Mathematical Biology and Bioinformatics. 2012. V. 7. №2. P. t16-t24.
URL: http://www.matbio.org/2012/Sobolev_7_t16.pdf
Translation from original Russian text © Sobolev O.V., Lunin V.Y., 2012 published in Matematicheskaya biologiya i bioinformatika, 2012, V. 7. №2. P.692-702. URL:http://www.matbio.org/2012/Sobolev_7_692.pdf

UDC: 548.737

The use of Refmac crystallographic refinement program for the detection of alternative conformations in biological macromolecules

Sobolev O.V.*, Lunin V.Y.**

Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia

Abstract. The analysis of the shifts of atomic centers in unrestrained crystallographic refinement of macromolecular structure enables the detection of alternative conformations of polypeptide chain fragments. Decision-making procedures of the presence or absence of alternative conformations for a particular fragment are based on statistical analysis of atomic shifts obtained during the trial unrestrained refinement of large number of protein structures. The analysis also showed that probability distributions of atomic shifts appear to depend on the program used for refinement, so the decision making procedures should be adapted to the refinement program. The construction of databases with atomic shifts, their analysis and the development of automatic decision-making procedures for detection of alternative conformations with the use of popular refinement program Refmac are described in this paper.

Key words: *structure of biological macromolecules, crystallographic refinement, alternative conformations.*

INTRODUCTION

One of the features of crystals for X-ray diffraction studies is that they contain considerable amount of the solvent (50% on the average). Side chains of amino acids exposed to the solvent usually possess increased mobility. As a result, such side chains may have different (alternative) conformations in different copies of molecules even in the crystal form. For high quality crystals this uncertainty converges to a small number of alternatives which are included into the model with different weights (occupancies) reflecting the frequency of particular conformation in the crystal under study. At present, the cases of two or three alternative conforamations (ACs) are studied. ACs may be introduced not only for side chains but also for main chains [1-3]. Inclusion of ACs in the model increases the accuracy of the model and the accuracy of structure factor phases, thus providing more reliable electron density maps over the whole unit cell. Moreover, in many cases ACs are related to protein function [2,4-6]. It has been noted that the number of residues that can be modeled in ACs increases with the resolution of the experimental data [1,3,7]. Detection of residues that are present in ACs is time-consuming and demands visual analysis of electron-density maps and is subjective in character. Previously we have suggested decision-making procedures to facilitate this work.

The procedure of crystallographic refinement from the mathematical point of view is a minimization procedure that leads to the best correspondence between the model and experimental data and correspondence between the model and some a priori (stereochemical) requirements. Usually, a target function for crystallographic refinement of biological macromolecules contains two parts. The first part of this criterion ensures correspondence

* oleg@impb.psn.ru
** lunin@impb.psn.ru

between model and X-ray data. The second part of the criterion puts restraints on model geometry. The presence of the second part compensates for the low data-to-parameter ratio and stabilizes refinement. Unrestrained refinement is unstable at middle resolutions of experimental data. At the same time there is an evidence that well-ordered regions of a structure can be refined without stereochemical restraints in the last stages of refinement if the resolution of the experimental data is high enough [8-10]. Sometimes it is possible to refine the whole model without stereochemical restraints [5]. It was noted long ago that even at atomic resolution poorly ordered residues deteriorate significantly in unrestrained refinement while the ordered residues are stable [8].

The basis of suggested method for identification of ACs is the hypothesis that significant deterioration of polypeptide chain fragments during unrestrained refinement indicates the AC presence [11]. Therefore unrestrained refinement of the structure may be used to get information about degree of disorder even if unrestrained refinement itself does not lead to stereochemically sensible model.

Previously we studied atomic mobility during unrestrained refinement of individual structures and also have conducted statistical analysis of atomic mobility during unrestrained refinement of 203 structures with resolution better than 1.2 Å [12]. This research showed that analysis of atomic mobility in unrestrained refinement may be used for identification of residues that most probably possess ACs. Methods for visual and automatic analysis of atomic shifts and automatic procedures were suggested. These methods were implemented as computer programs Shift_plot and AC_prediction [13]. They process the results of unrestrained refinement with one of the most popular program for crystallographic refinement phenix.refine [14].

