

SUMMER STUDENT REPORT

DESY HAMBURG 2012



Parameters for the NCS-detection and extension in ARP/wARP that improved all models in a preselected set of proteins

Author : Vladimir Lazarenko

Supervisor : Victor Lamzin

6.08.2012

Contents

- **Introduction**

1. General method
2. ARP/wARP

- **Method**

1. NCS
2. Find parameters

- **Results**

- **References**

- **Discussion**

- **Acknowledgements**

Introduction

The Identification of the function of macromolecules based on knowledge of the spatial (tertiary and quaternary) structure is one of the main objectives in the field of protein crystallography. Nowadays, the most accurate model of the spatial structure can be deduced by X-ray crystallography, which is based on the analysis on the diffraction obtained by exposing a crystal of the macromolecule to X-rays.

X-ray crystallography

1) Crystallization

To apply the method of X-ray diffraction one has to obtain a crystal of the considered protein. The process of protein crystallization should be free of possible contaminants. It uses different types of chromatography. Before crystallization can be attempted the protein has to be purified. One problem is, that different proteins crystallize in different conditions. The factors that vary most between different crystallization conditions are the temperature, ionic strength, alcohol content, polyethylene glycol, and pH value (these are also called precipitants). The crystallization itself can be achieved by different methods based on diffusion through the vapor phase or through special membranes and gels. In the first case, a drop of the protein solution is suspended (hanging) over the solution excluding the protein. The water in the protein drop will vaporize allowing an increase of the amount of precipitant. Since the system is in equilibrium, the crystal will grow as soon as the right ratio of water and precipitant emerged. In the second case, the protein is in a capillary placed in a test tube with the solution, in this case, precipitant will diffuse will through a gel or membrane that covers the capillary. However, due to many different factors and problems, it is not possible to crystallize every protein.

2) Data collection

Due to the large size of the protein molecules, they crystallize in large unit cells (about 100 Å), and, since a large number of reflections is required to achieve high resolution these cells repeat thousands of times. Hence, the larger the protein, the more exposure of its crystals to X-ray beams is required to obtain the necessary amount of reflections. However, a long exposure leads to radiation degradation. The use of low

temperature (about 100 K) provides a reduction of radiation damage and increased data accuracy. However, due to freezing of the protein crystal ice may form, which will alter the diffraction pattern. To circumvent this problem one uses 'shock' freezing (rapid cooling to very low temperatures). The same solvent is added to the special agents to prevent formation of ice crystals. Protein crystals are not perfect and often have a mosaic structure. As a consequence, the Bragg reflection does not occur at a certain angle, but some angular range. Thus, the crystal must be slightly rotated, so that the diffraction of all elements in the crystal can be recorded.

3) The phase problem

By collecting diffraction patterns one only obtains the intensity of the structure factors (the final result is a table with three indices representing each reflection and its intensity value). To obtain the distribution of the electron density of the macromolecule, the phase of the diffracted X-ray is also required. However, it cannot be directly measured. This is the so-called phase problem, which is addressed by a variety of mathematical techniques. The primary methods of solving the phase problem are isomorphous replacement, anomalous scattering or molecular replacement.

4) Refinement of the structure

In the usual case one can use the aid of molecular graphics programs, such as Coot, and structure refinement programs, such as Refmac [1]. In Coot, the user can manually build a model of the protein described by the electron density map, and in accordance with the difference electron density maps, can replace certain parts of the structure. Next to the new coordinates are recalculated Fourier coefficients and the construction of a new electron density map. The easier way to do this, is using an automated model building procedure (which combines model building and refinement), such as ARP/wARP [2].

In this work, my task was to improve a module of the ARP/wARP model building, which uses the automatic detection of non-crystallographic symmetry (NCS) to improve models at medium-to-low resolution. In the following I will give a brief overview on ARP/wARP, NCS and the module under consideration.

ARP/wARP

ARP/wARP is a software suite to build macromolecular models into X-ray crystallography electron density maps. The main idea of this software is – **the model consists only of what is found in the electron density map**. The main features of the ARP/wARP software suite are described in the next section.

Free atoms and hybrid models. A crystallographic electron density map is always sampled to a regular grid. Essentially, the main part of model building is to converse the information of the electron density map to a crystallographic molecular model, made by atoms with known chemical identity. As a first step of building a model, ARP/wARP converses the map information to a set of ‘free atoms’ that have no chemical identity. These free atoms are placed at every position in the electron density map, where the density is high enough to support an atom. In addition to that, they are placed in a way that retains a protein-like shape. As model building and refinement proceed, some free atoms gain chemical identity (they are identified as part of a protein chain), but some atoms will remain free. This mixture is called the ARP/wARP hybrid model, which combines two sources of information: it incorporates chemical knowledge from the partially built protein model, whereas its free atoms continue to interpret the electron density in areas where no model is yet available. Finally, the atomic positions of the free atoms are used as guides for building the protein main chain into the electron density maps. This also allows implementing computationally more efficient algorithms.[3]

