and nonplanar omega angles are calculated using *phenix.omegalyze*. If the omega angle (ω) is within 30° of 180° or 0°, the peptide is deemed as a *trans*- or as a *cis*-peptide, respectively. All other ω are denoted as *twisted*. Residues with *cis* peptides or *twisted* ω are displayed using blue and orange, respectively.

The ω peptide angle is overwhelmingly *trans* in proteins. The occurrence of a *cis*-peptide is likely to be conserved between structures of the same protein. This is not the case for Glu559 of the PDB entry 3BoF, which is *twisted* in chain A and *trans* in chain B. Visual inspection of the residue in *Coot* shows that the density for this residue, the penultimate residue in each chain, is poor compared to the rest of the model.

Interestingly, the pair of *twisted* peptides, Ser412, have a large amount of difference density. Refining these residues to the *trans* conformation, a feature available in *phenix.refine* using *the apply_-cis_trans_specification* parameter, results in a better fit to density and a reduction in the overall angle rmsd from 1.01° to 0.92°.

Missing atoms. If atoms or entire residues are missing in one chain while present in at least one other chain, they will be listed in the results table. Residues missing side-chain or backbone atoms are shaded orange and red, respectively, and the names of the missing atoms are shown in the cell. If the entire residue is missing, the cell is colored blue while matching complete residues are highlighted in green.

Entire residues are sometimes not modeled when there is no clear density to support their location, such as at chain termini and in loops.



	1370 LEU	1372 GLU	1373 ALA	1374 ALA
4f1u.pdb:A	---	---	---	---
4f1v.pdb:A	N,CA,CB,CG,CD1,CD2	all atoms	all atoms	all atoms

	1057 HIS	1317 HIS
4f1u.pdb:A	HD1	HD1
4f1v.pdb:A	HE2	HE2

Figure 7. (**A**) Results from the missing atom analysis for two PfluDING models. The cell width can be expanded to show all missing atom names, such as in the case of Leu1370. (**B**) Results from the histidine protonation analysis for two PfluDING models.

Similarly, it may occur that side-chain atoms cannot be placed* which happens typically for residues with long side-chains located at the protein surface (for example, Lys or Glu). However, it is also possible that atoms are misplaced unintentionally during file conversions, manual inspection or switching between different programs. To address these issues, the "missing atoms" tab gives a concise summary of missing atoms and residues for all input models. It should be noted that the analysis is based on the input models or the input sequence. Furthermore, at present hydrogen atoms are omitted from the missing atom analysis.

Figure 7(A) shows the missing atoms results table for two models of the protein PfluDING. Residues Glu1372, Ala1373, and Ala1374 are missing in model 4F1V. They are located at the C-terminal and there is no clear electron density to support their location, which explains why they were not modeled. Furthermore, Leu1370 lacks the atoms N, CA, CB, CG, CD1, and CD2 of the B-conformation (i.e., the C, O, and HA atoms of conformation B are present). It can be noted that both of the neighboring residues of Leu1370 adopt an alternative conformation. It is likely that the missing atoms of the B conformation of Leu1370 were part of a double conformation, which was not entirely deleted (as the C, O, and HA atoms remained) during manual model building.

Histidine protonation. Histidine residues have three different protonation states: either one or both of the ND1 and NE2 nitrogen atoms of the imidazole can be protonated. The results table displays different protonation states of histidine, if hydrogen atoms are present in at least one of the input models. Hydrogen atoms are only seen experimentally at very high resolution (better than 1 Å) but it is possible to infer the position of the histidine hydrogen atoms from their parent heavy atoms. Hydrogen atoms are therefore routinely added to lower resolution models to improve refinement and the analysis of geometry clashes. However, when the electron density of the H atoms is not visible, it often cannot reliably be deduced from the analysis of suitable H-bond donors or acceptors in the vicinity. Automatic procedures might therefore produce inconsistent models.