The aim of this work was the study of atomic mobility in unrestrained refinement in another popular refinement program Refmac [15]. Parameters for the decision-making procedures for analysis Refmac unrestrained refinement were obtained during this research. The data obtained was incorporated in AC_prediction program making it compatible with Refmac refinements.

1. MATHERIALS AND METHODS

1.1. Choosing and preparation of the models for the research

Two databases were composed from structures deposited to PDB [16] to study the mobility of atoms in unrestrained refinement and derive parameters for decision-making procedures. The first database corresponded to resolution between 1.1 Å and 1.2Å; the second database corresponded to resolution better than 1.1Å. Other conditions are listed in Table 1. The limits were set for *R*-factor values of the selected models to choose only high-quality structures which were studied carefully, so that the alternative conformations were assigned reliably.

The process of preparing the database consisted of three stages: model preparation, the trial unrestrained refinement, and analysis of atomic shifts. The goal of the first stage was to 'restore' the model as it could be before alternative conformations were included. For this purpose each residue was left in the model in single conformation with larges occupancy. Afterwards, all occupancies were set to 1. The standard refinement was performed for this model with Refmac (10 cycles) using anisotropic ADP, riding hydrogens and water molecules. At the second stage the Trial unrestrained refinement (TUR) consisted of 10 cycles was performed for this model with Refmac. At the third stage we have calculated atomic shifts after the TUR for all atoms and saved them in the databases along with supplementary information such as PDB-code of the structure, number of the atom and the residue, alternative conformation identifier, ADP value *etc*. We used an automated processing procedure which failed to process some of the selected models at different stages of the preparation. Main reasons for these failures were the inability to convert automatically cif-file to mtz format, or the inability to produce automatically cif-file for a ligand. Some of refinements ended with rather high *R*-factor values after the first stage. Structures with R_{work} values greater than 0.15 were excluded from the both databases.

Table 1. Database statistics

	Database 1	Database 2
Conditions		
Resolution limits	$1.1 \le d \le 1.2$Å	$d < 1.1$Å
R-factor limits		
R_{work}	0.13	0.12
R_{free}	0.16	0.15
Statistics		
Structures selected from PDB	189	102
Structures prepared for tests	127	59
Non-H atoms in the database	243079	122566
Atoms in AC in the database	14851	12649

Nevertheless, the number of models that were processed successfully was sufficient to continue the research.

1.2. Decision-making procedures

Previously we have developed several decision-making procedures for automatic classification of residues on residues in single conformation (SC) and residues with alternative conformations (AC) based on results of unrestrained refinement. Each decision-making procedure utilizes the results of trial unrestrained refinement (TUR) and generates a list of residues that are most likely present in ACs. In present research TUR consist of 10 cycles of Refmac refinement with anisotropic ADP, riding hydrogens and water molecules. TUR should be carried out for a structure with good resolution of experimental data (better than 1.2Å) which is already well-refined (to the *R*-factor around 0.15)

1.2.1. Integral shift measure

In unrestrained refinement each atom gets an individual shift. To discuss alternative conformations in terms of residues it is convenient to use some integral measure that reflects mobility of a group of atoms (e.g. main-chain or side-chain atoms of a particular residue) rather than of individual atoms. We used two such measures, namely mean atomic shift (calculated as the arithmetic mean of all shifts in atom group) and maximal atomic shift in atom group.

1.2.2. Threshold criteria

The first way for producing a list of potential AC residues (residues that are most likely present in alternative conformations) is to compare integral shift values obtained in the TUR for each residue against some predefined threshold value. If the integral shift value is greater than the threshold then the residue is classified as an AC residue otherwise as SC. The threshold value is a parameter of the decision making procedure. It may be adjusted to give the best prediction quality. Different kinds of atoms (e.g. side and main chain atoms) may have different mobility in unrestrained refinement. Therefore, threshold values may be set differently for different kinds of atoms.