Main chain. Main chain tracing in ARP/wARP uses all available atoms of the hybrid model as potential C^α atoms. Peptides between potential C^α pairs are recognized by matching the electron density that surrounds each potential C^α pair to the one precomputed from true C^α pairs from known structures. The recognized peptides are subsequently assembled into linear polypeptide chain fragments using a limited depth-first graph-search algorithm, where the main chain is built up from overlapping sets of four C^α fragments that are selected to match conformations observed in the Protein Data Bank (PDB). Chain

fragments including partial ‘guessed’ side chains are refined to fit the electron density using the steepest descent algorithm.

Side chains. The protein chains are subsequently docked in sequence with side chains built in the best rotamer configuration and refined in real space using an implementation of the downhill simplex algorithm. This allows the torsion angles of each side chain to be gradually changed in a stepwise manner, so as to fit atoms to the electron density, while keeping the side-chain bonded atom distances and angles intact.

Loop building. After sequence docking, the missing parts of the model can be easily identified. Using this knowledge and a distribution of five C^α fragments that have been derived from known structures, many structurally likely conformations are constructed and the ones that fit best the electron density are chosen. Incorporating previous information allows building in low-density regions.

Secondary structure recognition. At a resolution of 3.0 Å and lower, where electron density maps lack atomic features, ARP/ wARP uses a different algorithm to build protein helices and strands. Sparse map grid points with about 1 Å spacing are selected as potential C^α atoms on the basis of their density. They are then fed into a complex scheme of successive filtering steps that yield fragments of appropriate helical or stranded conformations. These are used to generate candidate trace ensembles that then undergo averaging. Finally, peptide backbone and C_β atoms are added, the secondary structural chain fragments are subject to real space refinement and the most likely chain direction is selected. The procedure has been designed to work at resolution down to 4.5 Å.

Ligand building

When the protein structure is (nearly) completed, smaller compounds—ligands, cofactors—bound to the protein are modeled in the difference electron density map. First, regions of difference density that have approximately the same volume as the ligand are identified. Subsequently, numeric features of the density region and its sparse representation (similar to the one used to build free atoms for chain tracing) are used to produce an ensemble of putative ligand structures to

best fit the local density. The single best model is chosen after restrained (steepest descent) real-space refinement of all candidates in the ensemble.

Cocktail screening. This technique compares the shapes of difference electron density blobs with the shapes of compounds from a list (cocktail) of ligand candidates. The ligand that fits best is selected for further construction of the ensemble and subsequent restrained refinement.

Solvent building

After the protein part of the model is complete, either manually or using automated software, a solvent structure can be constructed in a difference electron density map. The protein part of the model is not rebuilt. Apart from van der Waals repulsion, no restraints are applied to the refinement of solvent even if the protein part is highly restrained. Therefore, ordered solvent comprises on average about 10% of the model; improvement of solvent indirectly improves the density corresponding to the protein part. The output is the protein model with the solvent molecules transformed with symmetry operations to lie around the protein.

Iterations

Building protein chains or solvent with ARP/wARP proceeds in an iterative fashion. When the quality of the (partially built) model is sufficiently high, the phases improve overall and result in an enhanced electron density where a more accurate and more complete model may be built. In essence, ARP/wARP, like human crystallographers, links model building and refinement together into a unified process that iteratively proceeds toward the final macromolecular model. An important component within iterations is the model update. Parts of the existing model located in weak density can be removed and new atoms added where the density acquired pronounced features.

A flowchart of ARP/wARP is shown in the Figure 1.

