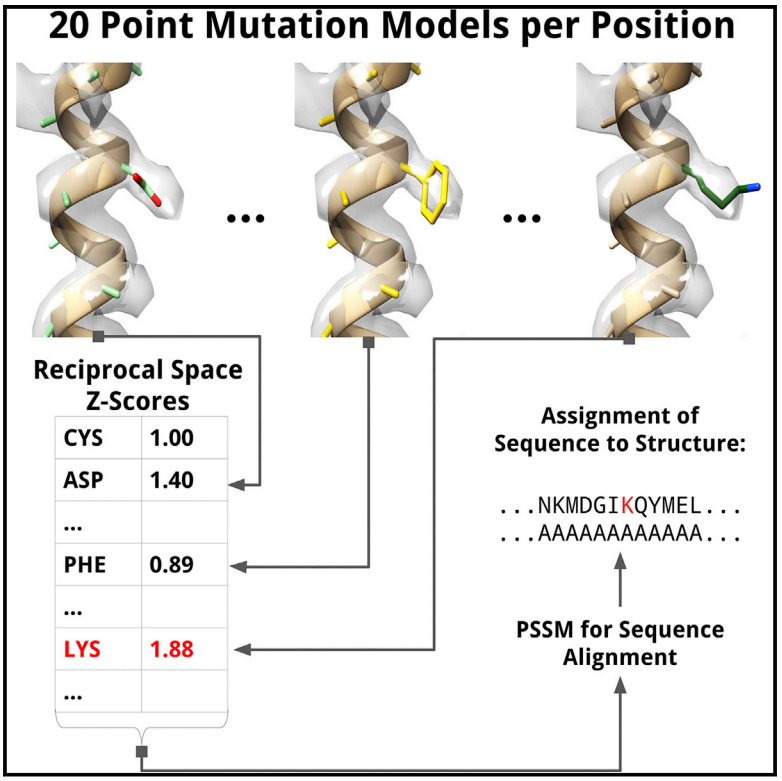
Resource

Structure

Automatic Inference of Sequence from Low-Resolution Crystallographic Data

Graphical Abstract



Highlights

- Reciprocal-space measures are sensitive enough to identify specific residue types
- Residue-type preferences are combined to thread a sequence onto the backbone
- An application for sequence assignment at 3.5–5.0 Å resolution

Authors

Ziv Ben-Aharon, Michael Levitt, Nir Kalisman

Correspondence

nirka@mail.huji.ac.il

In Brief

Ben-Aharon et al. find that certain crystallographic measures are more informative than previously assumed. They use these findings to solve a difficult technical problem in low-resolution crystallography: the identification of the amino acid types along the protein backbone.

Ben-Aharon et al., 2018, Structure *26*, 1546–1554
November 6, 2018
https://doi.org/10.1016/j.str.2018.08.011

CellPress

CellPress

Structure
Resource

Automatic Inference of Sequence from Low-Resolution Crystallographic Data

Ziv Ben-Aharon,[1] Michael Levitt,[2] and Nir Kalisman[1,3,*]
[1]Department of Biological Chemistry, The Hebrew University of Jerusalem, Jerusalem 91904, Israel
[2]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305, USA
[3]Lead Contact
*Correspondence: nirka@mail.huji.ac.il
https://doi.org/10.1016/j.str.2018.08.011

SUMMARY

At resolutions worse than 3.5 Å, the electron density is weak or nonexistent at the locations of the side chains. Consequently, the assignment of the protein sequences to their correct positions along the backbone is a difficult problem. In this work, we propose a fully automated computational approach to assign sequence at low resolution. It is based on our surprising observation that standard reciprocal-space indicators, such as the initial unrefined R value, are sensitive enough to detect an erroneous sequence assignment of even a single backbone position. Our approach correctly determines the amino acid type for 15%, 13%, and 9% of the backbone positions in crystallographic datasets with resolutions of 4.0 Å, 4.5 Å, and 5.0 Å, respectively. We implement these findings in an application for threading a sequence onto a backbone structure. For the three resolution ranges, the application threads 83%, 81%, and 64% of the sequences exactly as in the deposited PDB structures.