1.2.3. Likelihood-based criteria

The second type of the decision-making procedure is based on the maximum likelihood principle. In this case we use two types of probability distribution for an integral shift value (for example, for the mean atomic shift). The first distribution corresponds to the residues present in a SC and the second to residues present in ACs. These distributions were derived empirically using the database described in Sec. 1.1. The distributions may be different for different kinds of atoms. To make a decision concerning a particular residue the integral shift value is calculated as

well as the probabilities to meet this value for the SC-case and for the AC-case. If the probability is larger for the AC-case, then the residue is classified as AC-residue. Otherwise, it is classified as SC residue.

1.3. Testing of decision-making procedures

The suggested decision making procedures may be considered as binary classification statistical tests, and we used standard statistical approaches to estimate their quality. Both databases have been divided into two equal parts to adjust procedures parameters and evaluate procedures performance. The first part (training sample) was used to find appropriate cutoff levels and calculate empirical distributions of the shifts. The second part (test sample) was used to evaluate the quality of the procedures, supposing PDB assignments of SC or AC as the 'true answers'. For each residue we consider the result of the test as positive if the decision making procedure gives the prediction 'the residue is present in AC' and as negative if the prediction is 'the residue is present in a SC'. We assume that the procedure gives a true prediction if the result matches structure deposited in PDB, *i.e.* either the criterion indicates the presence of AC and the model deposited in PDB does contain alternative conformations for the residue, or the procedure indicates SC and this residue is present in the model in a single-conformation. Other variants are considered as false answers of the procedure. We assume that a residue has alternative conformation of main chain or side chain if it has at least one AC-atom in that chain. Alanine and glycine residues were excluded from the evaluation of the quality of the side chain prediction.

1.3.1. Statistical criteria

The two main characteristics that were used to evaluate the quality of the statistical test are type 1 error and type 2 error (fp_rate, fn_rate):

$$\text{fp\_rate} = \frac{FP}{TN + FP}, \qquad (1)$$

$$\text{fn\_rate} = \frac{FN}{TP + FN}, \qquad (2)$$

where *TP* (true positive) is a number of correctly predicted AC, *FP* (false positive) is a number of wrong 'AC'-predictions, *TN* (true negative) is a number of correctly predicted SC, *FN* (false negative) is a number of wrong 'SC'-predictions.

The values of fp_rate and fn_rate are competitive values and depend on the threshold value. Fig. 1a shows fp_rate and fn_rate *versus* threshold value for criterion based on mean atomic shifts for side chains. The two curves may be combined into one Receiver Operating