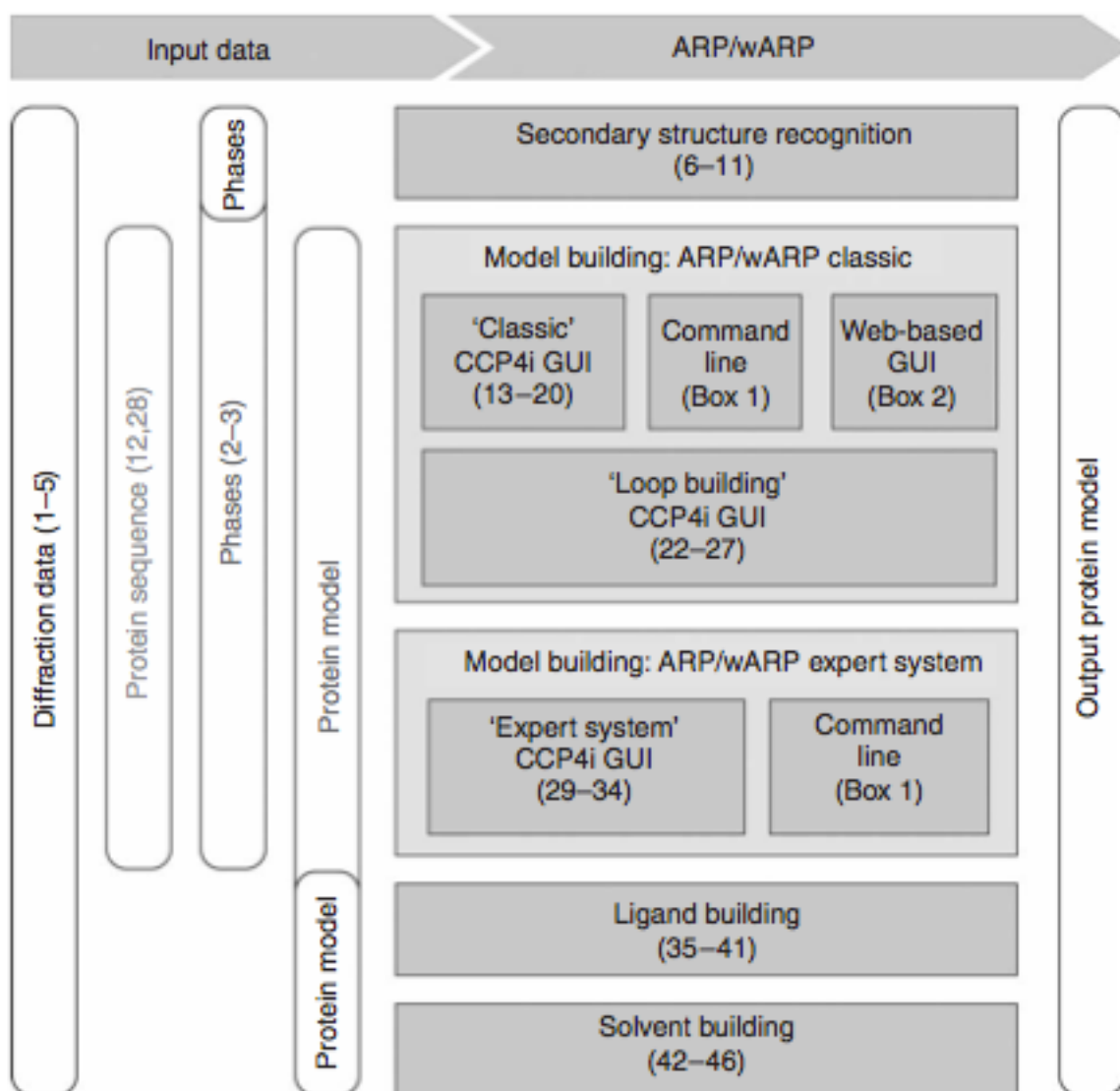


Figure 1 | A flowchart of the ARP/wARP procedure. The arrow on top indicates the flow of the data. ARP/wARP modules are labeled in the middle in gray-shaded boxes; the numbers in parentheses refer to the steps in Procedure that describe them. The rounded rectangular boxes to the left represent input data (black for required data, light gray for optional input—the sequence—and medium gray for alternative input—the phases or a model) and those to the right represent the output data. The vertical span of the input/output boxes refers to the procedures they are connected to in the middle[2]

Methods

NCS

It is a curious fact that many proteins prefer to form crystals with multiple copies in the asymmetric unit. A recent statistical survey found that this happens in about one-third of all crystals [4]. More than 50% of the structures in the current release of the PDB contain this additional structural information, which can be used as an additional (intrinsic) information. This peculiarity is called noncrystallographic symmetry (NCS). The NCS order may be as high as 60; this results in 70% of all structural fragments in the PDB being involved in an NCS relation. Two types of NCS can be distinguished (Figure 2).

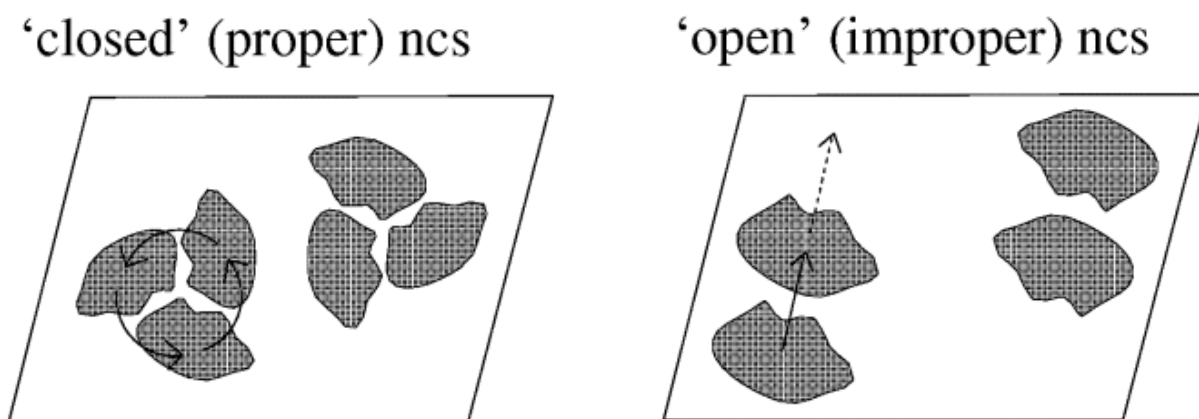


Figure 2. Types of noncrystallographic symmetry.