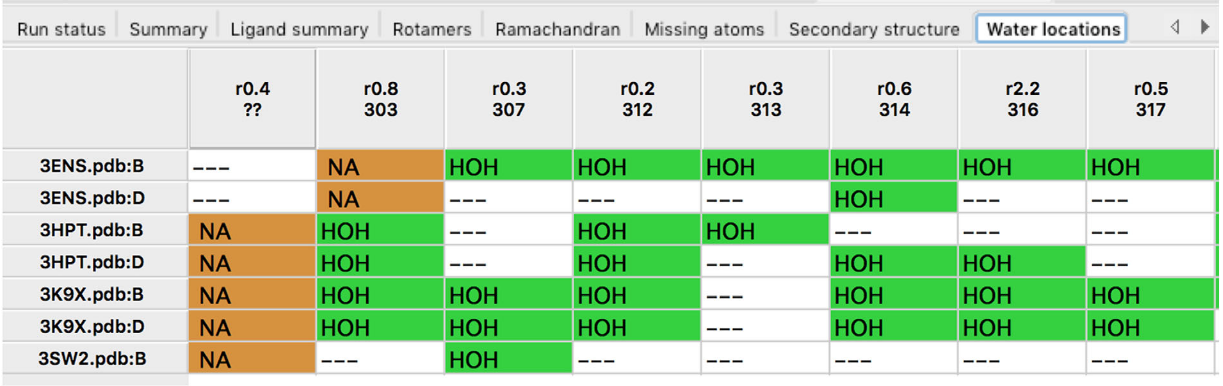
Figure 7(B) shows the results table for the comparison for two models of the protein PfluDING that contains two histidine residues. The comparison shows that the ND1 atoms of His57 and His317 are protonated in model 4F1U, while the NE2 atoms are protonated in model 4F1V. It seems likely that NE2 should be protonated in both structures, as there are unfavorable interactions with an asparagine NH2 head group if ND2 is protonated.

Water molecules. Differences in solvation site between models are of interest, as they may reflect genuine changes in structure or could be a result of misidentification of water and/or ion atoms.

In a model file, water molecules can be either associated with a macromolecular chain (this assignment is currently performed upon deposition of a model to the wwPDB[11]) or they may be placed in a separate chain (this is common practice in model building and refinement programs). The latter kind of model is analyzed using a simple procedure to assign water molecules to protein chains. Water molecules (and ions and ligands) are associated with a protein chain if they are located within 3.5 Å distance of one or more protein atoms. Once a molecule has been assigned to a chain, it cannot be moved to another protein chain. This procedure—if necessary—is carried out automatically by the *Structure Comparison* tool prior to the water cluster analysis (see Methods). The analysis results in a list of clusters containing up to N water molecules, where N is the number of superposed chains. Any cluster with less than N water molecules is displayed in the results table. Clusters with heterogeneous compositions of water and ions are particularly interesting and are reported at the beginning of the table. Cells representing a water molecule are highlighted in green and marked with the corresponding residue name used for water in the model (such as "HOH"), while the cell of the chain with no corresponding water is white and otherwise left blank. Cells containing ions are highlighted in orange.

Figure 8 shows the water results table for the factor Xa structures, with the sodium locations in the two left columns. Note that the column header of the second column shows a symbol, r0.8, that indicates the approximate radius of the cluster

*It should be noted that this approach is often debated: www.mail-archive.com/ccp4bb@jiscmail.ac.uk/msg20268.html



	r0.4 ??	r0.8 303	r0.3 307	r0.2 312	r0.3 313	r0.6 314	r2.2 316	r0.5 317
3ENS.pdb:B	---	NA	HOH	HOH	HOH	HOH	HOH	HOH
3ENS.pdb:D	---	NA	---	---	---	HOH	---	---
3HPT.pdb:B	NA	HOH	---	HOH	HOH	---	---	---
3HPT.pdb:D	NA	HOH	---	HOH	---	HOH	HOH	---
3K9X.pdb:B	NA	HOH	HOH	HOH	---	HOH	HOH	HOH
3K9X.pdb:D	NA	HOH	HOH	HOH	---	HOH	HOH	HOH
3SW2.pdb:B	NA	---	HOH	---	---	---	---	---

Figure 8. Water analysis results table for the factor Xa group of proteins. Note that the sodium locations are in the two left columns.

rounded up to the nearest tenth of an Ångström. The number 303 is the residue ID of the water molecule in the reference structure.