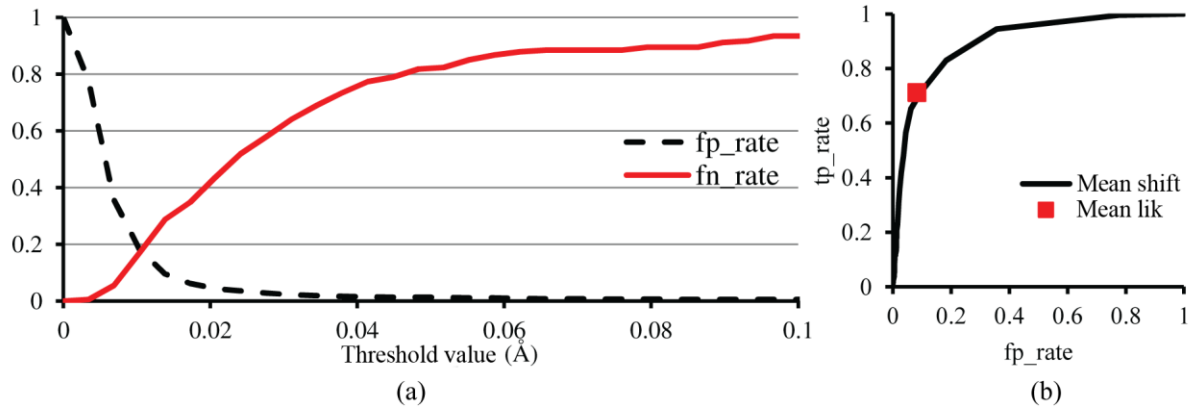

Fig. 1. Probabilities of type 1 error (black dashed lihe) and type 2 error (red solid line) *versus* threshold value (a), ROC-curve and marker for decision-making procedure based on maximum likelihood (b). Criterion based on analysis of mean atomic shifts for side inner chains (test sample, 0.099 < *Rwork* < 0.129).

Characteristic (ROC) curve (Fig.1b). Every point on the ROC curve corresponds to some threshold value and has the coordinates fp_rate and (1-fn_rate). The curve shows dependence of a 'benefit' (the proportion of correctly predicted ACs) on a 'cost' (the proportion of wrongly predicted SCs). A good decision-making procedure should find a large percent of AC-residues and have low rate of false alarms. Evaluation of the quality of likelihood-based criteria may be reflected as the single point in ROC-space because they do not have any parameters to adjust.

We have used two more values for evaluation of statistical criteria: positive predictive value (PPV) and negative predictive value (NPV):

$$PPV = \frac{TP}{TP + FP}, \tag{3}$$

$$NPV = \frac{TN}{TN + FN}. \tag{4}$$

In our case *PPV* gives a share of correct predictions among the all ‘AC’-predictions and *NPV* gives a share of correct predictions among the all ‘SC’-predictions.

1.3.2. Summary characteristics

For threshold-based criteria the values type 1 and 2 errors are competitive values and depend on the threshold value. The same is valid for *PPV* and *NPV*. As a compromise, a single-value measure of goodness can be defined as the 'balanced accuracy':

$$bACC = \frac{(1 - \text{fp\_rate}) + (1 - \text{fn\_rate})}{2} \tag{5}$$

as suggested earlier in CASP competitions in section “Identifying disordered regions in target proteins” [17]. In threshold-based procedures *bACC* depends on the threshold value, and the choice of the threshold was optimized to have the maximal possible balanced accuracy.

Another summary value used for evaluation is the Area Under the Curve (*AUC*) calculated for the ROC curve. This value can be calculated only for the threshold-based procedures. A large *AUC* value means that we can choose a threshold resulting in small probabilities of the both types of errors simultaneously. In our study *AUC* value is equal to the probability to have a larger shift value for an AC-residue than for a SC-residue if these two residues are chosen randomly [18].