INTRODUCTION

Crystallographic datasets of great interest to structural biologists often have relatively low resolutions. At the resolution range of 3.5–4.5 Å, the electron density usually allows reliable tracing of the protein backbone. However, it reveals very little information on the identity and conformation of the side chains at the positions along the backbone (Figure 1). This is problematic for sequence assignment because current assignment tools rely on correlations between the local electron density and the side-chain geometries to infer the likely residue type at a backbone position (Zou and Jones, 1996; Holton et al., 2000; Terwilliger, 2003; Cohen et al., 2004; Cowtan, 2008). Not surprisingly, all these tools show a very sharp drop in performance for resolutions worse than 3 Å. Consequently, crystallographers working at low resolutions run the risk of assigning stretches of sequence to their incorrect positions on the backbone. This problem is most severe when the protein complex comprises several sequences that are highly similar to each other. In such cases, an entire protein subunit might be erroneously assigned onto the wrong backbone trace (Dekker et al., 2011; Kalisman et al., 2013). These difficulties have led to the release of incorrect crystallographic structures or to initial uncertainty as to their correctness (Fleishman et al., 2004; Maeda et al., 2009).

To assign sequences onto low-resolution backbone traces, crystallographers generally use one or more of the following three approaches: molecular replacement, heavy-atom additions, and identification of large aromatic residues. Molecular replacement is currently the most common approach, wherein the sequence assignment is solved concurrently with the solution of the phases (Forneris et al., 2010; Lu et al., 2013; Shaya et al., 2014; Karakas and Furukawa, 2014; Bae et al., 2013; Wang et al., 2014). For this approach to work, an accurate model for a significant portion of the protein mass within the asymmetric unit must be available. Software such as PHASER (McCoy et al., 2007) can then determine the location and orientation of the relevant protein mass within the asymmetric unit through comparison of the Patterson maps. The assignment of sequence to backbone, which is known for the original model, is then transferred to the crystallographic model of the low-resolution dataset. While molecular replacement is effective, it is lacking in two ways. First, it does not address the assignment of sequence elements that are not part of the original phasing model, such as additional loops, domains, or entire protein subunits. Second, it does not provide independent crystallographic evidence for the correctness of the sequence assignment of the phasing model.

Heavy-atom additions are used primarily to solve the phase problem. As an added advantage, they are also used to assign the locations of certain amino acid residues along the backbone trace. This approach is based on the labeling of specific amino acids with certain heavy atoms that are not native to the protein structure. The crystallographic techniques of isomorphous replacement or anomalous scattering then determine the locations along the backbone of the heavy atoms. Consequently, the crystallographer is able to anchor specific residues along the backbone and use these anchor positions to assign the rest of the sequence. Applications of this approach often rely on incorporation of seleno-methionine into proteins (Bae et al., 2013; Özkan et al., 2014; Wang et al., 2014) or the soaking of the crystals in heavy-atom solutions (Aller et al., 2009; Lu et al., 2013). While providing direct evidence for certain residue identities, heavy-atom labeling has several drawbacks. It requires additional experimental effort from the crystallographer and might not always be applicable, such as for large complexes