An element which is independent in the sense of rotation is defined as 'proper'. An example would be a molecule exhibiting an N-fold axis, with each element rotated by $(360/N)$ degrees to the next one. 'Improper' NCS is referred to in the case of arbitrary rotation between two molecules in the same asymmetric unit.

The use of NCS has been an extremely valuable asset in crystallographic structure determination. Perhaps its most frequent application is in density modification, in which NCS averaging helps to improve and extend phases to higher resolution as well as to reduce bias in cases where initial maps have been derived from an incomplete model. The stereochemical information from the regions of the protein chain that have been defined as NCS-related is added to the prior probability distribution in order to be used together with the observed structure factors in refinement.

The recently released versions of the ARP/wARP software suite contain a module, which uses NCS to improve the model. This module is called PNSextender (Protein NCS-based Structure Extender) [5].

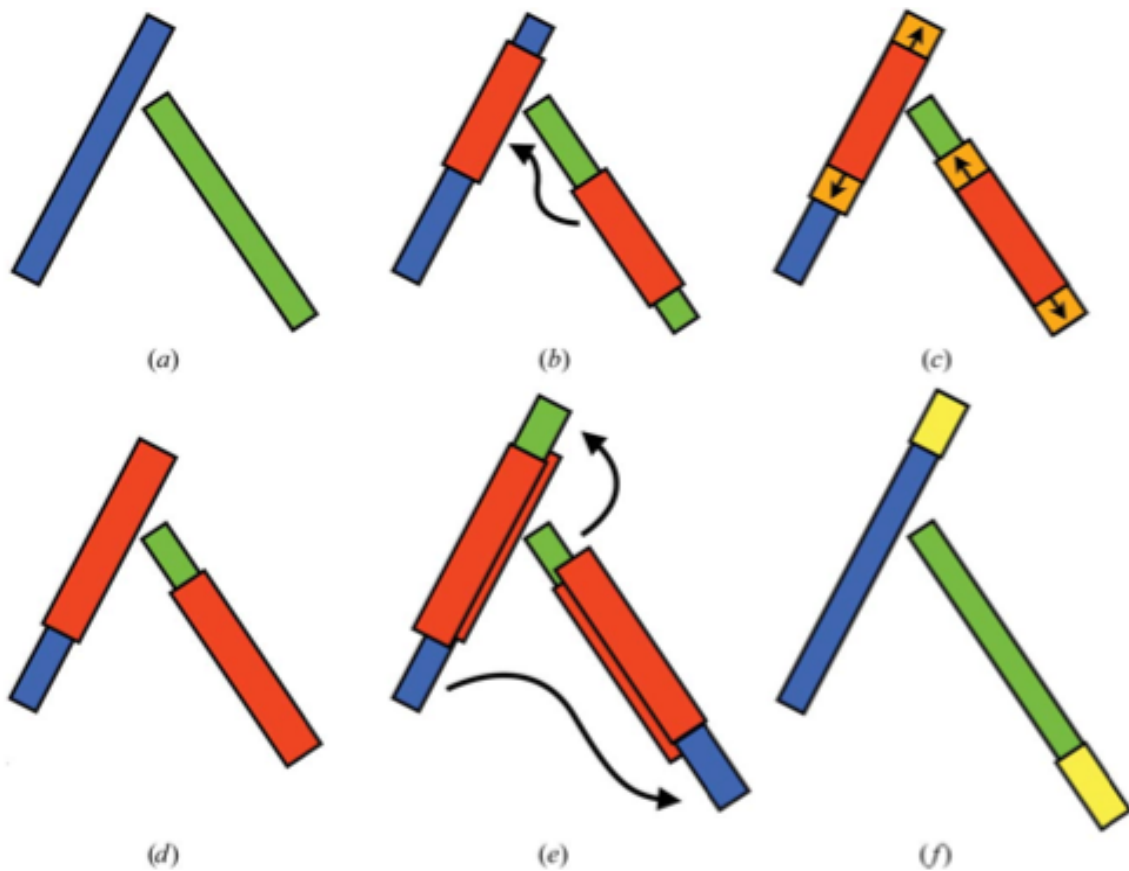


Figure 3. Workflow of the Protein NCS-based Structure Extender (PNS Extender). Intermediate partial models are examined for symmetric dependencies between stretches of two fragments (a). An initial match is found between green and blue regions (b, red blocks). The initial match is extended in both directions of the chain fragments (c, orange blocks). Once the extension is finished and the r.m.s.d. between the extended matches (red blocks in d) is still below the acceptance threshold, each extension (e, overlaid blocks) is NCS-transformed, as shown by arrows, onto the other fragment. Finally, longer extended green and blue fragments are obtained (f) and their extended parts (f, yellow blocks) are input as guides for protein-chain tracing.[5]