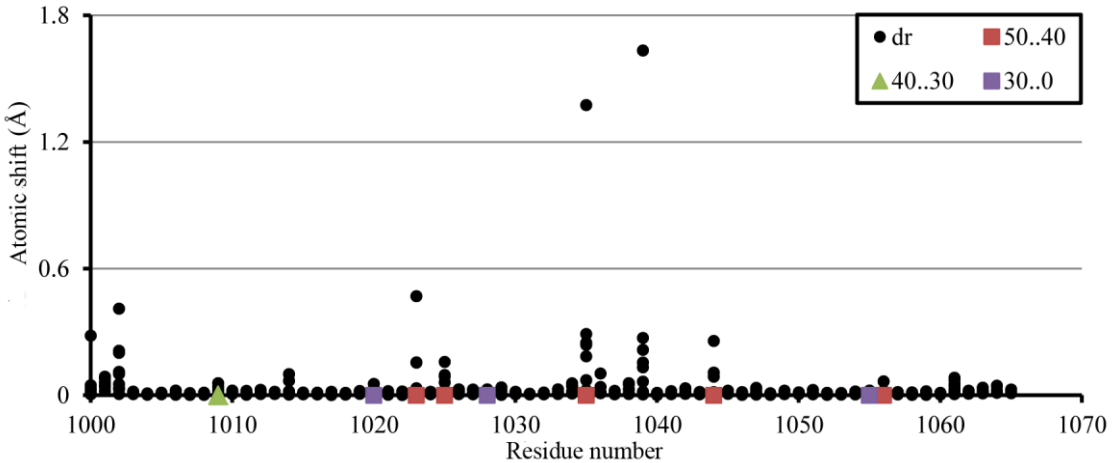

Fig. 2. Atomic shifts obtained in TUR for 1hg7 structure. Each column of dots represents shifts of non-hydrogen atoms of one residue. The residues are numbered as in the PDB file. Markers indicate residues that were defined by the authors of the structure as AC residues. The colour and shapes of the markers reflects the refined values of occupancies (in percent) for the less populated conformation.

2. RESULTS AND DUSCUSSION

2.1. Diagrams of atomic shifts

Fig.2 shows the diagram of atomic shifts plotted for 1hg7 [19] entry. The markers in this figure highlight the residues, for which the authors of the structure have introduced AC. The colour and shape of the marker reflect the refined value of the occupancy for the less populated conformation. It is easy to see that in general, ACs were introduced for residues that get significant atomic shifts in the unrestrained refinement. Moreover, AC with near to equal occupancies were found for residues that reveal larger atomic shifts in comparison with residues possessing of a minor occupancy (e.g. 0.3) for one of alternative conformations. Some residues with AC in PDB model do not reveal themselves by large atomic shifts. Such diagram gives overall impression about disorder of different areas of the structure. The final decision about the significance of atomic shift magnitudes should be made by the user. The procedures described below help to automate this decision.

2.2. Atomic mobility in unrestrained refinement

All atoms in the databases can be divided in two types, namely SC-atoms and AC-atoms. Fig. 3a shows empirical probability distributions of atomic shifts for these atom types for structures with resolution 1.1–1.2Å. These distributions are significantly different. This observation provides the possibility to distinguish these atom types by their atomic shifts in TUR. The distribution for SC-atoms has more pronounced peak situated on the left side of the graph. The distribution for AC-atoms is smeared implying greater atomic shifts for AC-atoms.

Fig. 3b shows mean atomic shifts for different types of atoms in unrestrained refinement of structures with resolution 1.1–1.2Å. Mean values are different for different types of atoms. Moreover, mean atomic shifts increase with *R*-factor value. These results are in good agreement with previously obtained when phenix.refine was used for refinement [12].

2.3. Evaluation of decision-making procedures

Five decision-making procedures were tested in the present research: two threshold-based for mean (mean) and maximal (max) atomic shift and three likelihood-based for mean (mean_lik), maximal (max_lik) and individual (shift_lik) atomic shift. The results are presented in Tables 2 and 3. The tables 4 and 5 summarize the results of evaluation of threshold criterion based on mean atomic shifts. Fig. 4 shows ROC-curves for four different atom types. The errors obtained for likelihood based criteria and best threshold choice (best_thr) are shown by markers.