Figure 9 shows the sodium ions in all models of factor Xa group. Models 3HPT and 3K9X have the same arrangement of sodium ions and water molecules [see Fig. 9(A)]. Model 3ENS contains a sodium ion near the water locations of the first two models and there is a positive density peak near the sodium locations of models 3HPT and 3K9X [Fig. 9(B)], strongly suggesting that it actually occupies the center position and not the modeled one. In 3WS2, the sodium is at the same position as in the first two models but is missing the two water molecules for which positive density peaks appear in the difference map [Fig. 9(C)].

The results from the water analysis can also be used to investigate the possibility of a side-chain rotamer misfit into a water site or even the unnecessary use of alternative locations for a side-chain when the solvent model should in fact be updated.

Ligands. Many proteins are co-crystallized with ligands or molecules from the crystallization agent or buffer solutions. The development of new small molecule therapeutics often relies on the detailed analysis of ligand/macromolecule complexes and their associated densities. The differences between complexes can be particularly important for proposing changes to the small molecules to improve their efficacy or specificity. With the *Structure Comparison* tool, all ligands and small molecules can be localized to compare the molecule's position and orientation as well as any conformational changes in the protein.

For each ligand, the center of mass is calculated and a cluster analysis (similar to the method employed for water molecules, see Methods) is carried out. The results table lists all ligand clusters, making it possible to easily see which ligands are located at similar positions.

Figure 10 shows the four ligands (nine instances) from the factor Xa structures. The Fi1 ligand in 3WS2 has the largest difference in chemical structure from the others but has a convincing fit to its density. As with the water location table, a radius is provided in the table based on the center of mass of each of the ligands to give a measure of the superposition.

Atomic displacement parameters. Isotropic ADPs are extracted from each structure. The per-residue ADP averages for all atoms, main-chain atoms only and side-chain atoms are calculated. Because the ADPs for different structures may be on significantly different scales depending on resolution or crystal form, the ratio of the local average to the mean for the entire chain is also calculated, providing a normalized plot. The plot can be saved as an image and the values can be saved to a CSV file. The plot also displays points along the bottom corresponding to the residues appearing in the Ramachandran and rotamer tabs so that the relationship between high ADP values and validation outliers or modeling mismatches can be explored.

The ADP of an atom or a group of atoms represents, in part, small-scale static and dynamic disorder (as opposed to occupancies that represent the same but large-scale). Differences in ADP can be therefore indicative of different mobility, or local order in the structure. However, higher ADPs can also reflect model errors, as they typically increase when there is no strong density to support the modeled position.

Figure 11(A) shows the ADP profile (normalized by chain) for three models of heart fatty acid-binding protein (H-FABP). Overall, the profiles look very similar, but there are two residues where the ADPs differ significantly: Met1 in 3WXQ and Asp110 in 3WVM. Figure 11(B) shows a superposition of all three H-FABP models, as well as density maps for 3WVM in the vicinity of Asp110. The OD2 atom of the Asp110 side-chain is not optimally placed, as it is covered with negative difference density. The orientation of the Asp 110 side-chain is most likely similar to that in model 4TJZ (cyan), as