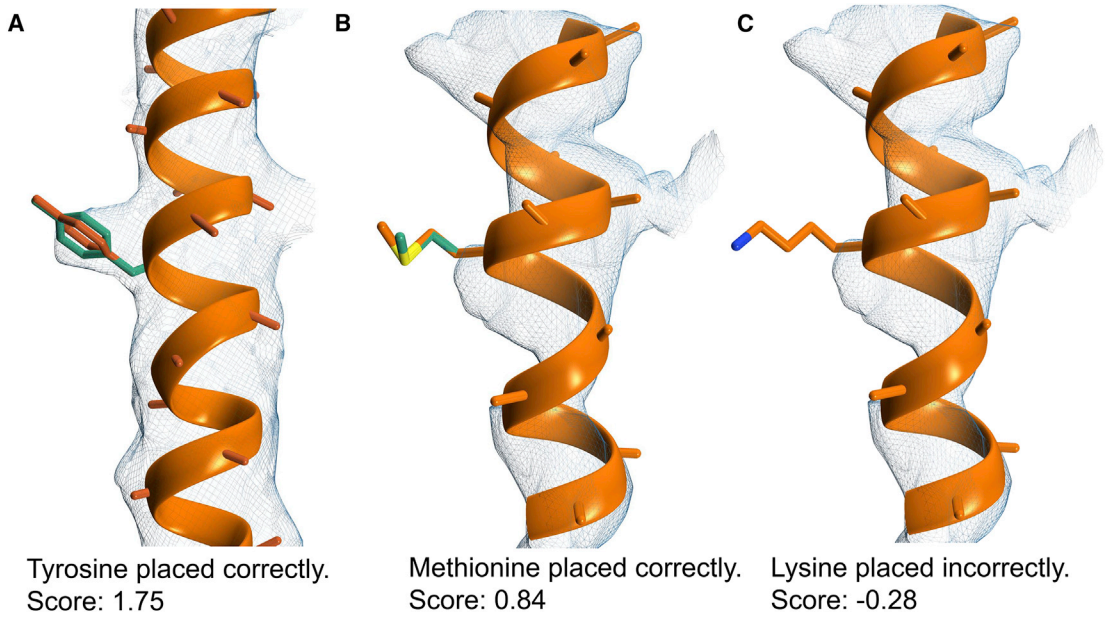


Figure 1. Electron Density Maps at 4.0-Å Resolution Reveal Very Little Information on the Side-Chain Identity Even for a Well-Ordered Helical Backbone

Our application can identify the correct side chain based on a scoring scheme that relies on correlation to the structure factors in reciprocal space.

(A) A backbone region from PDB: 4LTP with the side chain of Tyr 162 shown in green and its predicted rotamer by SCWRL4 in orange. Out of the 20 amino acids, our application assigns the highest score in that position to a tyrosine.

(B) A similar example from PDB: 3T51 shows the side chain of Met 696. For this position, our application also correctly assigns the highest score to a methionine.

(C) When a lysine is assigned incorrectly at the same position as in (B), the score is poor in spite of the general structural similarity between the side chains of lysine and methionine.

that are purified intact (Dekker et al., 2011; Muñoz et al., 2011). Additionally, the labeled positions might not spread evenly along the backbone, making the assignment of some positions difficult.

Finally, the electron density of aromatic residues is also used (Aller et al., 2009; Dekker et al., 2011; Lu et al., 2013), although rarely as the sole evidence for assignment. Because of the size and structure of their side chains, the electron density of aromatic residues is often noticeable on an otherwise featureless backbone (Figure 1A). This approach does not require additional experimental work and aromatic residues are generally abundant along the sequence. Unfortunately, the association between electron density and aromatic side chains is not always clear at low resolutions. An aromatic side chain might not show a clear density and conversely, strong density can accompany a non-aromatic residue. These difficulties make locating aromatic residues a rather subjective and error-prone approach for sequence assignment.

In this work, we propose a fourth approach to assign a sequence to a backbone, which is computational and completely automatic. It is based on our finding that standard crystallographic indicators in reciprocal space, such as the R value, are sensitive enough to detect the residue type of even a single backbone position. We show that for many of the positions, we can assign a score for the correct amino acid type that is better than that of the other 19 types (Figures 1B and 1C). We rely on this observation to mutate each backbone position to the 20 amino acid types and calculate a fit score for each mutation. Based on this preference, we develop a computational scheme to thread the entire sequence onto the backbone.

RESULTS

We first demonstrate how the correct residue type at a single position along the backbone can be inferred. Similar to existing tools, we mutate each position into the 20 amino acids, score the models, and then predict the residue type to be the best-scoring mutation (Holton et al., 2000; Terwilliger, 2003; Cohen et al., 2004; Cowtan, 2008). What we do differently is the use of reciprocal-space measures to score the mutations rather than relying on real space correlations to the electron density.