The first step of the PNSextender (Figs. 3a and 3b) involves an analysis of the partially built protein-chain fragments for possible symmetry-related dependencies. Each stretch of a fixed number of C^α atoms of each chain fragment is least-squares superposed with each stretch of the same length of every other fragment. To find the longest continuous region of the NCS match between two fragments, we adjust each initial overlapping stretch (as shown in Fig. 3b) by extending the matching region in both directions along the chain (Fig. 3c). During the extension we recompute the r.m.s.d. over the increased length, L_{ext} . Should the r.m.s.d. exceed a predefined threshold of $0.2 L_{\text{ext}} \text{ \AA}$, the

inspected NCS match is not considered further. This helps to reduce false positives by avoiding arbitrary or unlikely matches. Once extension is complete, the remaining ‘tails’ (Fig. 3d, blue and green ‘leftover’ tubes) are considered on both sides of the overlap region. All C^α atoms from the tails of each fragment are NCS-transformed to the end part of the corresponding fragment (Fig. 3e). Should there be stereo-chemical clashes (defined as two atoms being at a distance of less than 0.7 Å from each other) between an NCS-transformed atom and any other atom from existing protein-chain fragments, the former is deleted.

Find parameters

All proteins are different, but for ARP/wARP we need to find a set of parameters, which delivers the best result for all possible proteins. The other option would be to test each protein with many different combinations of parameters, which will result in long computations and is thus undesirable. The changeable parameters are described in the following: r.m.s.d (the r.m.s.d superposed fragments have to have so that they are deemed to be NCS related), number of found extensions that are fed back into the ARP/wARP model building process, and the length of fragments than are taken for automatic ncs-detection. My main work was to find the best combination. In this work the range of for these parameters were the following, for the r.m.s.d 0,4Å – 0,8Å, for the length of fragments 1000 – 3000 (where 1000 and 3000 are flags to set the length to either the average length of all partially built chain fragments or the 50th percentile) and 1 to 4 found extensions to be fed back into the ARP/wARP model building process. This lead to 32 possible sets of parameters (4 x 2 x 4), testing all the sets for one protein case took about one day. To deduce the resulting values for each case (number of chains, number of residues and R-factor) from log-file, I have written a Perl script. The results are presented in the next part.

Results

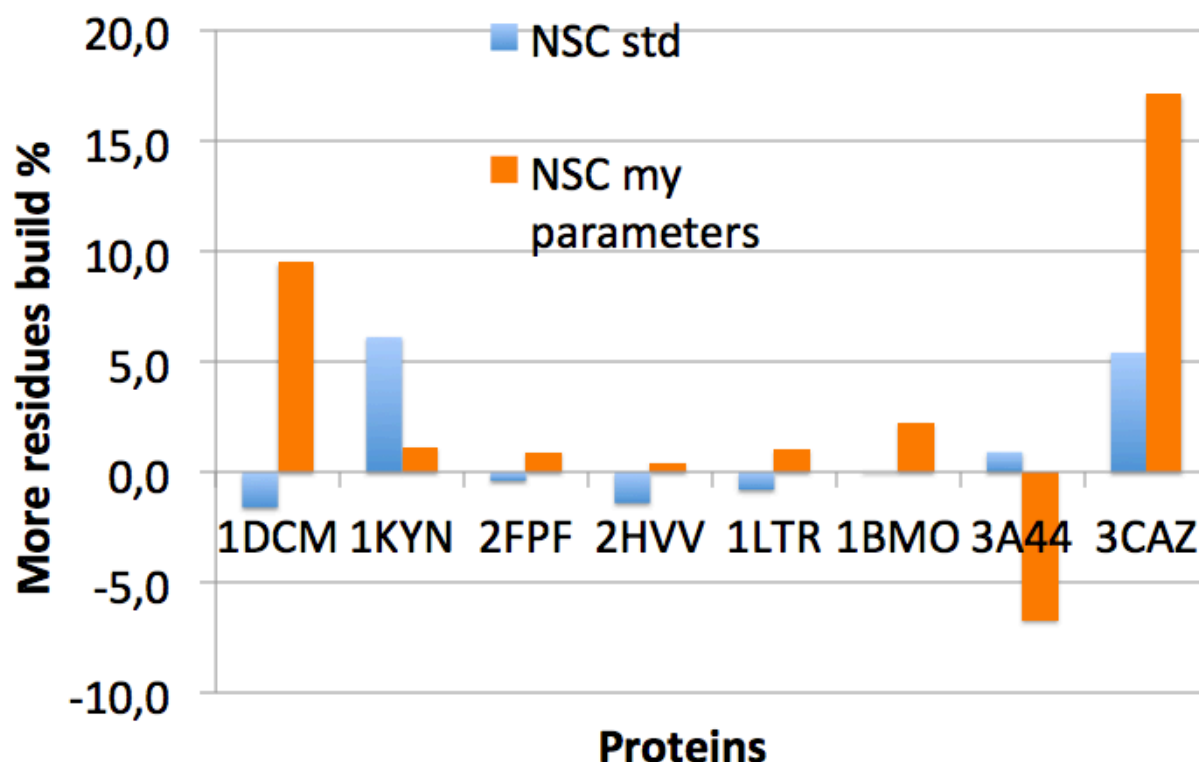
The test set comprised eight proteins from that had been selected from the PDB, which have NCS and low resolution (2,6 – 3,35 Å). On these proteins I have tested the performance of ARP/wARP, firstly standard protocol and after that, using the one incorporating the PNSextender. To understand which parameters are the best one for all test cases, I have taken a closer look at the number of chains, residues and the R-factor. In more detail, the number of residues was divided by the number of chains, the result was then again divided by the R-factor. For each set of parameters, these values were summed up and the best value was chosen as defining the best set of parameters. The best parameters were an r.m.s.d. threshold of 0.7 Å, the flag for the length of fragments 1000 (using the 50th percentile of the lengths of all partially built protein chains) and number of found extensions that are fed back into the arp/warp model building process (2). Statistics are presented in Table 1.