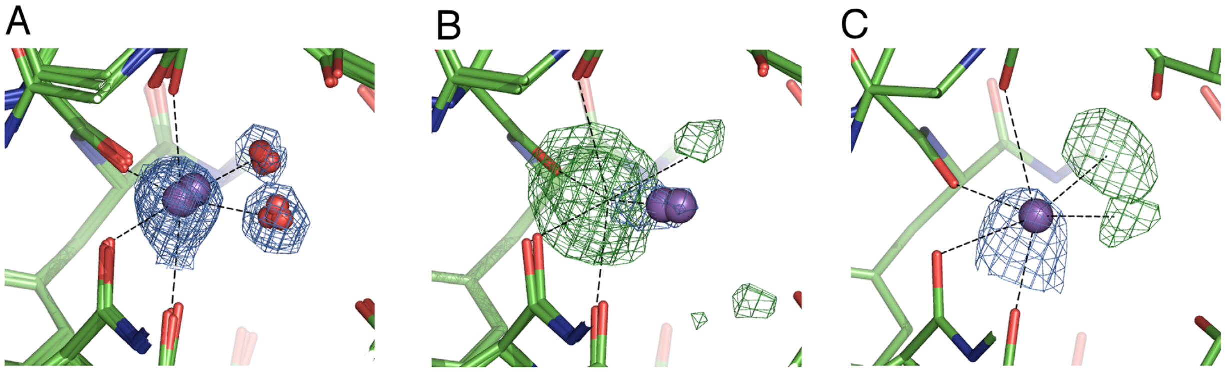


Figure 9. Sodium ions (violet spheres) in all structures of the factor Xa group; (**A**) Models 3HPT and 3K9X; (**B**) Model 3ENS; (**C**) Model 3WS2. The mF_{obs}–DF_{model} (green) and $2mF_{obs}$–DF_{model} electron density maps are contoured at 3σ and 2σ, respectively, except in the case of 3SW2:B, where the contours are 4σ and 1.5σ. Water molecules are displayed as red spheres.

suggested by a positive difference density peak and a $2mF_{obs}$–DF_{model} peak in the vicinity of the OD1 and OD2 atoms of Asp110 in 4TJZ. Furthermore, the side-chain rotamer is flagged as an outlier for Asp110 in 3WVM. Along with the density and the rotamer angle, the high ADP of this residue compared to the other models is highly suggestive of a sub-optimal orientation of Asp110. Similarly, inspection of the electron density maps near Met1 suggests that its orientation in 3WXQ can be optimized (not shown). Also, the Ramachandran angle of Met1 is in the outlier region, which further implies that its initial orientation can be improved. The ADP profile can therefore help to identify regions, which should be inspected.

Interaction with graphics programs

A key component of the *Phenix* GUI is the ability to control a *Coot* or *PyMOL* session through Python extensions to these programs.[31] After the initial calculations are complete, all models and maps are loaded into one or both viewers. Double-clicking on any non-empty cell in the result tables immediately centers the view(s) on the appropriate residue. It should be noted that clicking on an empty cell (often designated by "—") will not change the view.

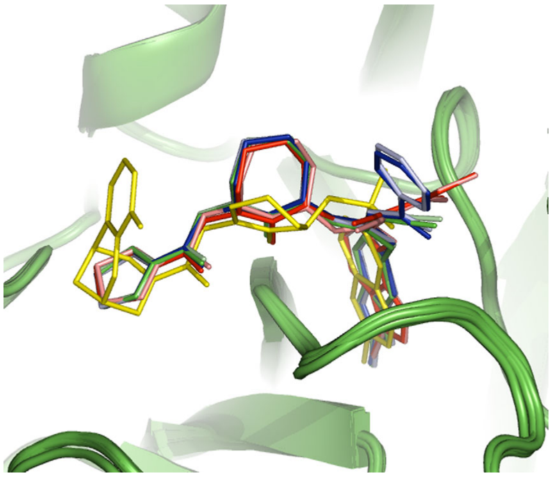


Figure 10. Superposition of the four ligands (nine instances) from the factor Xa group. Yellow: FI1 in 3SW2 chain B; Red: ENS in 3ENS chain B; Pink: ENS in 3ENS chain D; Dark green: YET in 3HPT chain B; Light green: YET in 3HPT chain D; Dark blue: MBM in 3K9X chain B; Light blue: MBM in 3K9X chain D.

Control panels within *Phenix* toggle the visibility of individual objects and enable additional actions such as recovery of edited models from *Coot* in their original orientations or display of all divergent residues for a specific analysis in *PyMOL*.

The *PyMOL* Control Window has an additional pull-down choice that will select and highlight only those residues listed in the corresponding tab. This can be very helpful in viewing only the ligands and producing publication ready images.