The inputs to our approach are the crystallographic structure factors and the coordinates of the protein backbone atoms in the asymmetric unit. All other elements of the model, which are designated by "HETATM" in the PDB format, are omitted in the analysis that follows. We convert all residues into alanines and place the Cβ atoms using SCWRL4 (Krivov et al., 2009). Next, we mutate each backbone position in turn into the 20 amino acid types and use SCWRL4 to position the side-chain rotamers. The temperature factors of the side-chain atoms are set to be the same as that of the residue Cα. The result is a set of 20 models that are all identical except for the mutations at the position in question (as in Figures 1B and 1C). For each model, we use the SFCheck program (Vaguine et al., 1999) to calculate its fit to the structure factors. SFCheck outputs two measures in reciprocal space that report the quality of fit: the R value and

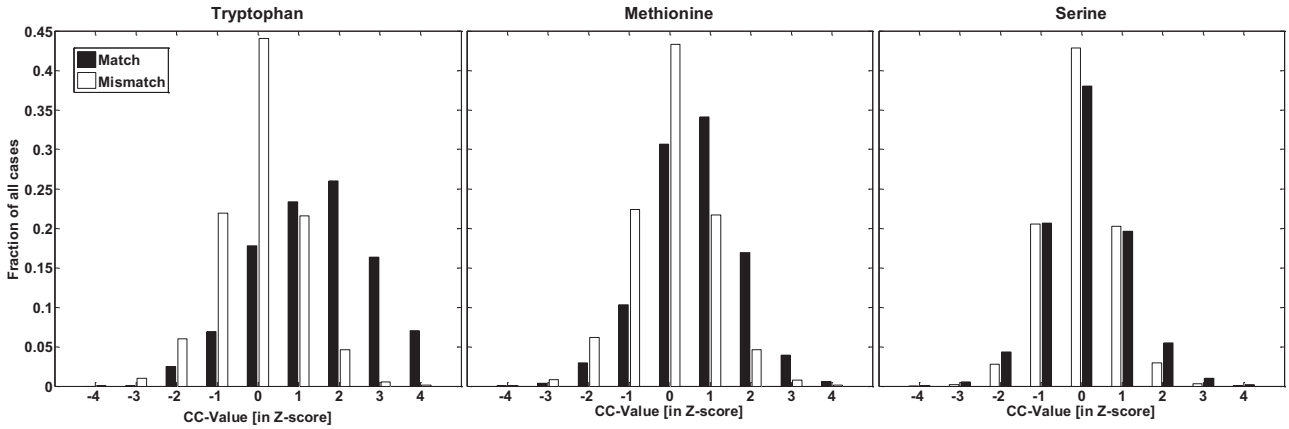


Figure 2. Histograms of Reciprocal-Space Correlation Coefficient (CC) Values for All Backbone Positions in the 3.8–4.0-Å Dataset

For a given amino acid type: match (black bars) shows values for positions where the modeled amino acid was the same as that reported in the PDB entry; mismatch (white bar) shows all other positions. Higher CC values indicate a better fit between a model and the crystallographic structure factors. The histograms for tryptophan, methionine, and serine amino acids are shown as typical examples for strong, moderate, and negligible CC-value signal, respectively.

the correlation coefficient (CC) value. We chose to work with the CC value, which performed slightly better than the R value (data not shown). The mutation scan gives an N-by-20 matrix of CC values for each chain (where N is the number of residue positions along the chain).

It is meaningless to directly compare the CC values between models that do not have exactly the same number of atoms. To enable comparison between the mutations at a certain position, we convert the CC values into an N-by-20 matrix of *Z* scores. This was done by calculating the mean and SD of each column in the CC-value matrix and converting that column accordingly (Terwilliger, 2003). The *Z* scores in each row can now be directly compared for the best-fitting mutation. Moreover, *Z* scores from chains belonging to different crystallographic datasets can now be pooled together for a comprehensive statistical analysis. We applied this approach to a set of crystallographic structures in the 3.8–4.0-Å resolution range and obtained 331 matrices with a total of 117,050 rows, each with 20 CC values (expressed as *Z* scores). This extensive dataset shows a clear CC-value signal that reports on the amino acid preference of each position (Figure 2). For example, our dataset contains 1,518 positions that are assigned as tryptophan in the deposited PDB models and 115,532 positions that are not tryptophan. When mutated into side chains of tryptophan, the positions that were originally reported as tryptophan (''Matches,'' black bars in Figure 2A) had a median *Z* score of 1.45. This is markedly higher than the *Z* scores obtained by all the other positions (''Mismatches,'' white bars in Figure 2A) that have a median *Z* score value of 0.0. The *Z* score can therefore indicate quite significantly whether a position is a tryptophan.