	1DCM	1KYN	2FPF	2HVV	1LTR	1BMO	3A44	3CAZ
More residues built (PNSext result compared to ARP)	9,52	-4,32	0,87	0,36	1	2,20	-6,67	17,12
Model completeness of PNSext result	54,76	56,60	81,34	76,09	89,73	79,61	37,77	44,22
%difference of main in residues/chains ARP	64,29	1,14	-4,66	53,85	37,72	8,08	-4,07	24,01
Improvement or R-Factor	-0,0037	-0,0068	-0,0077	0,0252	-0,0042	-0,006	0,02	0,012
Model completeness of ARP result	50	59,15	80,63	75,82	88,85	77,9	40,47	37,76
Improvement in Model completeness (PNSext vs. ARP)	4,76	-2,55	0,70	0,27	0,88	1,72	-2,7	6,46

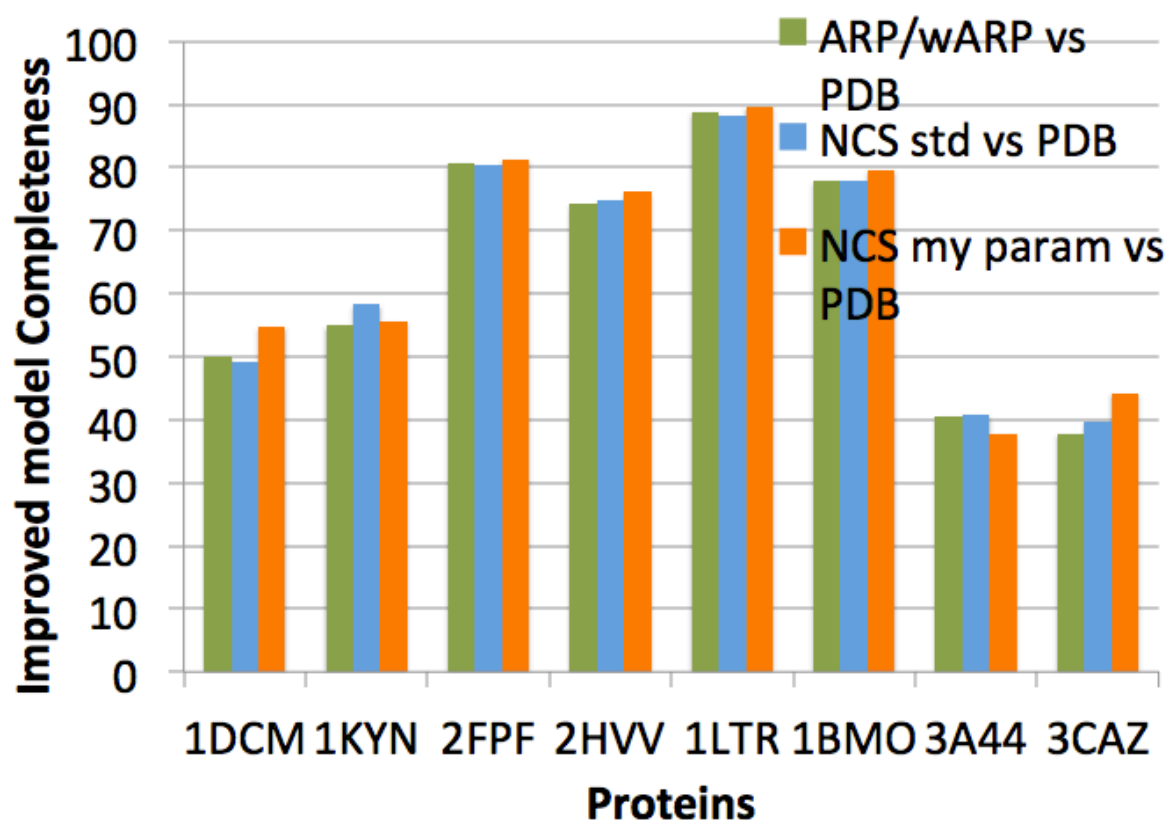
Table 1. Statistics of the selected parameters (best parameters).