Methods

Water clustering

A clustering algorithm is applied to find regions that have localized water molecules. Distance criteria are used to assign the water molecules to clusters that represent a preserved water location. The clustering scheme is based on an hierarchical algorithm[32] with the addition of a termination criterion. The algorithm begins with each water molecule forming a cluster of one molecule. Looping over the list of distances between clusters (initially water molecules) from shortest to longest, it is determined if there is no water molecule from a different chain in the neighboring cluster, that is, the intersection of the sets of chain ID from each cluster is empty. If so, the two clusters are merged into a single cluster. This cluster will consist of, at most, a water molecule from each chain. This could be termed a trans-chain droplet.

If there is a water molecule from the same chain in a neighbouring cluster, both clusters are "frozen" so that no more water molecules can be added to either. The cluster cannot expand because there is a

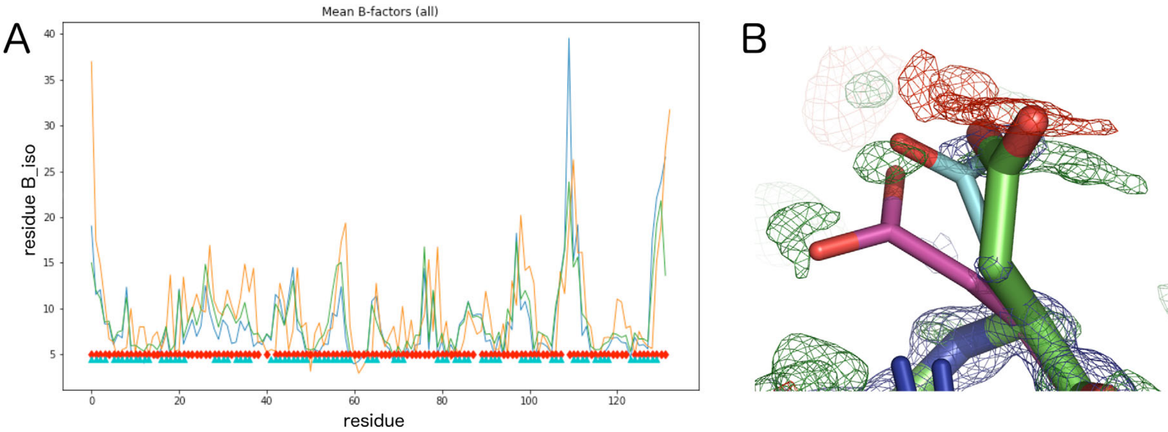


Figure 11. (**A**) ADP profile of three H-FABP models, as it is represented in the ADP results tab; 3WVM: blue, 3WXQ: orange and 4TJZ: green. The red diamonds and blue triangles designate residues with rotamer and Ramachandran differences, respectively. (**B**) Asp110 in three H-FABP models, 3WVM: color by atom, 3WXQ: violet and 4TJZ: cyan, and electron density maps for 3WVM. The positive (green) and negative (red) mF_{obs}–DF_{model} maps are contoured at ±2.5σ. The $2mF_{\mathrm{obs}}$–DF_{model} electron density is represented in blue at 1.5 σ contour level.

water molecule in the same real space that should belong to another cluster.

The result is a list of clusters containing between one and the number of superposed chains water molecules. If the cluster has a water molecule from each chain, it is preserved. However, if there are fewer water molecules in a cluster than the number of chains, the non-uniformity requires attention.

The clustering algorithm is not unique to water molecules. Ions can be included in the analysis and the resulting clusters may therefore contain both water molecules and ions (but only either a water molecule or ion from each chain). Depending on the circumstances, mixed ion/water clusters can represent either badly placed ions, badly placed water molecules or a physically meaningful change in the environment between models.

Because the clustering algorithm is based on a single point in space, it can be also applied to small molecules and ligands by using the center of mass of each entity.

Map superposition

In the *Structure Comparison* program, the models are superposed using the *phenix.superpose_pdbs* tool. The rotation and translation operators obtained by least squares fitting of equivalent coordinates are saved and used in the map superpose step. An orthogonal box around the reference molecule is determined and both the molecule and corresponding map are shifted into that box. The rotation-translation operations are applied to the map corresponding to the moving model. Both reference and moving maps are defined on a regular grid. However, applying the rotation and translation to the moving map does not guarantee an exact superposition of grid nodes so tricubic interpolation is used to calculate the map values on the transformed map grid. The command line utility to superpose models and maps is *phenix.superpose_maps*.