The median *Z* scores of matches and mismatches for all the 20 amino acid types are summarized in Table 1. Nearly all amino acid types show a positive *Z*-score signal discriminating between matches and mismatches. As expected, the bulky and ringed amino acids (PHE, HIS, TRP, and TYR) show strong signals, which are consistent with their current role in user-guided sequence threading. It is surprising, however, to observe that other amino acid types (most notably GLY and MET) gave signals that are almost as strong. The overall positive signal implies that information is available on the sequence of any stretch of backbone, thereby eliminating the dependency on aromatic amino acids as sequence ''anchors.'' Another unexpected feature is the nearly negligible signal shown by the small amino acids (SER, THR, and VAL). At present, we cannot offer an explanation for their weak signals. It is certainly not because of modeling inaccuracy, since the branched THR and VAL amino acids are among the best-predicted side chains by SCWRL4, with over 95% rotamer accuracy (Krivov et al., 2009). It is also not likely to be size per se, since the smaller GLY shows one of the strongest signals. Note that the null signal of ALA is a direct result of our methodology (ALA mutations on an all-ALA background) and does not carry any meaning.

We point out two other remarkable features of our approach. First, the absolute magnitude of the CC-value signal is very small and apparent only in the fourth digit after the decimal point (Table 1, rightmost column). In fact, we had to ''hack'' the SFCheck source code in order to print six significant digits of the CC value in the output. Previously, these digits were thought to be insignificant, but as we show here, they carry valuable signal. A second observation is that the CC-value signal decreases very moderately with the chain length. The *Z* scores of chains longer than 1,000 residues are only slightly worse than those of much shorter chains (Table 2). Although we have only 16 such chains in our database, the moderate decrease does not seem to be an artifact of small sample size. This observation seems counterintuitive, as one might expect the signal to be inversely proportional to the total protein mass. However, the use of *Z* scores may largely compensate for the size of the system.

Table 3 lists the CC-value signals that we observed for resolutions worse than 4.0 Å. Encouragingly, a significant signal is still seen at resolutions of 4.0–4.5 Å. For the next resolution range, 4.5–5.0 Å, we observe a sharp drop in performance. We note, however, that the signal is still not negligible, and useful results can probably be obtained by combining our approach with external structural information. For example, we previously

Table 1. The Discrimination Power for Each Amino Acid Type According to the CC-Value Scan

	Median of Matches [*Z* Score]	Median of Mismatches [*Z* Score]	Total Number of Positions	Mean of CC Value	SD of CC Value
ALA	0.00	0.00	8,239	0.630	0.00000
CYS	0.45	0.00	1,616	0.626	0.00040
ASP	0.39	−0.01	6,053	0.625	0.00045
GLU	0.27	−0.01	7,661	0.622	0.00053
PHE	1.14	−0.02	5,167	0.620	0.00063
GLY	0.96	0.00	7,457	0.629	0.00015
HIS	0.85	0.00	2,272	0.620	0.00063
ILE	0.20	0.00	7,827	0.628	0.00032
LYS	0.10	0.00	6,706	0.625	0.00046
LEU	0.20	−0.01	11,994	0.627	0.00035
MET	0.60	0.00	2,705	0.621	0.00055
ASN	0.40	−0.01	5,345	0.626	0.00043
PRO	0.37	0.00	5,041	0.629	0.00025
GLN	0.39	−0.01	4,645	0.623	0.00049
ARG	0.37	−0.01	5,782	0.620	0.00057
SER	0.00	0.00	7,571	0.630	0.00020
THR	0.19	0.00	6,657	0.629	0.00027
VAL	−0.01	0.00	8,477	0.630	0.00022
TRP	1.45	0.00	1,518	0.620	0.00080
TYR	1.26	−0.02	4,317	0.620	0.00070
ALL	0.24	0.00	117,050	0.626	0.00037

Table compiled from 331 chains with resolution of 3.8–4.0 Å.