The following plots show the comparisons between the results obtained by the standard ARP/wARP protocol (coined on the plots as ‘ARP’), ARP/wARP using the PNSextender with standard parameters (on the plots ‘NCS std’) and ARP/wARP using the PNSextender with the ‘best’ parameters I have found during this project (on the plots ‘NCS

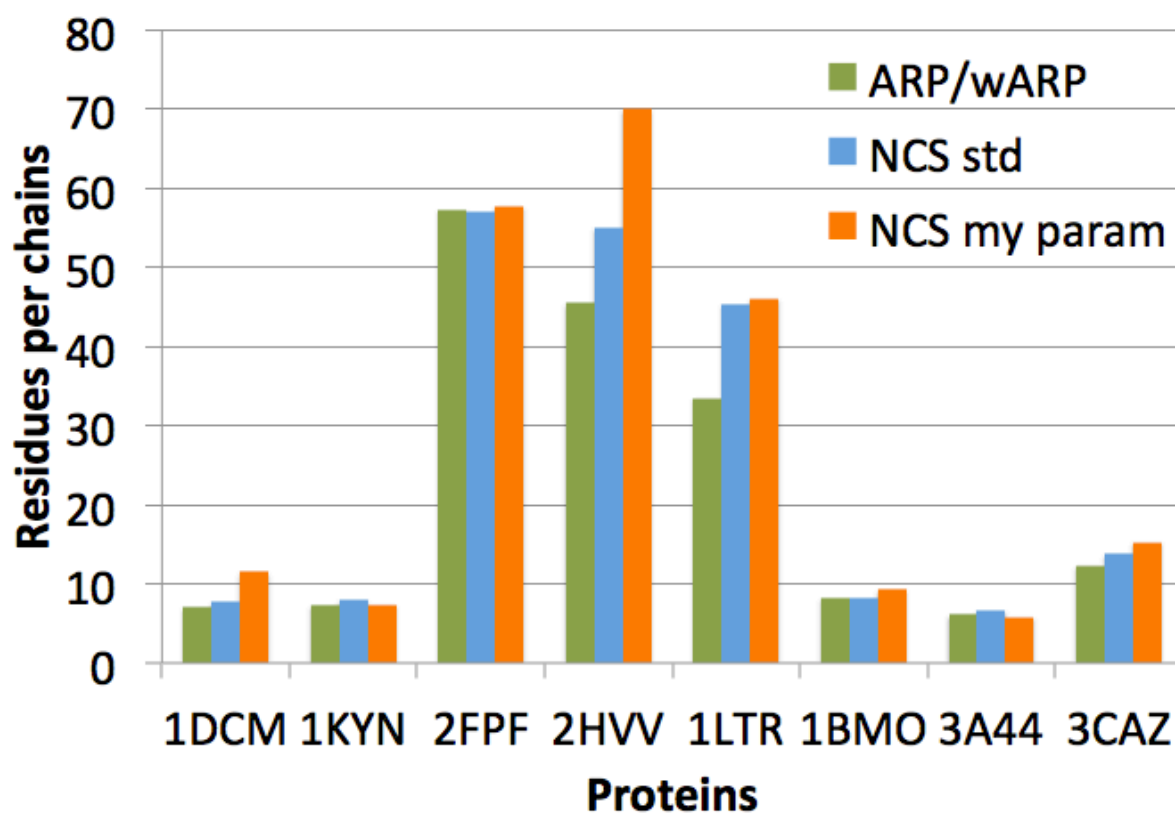
mine') (Plot №1 – Plot №4).