Conclusion

The comparison of similar models or NCS copies of chains can provide valuable information to improve model building and perform validation. Differences between models require in many cases some deeper consideration. They can highlight genuine heterogeneity originating from different crystal contexts (such as the rotamer change of Glu26 in 3SW2 to form a hydrogen-bond with a symmetry-related residue due to crystal packing) or conformational changes, which happen, that is, upon ligand binding.

On the other hand, differences can be also due to errors in model building (such as Arg143 in the factor Xa group that has a consensus for rotamer *mtt180* and also the density to support it), or highlight areas that cannot be clearly modeled, that is, when the experimental data does not provide any clear evidence for a particular conformation.[33] It can be noted that there is often more than one indicator of a problem such as in the example discussing the ADPs, where problematic residues are also flagged as rotamer or Ramachandran outliers.

With the advent of cryo-EM structures, it is envisage that this tool will be expanded to cover the comparison of structures and maps directly. Also, the neutron refinement structures can be included with the development of a deuterium aware missing atoms module.

In several examples used here, refining with a modern refinement program can correct a number of the problems but there are often differences that need further attention.

Download instructions: *Structure Comparison* is available from the *Phenix* release version 1.12.

A tutorial video, explaining how to use *Structure Comparison* and discussing results using example data, is available on the *Phenix* Tutorials YouTube Channel (www.youtube.com/c/phenixtutorials).

Acknowledgments

This research was supported by the NIH (grant GM063210 and its ARRA supplement) and the *Phenix* Industrial Consortium. This work was partially supported by the US Department of Energy under Contract DE-AC02-05CH11231. The algorithms described here are available in the *Phenix* software suite (http://www.phenix-online.org). The complete suite is freely available to academic users. Many of the individual components, such as the routines for map transformation, are incorporated into the open-source CCTBX library (http://cctbx.github.io).

We are grateful to Robert Edgar and the UCSF Computer Graphics Laboratory for the open-source programs MUSCLE and ksDSSP, respectively and for advice on their use. National Institutes of Health, GM063210;U.S. Department of Energy, DE-AC02-05CH11231.