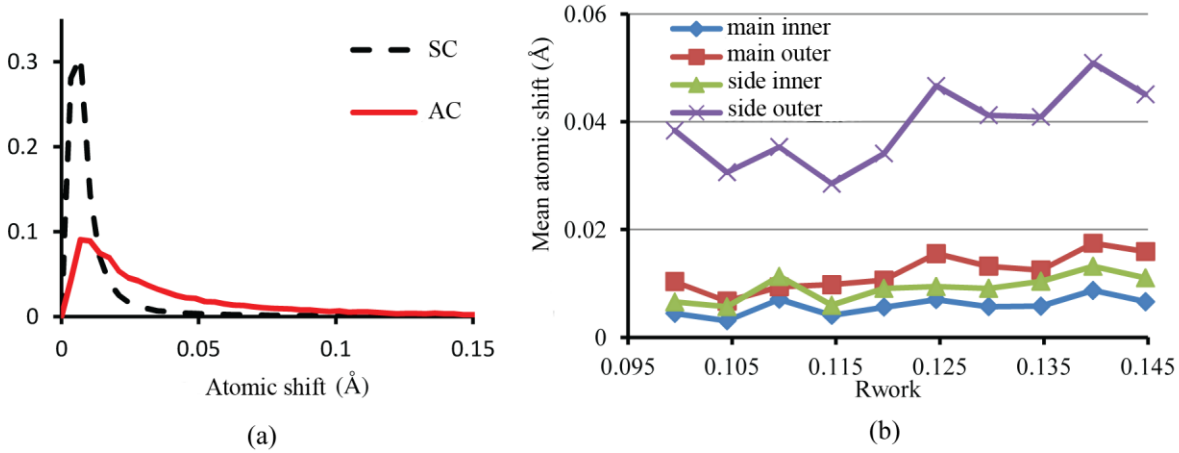

Fig 3. The distribution of atomic shifts for SC atoms (black dashed line) and for AC atoms (red solid line) for database 1 (a). Mean atomic shifts for different types of atoms in SC *versus* *R*-factor (database 1) (b).

The testing of automatic decision-making procedures showed that, as it was expected, the results are statistical in nature and contain errors of both types (false alarms and missing of AC). The achieved prediction quality is different for different parts of a structure. The best prediction was obtained for side chains situated inside the globule. All five procedures showed near to similar prediction quality for all of them. The prediction accuracy achieved at the moment does not give a reason to decide on the advantage of any of tested procedures.

Table 2. Balanced accuracy (*bACC*) for different critera (database 1, test sample)

Criterion type	Atom type					
	main inner	main outer	side inner	side outer	main (overall)	side (overall)
mean_shift	0.697	0.664	0.817	0.730	0.684	0.787
max_shift	0.687	0.670	0.797	0.715	0.693	0.773
mean_lik	0.650	0.619	0.792	0.687	0.640	0.743
max_lik	0.690	0.649	0.781	0.663	0.695	0.721
shift_lik	0.650	0.640	0.760	0.688	0.650	0.735

Table 3. The area under the ROC curve (*AUC*) for threshold-based criteria (database 1, test sample, $0.099 < Rwork < 0.129$)

Criterion type	Atom type			
	main inner	main outer	side inner	side outer
mean_shift	0.744	0.703	0.898	0.789
max_shift	0.746	0.708	0.881	0.777

Table 4. Evaluation results of threshold criterion based on mean atomic shifts for database 1 (1.1–1.2Å), test sample

Atom type	*bACC*	*fp_rate*	*fn_rate*	*PPV*	*NPV*	Fraction of AC
Main inner	0.697	0.184	0.422	0.079	0.986	0.027
Main outer	0.664	0.235	0.437	0.144	0.962	0.066
Side inner	0.817	0.190	0.176	0.207	0.987	0.057
Side outer	0.730	0.386	0.155	0.280	0.957	0.151
Main (overall)	0.684	0.202	0.431	0.107	0.978	0.041
Side (overall)	0.787	0.263	0.162	0.249	0.978	0.094

Таблица 5. Evaluation results of threshold criterion based on mean atomic shifts for database 2 (d < 1.1Å), test sample

Atom type	*bACC*	fp_rate	fn_rate	*PPV*	*NPV*	Fraction of AC
Main inner	0.841	0.115	0.204	0.312	0.985	0.062
Main outer	0.778	0.271	0.173	0.297	0.968	0.122
Side inner	0.868	0.114	0.149	0.455	0.981	0.101
Side outer	0.753	0.350	0.144	0.460	0.928	0.259
Main (overall)	0.822	0.169	0.188	0.304	0.980	0.083
Side (overall)	0.828	0.197	0.146	0.458	0.966	0.163