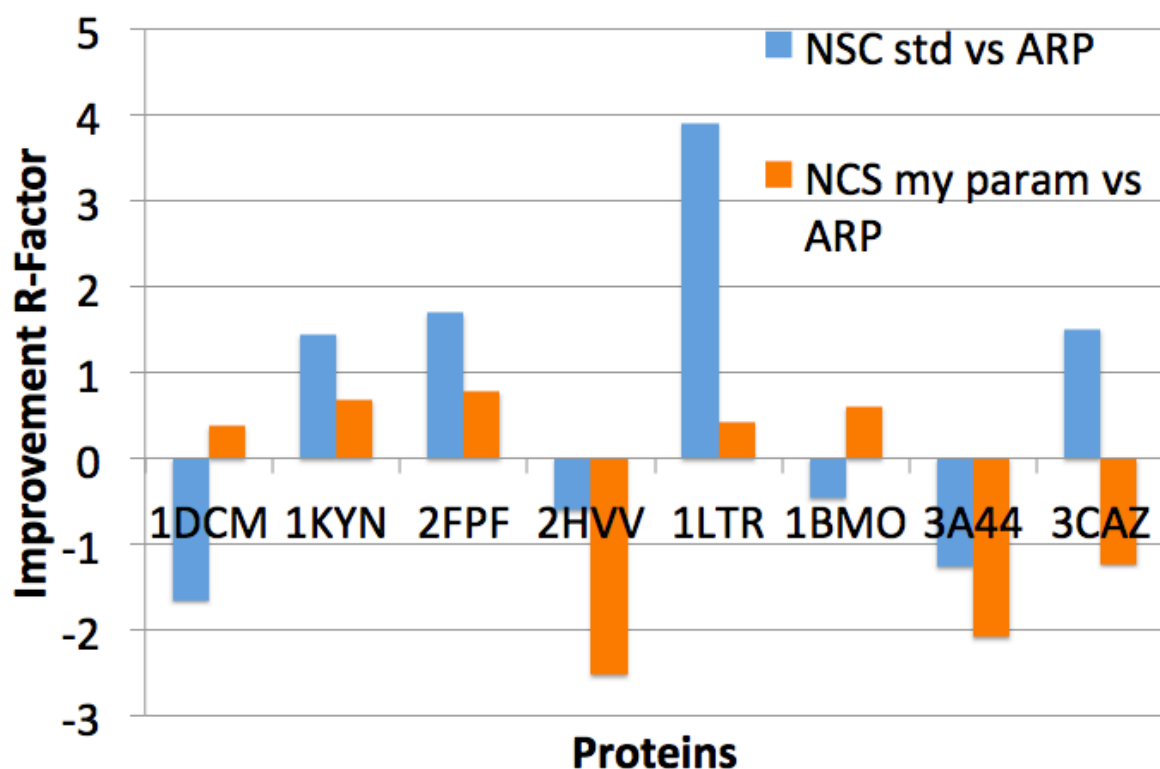
Plot №1. Difference in number of residues after work of NSC std and NCS mine.



Plot №2. Improvement in model in comparison with PDB.



Plot №3. Different in value of residues divided per chains in ARP, NCS std and NCS mine.



Plot №4. Improvement of R-factor in comparison NCS std with ARP and NCS mine with ARP.

Discussion

As is can be see in the plots, ARP/wARP using the PNSextender with the best parameters I have found generally work better. For some protein test cases, the model completeness decreases slightly. However this can be explained with a strong focus on improvement in R-factors. Also, after an analysis of these plots we can say, that the new parameters (NCSmine) give a better overall improvement than the previous parameters. Obviously, these results are not the final best ones. More testing on a broader dataset and really all possible parameters is needed for that (at least 100 structures). However, this would take a large amount of time (and cluster-access), which was not feasible for the short duration of my project. I hope, that continuing such kind of research will give to us the optimal parameters!

References

- [1] *REFMAC5* for the refinement of macromolecular crystal structures / G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long and A. A. Vagin / *Acta Cryst.* (2011). D67, 355-367
- [2] Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7 / Gerrit Langer, Serge X Cohen, Victor S Lamzin & Anastassis Perrakis / NATURE PROTOCOLS | VOL.3 NO.7 | 2008 | 1171
- [3] ARP/wARP and Automatic Interpretation of Protein Electron Density Maps / Richard J. Morris, Anastassis Perrakis, and Victor S. Lamzin / METHODS IN ENZYMOLOGY, VOL. 374
- [4] Not your average density / Gerard J Kleyweg and Randy J Read / © Current Biology Ltd ISSN 0969-2126
- [5] Use of noncrystallographic symmetry for automated model building at medium to low resolution / Tim Wiegels and Victor S. Lamzin / *Acta Cryst.* (2012). D68, 446–453

Acknowledgements

I would like to say a thank you to all the people who made this summer school possible for me, all people from DESY for their kindness to all summer students, all the members of the Lamzin group at the EMBL and all other people in EMBL for their willingness to help me in my training and work, especially my supervisor Victor Lamzin, Tim Wiegels, for his help and patience, and Johanna Kallio for her help at the beamline and in wet lab.