References

1. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012) Towards automated crystallographic structure refinement with phenix.refine. Acta Cryst D68:352–367.
2. Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA (2011) REFMAC5 for the refinement of macromolecular crystal structures. Acta Cryst D67:355–367.
3. Nicholls RA, Long F, Murshudov GN (2012) Low-resolution refinement tools in REFMAC5. Acta Cryst D68: 404–417.
4. Schröder GF, Levitt M, Brunger AT (2010) Super-resolution biomolecular crystallography with low-resolution data. Nature 464:1218–U146.
5. Schröder GF, Brunger AT, Levitt M (2007) Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. Structure 15:1630–1641.
6. Smart OS, Womack TO, Flensburg C, Keller P, Paciorek W, Sharff A, Vonrhein C, Bricogne G (2012) Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. Acta Cryst D68:368–380.
7. Hendrickson W (1985) Stereochemically restrained refinement of macromolecular structures. Methods Enzymol 115:252–270.
8. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Cryst D66:213–221.
9. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242.
10. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.
11. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Cryst D66:486–501.
12. Computational Crystallography Newsletter (2015) Comput Crystallogr Newsl 6:26–53.
13. Liebschner D, Dauter M, Brzuszkiewicz A, Dauter Z (2013) On the reproducibility of protein crystal structures: five atomic resolution structures of trypsin. Acta Cryst D69:1447–1462.
14. Liebschner D, Elias M, Moniot S, Fournier B, Scott K, Jelsch C, Guillot B, Lecomte C, Chabrière E (2009) Elucidation of the phosphate binding mode of DING proteins revealed by subangstrom X-ray crystallography. J Am Chem Soc 131:7879–7886.
15. Fogle EJ, Liu W, Woon S-T, Keller JW, Toney MD (2005) Role of Q52 in catalysis of decarboxylation and transamination in dialkylglycine decarboxylase. Biochemistry 44:16392–16404.
16. Shi Y, Sitkoff D, Zhang J, Klei HE, Kish K, Liu EC, Hartl KS, Seiler SM, Chang M, Huang C, Youssef S, Steinbacher TE, Schumacher WA, Grazier N, Pudzianowski A, Apedo A, Discenza L, Yanchunas J, Stein PD, Atwal KS (2008) Design, structure-activity relationships, X-ray crystal structure, and energetic contributions of a critical P1 pharmacophore: 3-chloroindole-7-yl-based factor Xa inhibitors. J Med Chem 51:7541–7551.
17. Shi Y, Zhang J, Shi M, O'Connor SP, Bisaha SN, Li C, Sitkoff D, Pudzianowski AT, Chong S, Klei HE, Kish K, Yanchunas J Jr, Liu EC, Hartl KS, Seiler SM, Steinbacher TE, Schumacher WA, Atwal KS, Stein PD (2009) Cyanoguanidine-based lactam derivatives as a novel class of orally bioavailable factor Xa inhibitors. Bioorg Med Chem Lett 19:4034–4041.
18. Shi Y, Li C, O'Connor SP, Zhang J, Shi M, Bisaha SN, Wang Y, Sitkoff D, Pudzianowski AT, Huang C, Klei HE, Kish K, Yanchunas J Jr, Liu EC, Hartl KS, Seiler SM, Steinbacher TE, Schumacher WA, Atwal KS, Stein PD (2009) Aroylguanidine-based factor Xa inhibitors: the discovery of BMS-344577. Bioorg Med Chem Lett 19:6882–6889.
19. Shi Y, O'Connor SP, Sitkoff D, Zhang J, Shi M, Bisaha SN, Wang Y, Li C, Ruan Z, Lawrence RM, Klei HE, Kish K, Liu EC, Seiler SM, Schweizer L, Steinbacher TE, Schumacher WA, Robl JA, Macor JE, Atwal KS, Stein PD (2011) Arylsulfonamidopiperidone derivatives as a novel class of factor Xa inhibitors. Bioorg Med Chem Lett 21:7516–7521.
20. Koutmos M, Pejchal R, Bomer TM, Matthews RG, Smith JL, Ludwig ML (2008) Metal active site elasticity linked to activation of homocysteine in methionine synthases. Proc Natl Acad Sci USA 105:3286–3291.
21. DeLano WL (2002) PyMOL 0.99.
22. Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. Proteins 40: 389–408.
23. Hintze BJ, Lewis SM, Richardson JS, Richardson DC (2016) Molprobity's ultimate rotamer-library distributions for model validation. Proteins 84:1177–1189.
24. Headd JJ, Immormino RM, Keedy DA, Emsley P, Richardson DC, Richardson JS (2009) Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place. J Struct Funct Genomics 10:83–93.

25. Lovell SC, Davis IW, Adrendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. Proteins 50:437–450.
26. Williams CJ (2015) Using C-alpha geometry to describe protein secondary structure and motifs.
27. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637.
28. Sawai H, Yamanaka M, Sugimoto H, Shiro Y, Aono S (2012) Structural basis for the transcriptional regulation of heme homeostasis in *Lactococcus lactis*. J Biol Chem 287:30755–30768.
29. Amemiya T, Koike R, Fuchigami S, Ikeguchi M, Kidera A (2011) Classification and annotation of the relationship between protein structural change and ligand binding. J Mol Biol 408:568–584.
30. Martin J, Letellier G, Marin A, Taly J-F, de Brevern AG, Gibrat J-F (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 5:17.
31. Echols N, Grosse-Kunstleve RW, Afonine PV, Bunkoczi G, Chen VB, Headd JJ, McCoy AJ, Moriarty NW, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Adams PD (2012) Graphical tools for macromolecular crystallography in PHENIX. J Appl Cryst 45: 581–586.
32. Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32:241–254.
33. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Adams PD, Moriarty NW, Zwart P, Read RJ, Turk D, Hung L-W (2007) Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models. Acta Cryst D63:597–610.