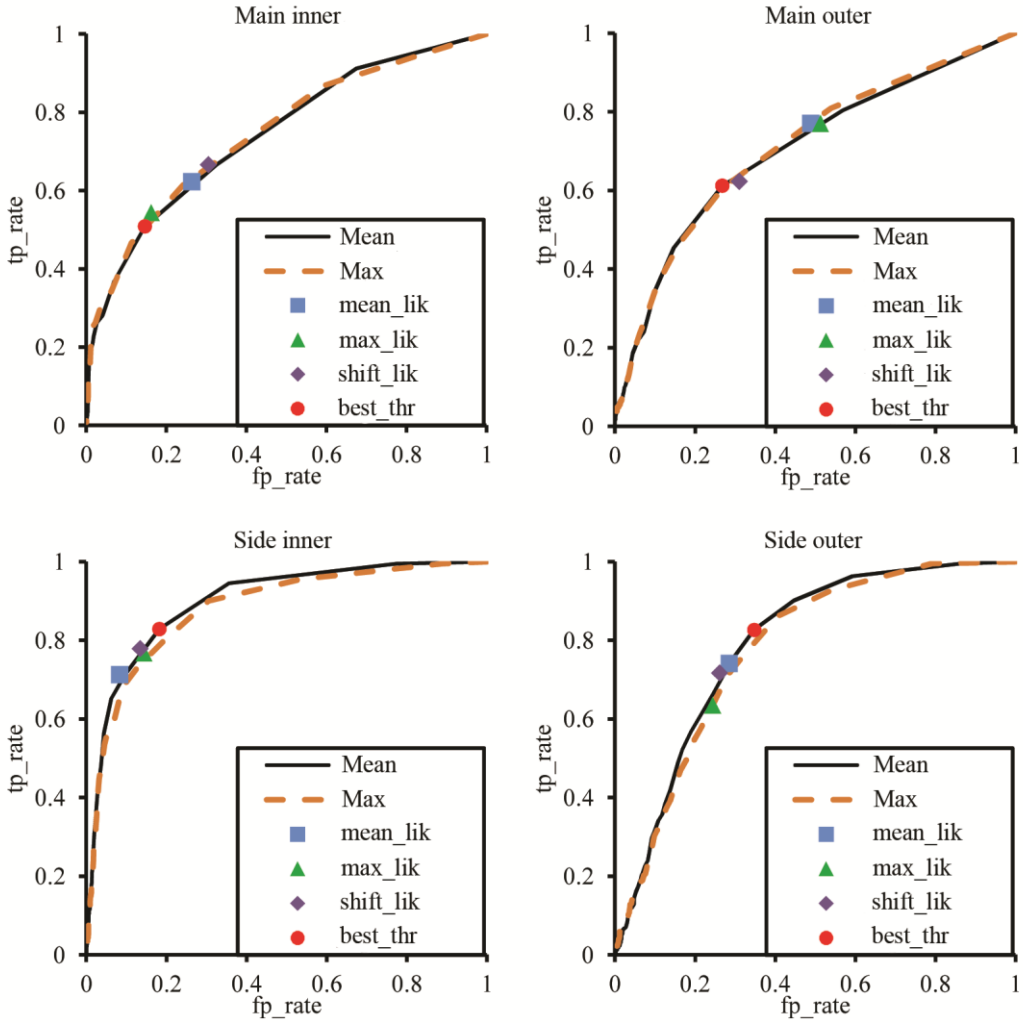

Fig. 4. Evaluation of decision-making procedures (database 1, test sample, $0.099 < Rwork < 0.129$) for different atom types. Solid and dashed lines correspond to mean shift (mean) and maximum-shift (max) threshold-based criteria, respectively. Squares, triangles and diamonds correspond to criteria based on the likelihood of mean (mean_lik) and maximum (max_lik) shifts and individual shifts of atoms (shift_lik). The circle (best_thr) marks the optimal threshold value corresponding to the mean-shift criterion.

CONCLUSION

A trial unrestrained refinement and subsequent analysis of atomic shifts can reveal the most unstable parts of the structure. These parts are prime candidates for the introduction of alternative conformations, therefore they should be checked first with electron density maps. Although the proposed formal decision-making procedures do not allow exact prediction which residues are present in alternative conformation, they provide an opportunity to reduce the amount of work for the visual analysis of the maps.

New threshold values and empirical probability distributions were obtained. They extend methods developed before and allow using Refmac for refinement. This data was added into the programs for analysis of atomic shifts. The programs and documentation are available from the following address: www.impb.ru/lmc/programs/ac_prediction/

This work was supported by RFBR, grant № 12-04-31096.

REFERENCES

1. Howard E.I., Sanishvili R., Cachau R.E., Mitschler A., Chevrier B., Barth P., Lamour V., Van Zandt M., Sibley E., Bon C., Moras D., Schneider T.R., Joachimiak A., Podjarny A. *Proteins.* 2004. V. 55. P. 792–804.
2. Ševčík J., Dauter Z., Wilson K.S. *Acta Crystallographica D.* 2004. V. 60. P. 1198–1204.
3. Wang J., Dauter M., Alkire R., Joachimiak A., Dauter Z. *Acta Crystallographica D.* 2007. V. 63. P. 1254–1268.
4. Rypniewski W.R., Østergaard P., Nørregaard-Madsen M., Dauter M., Wilson K.S. *Acta Crystallographica D.* 2001. V. 57. P. 8–19.
5. Getzoff E.D., Gutwin K.N., Genick U.K. *Nat. Struct. Biol.* 2003. V. 10. P. 663–668.
6. Kursula I., Wierenga R.K. *J. Biol. Chem.* 2003. V. 278. P. 9544–9551.
7. Addlagatta A., Krzywda S., Czapinska H., Otlewski J., Jaskolski M. *Acta Crystallographica D.* 2001. V. 57. P. 649–663.
8. Dauter Z., Sierer L.C., Wilson K.S. *Acta Crystallographica D.* 1992. V. 48. P. 42–59.
9. Dodd F.E., Hasnain S.S. *Acta Crystallographica D.* 1995. V. 51. P. 1052–1064.
10. Koepke J., Scharff E.I., Lücke C., Rüterjans H., Fritzsch G. *Acta Crystallographica D.* 2003. V. 59. P. 1744–1754.
11. Sobolev O.V., Lunin V.Y. *Mathematical Biology and Bioinformatics*. 2008. T. 3. C. 50–59. URL: http://www.matbio.org/downloads/Sobolev2008(3_50).pdf (accessed 21 December 2012).
12. Sobolev O.V., Lunin V.Y. *Acta Crystallographica D.* 2012. V. 68. P. 1118–1127.
13. Sobolev O.V. *Computational Crystallography Newsletter.* 2012. V. 3. P. 32–34.
14. Afonine P.V., Grosse-Kunstleve R.W., Echols N., Headd J.J., Moriarty N.W., Mustyakimov M., Terwilliger T.C., Urzhumtsev A., Zwart P.H., Adams P.D. *Acta Crystallographica D.* 2012. V. 68. P. 352–367.
15. Murshudov G.N., Skubak P., Lebedev A.A., Pannu N.S., Steiner R.A., Nicholls R.A., Winn M.D., Long F., Vagin A.A. *Acta Crystallographica D.* 2011. V. 67. P. 355–367.
16. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. *Nucleic Acids Res.* 2000. V. 28. P. 235–242.
17. Noivirt-Brik O., Prilusky J., Sussman J.L. *Proteins.* 2009. V. 77. Suppl. 9. P. 210–216.
18. Fawcett T. *Pattern Recognition Letters.* 2006. V. 27. P. 861–874.
19. Antson A.A., Smith D.J., Roper D.I., Lewis S., Caves L.S., Verma C.S., Buckley S.L., Lillford P.J., Hubbard R.E. *J. Mol. Biol.* 2010. V. 305. P. 875–889.