Table 2. Surprisingly, the Magnitude of the CC-Value Signal Decreases Only Slightly with the Size of the Structure

	Median of Matches [*Z* Scores]		
	All (331 Chains)	Length <250 (152 Chains)	Length >1,000 (16 Chains)
ALA	0.00	0.00	0.00
CYS	0.45	0.46	0.57
ASP	0.39	0.41	0.40
GLU	0.27	0.24	0.23
PHE	1.14	1.04	1.25
GLY	0.96	0.93	0.81
HIS	0.85	0.72	1.05
ILE	0.20	0.20	0.12
LYS	0.10	0.09	0.10
LEU	0.20	0.16	0.15
MET	0.60	0.41	0.59
ASN	0.40	0.33	0.37
PRO	0.37	0.36	0.27
GLN	0.39	0.33	0.39
ARG	0.37	0.37	0.44
SER	0.00	0.00	0.00
THR	0.19	0.21	0.13
VAL	−0.01	−0.05	0.00
TRP	1.45	1.49	1.47
TYR	1.26	1.19	1.37
ALL	0.24	0.22	0.19

Here we compare chains shorter than 250 residues with long chains of more than 1,000 residues. Resolution is 3.8–4.0 Å.

demonstrated that reciprocal-space measures on crystallographic data of 5.5 Å resolution can correctly assign the sequences of the eukaryotic chaperonin containing TCP-1 (CCT) chaperonin to their backbone traces (Muñoz et al., 2011; Kalisman et al., 2013). Yet, for this assignment we were required to rely on the known subunit arrangement of the CCT. In the future, we believe that emerging structural methodologies, such as correlated mutations or cross-linking coupled to mass spectrometry, could be used to augment our approach at very low resolutions.

One might be concerned that the CC-value signal arises from "residue memory" in the backbone coordinates that we use as input. To test this, we perturbed the backbones of the chains in the following way: the entire chain was reassembled from fragments of length five, which originated from crystal structures with no sequence similarity (Levitt, 1992). The reassembled models were then refined (as an all-ALA chain) according to the ENCAD force field (Levitt, 1983; Levitt et al., 1995) without reference to the crystallographic data. The resulting models showed good stereochemistry and had an average deviation of 0.72 Å from the original backbone trace (Table S1). While we expect that this perturbation removed any residue memory in the backbones, it led to a decrease of only 30% in the average *Z* score signal. In addition, we found that it had little effect on our ability to correctly thread the sequences on the backbones as we discuss below. Thus, most of the CC-value signal is not coming from any residue bias in the backbone. It is also encouraging to see that our approach is insensitive to backbone perturbations of 0.7 Å.

Support Vector Machine Enhances the Amino Acid Identification

We now consider a more elaborate inference scheme that builds on the CC-value signals. Specifically, we want to use the information in all the 20 CC values available for each backbone position. To that end, we trained 20 support vector machine (SVM) two-state linear classifiers, one for each amino acid type. Each classifier aims to determine if a certain position is either a certain amino acid or anything but that amino acid (e.g., TRP/non-TRP). The classifiers were trained and tested on a subset of the crystallographic structures in the 3.8–4.0-Å resolution range. We used a balanced set of positions for the training and testing of each classifier (Wei and Dunbrack, 2013). That set included all the positions matching the amino acid and an equal number of random positions that were not the amino acid. The set was divided into two equal subsets for training and for testing. Random guess on these subsets would result in 50% accuracy. Table S2 shows that the classifiers were considerably more successful than random with average accuracies of 66.5% and 67.2% on the test and training subsets, respectively. The nearly equal success rates obtained for training and testing indicates that the classifiers were not overtrained.

Table 3. The CC-Value Signal for Datasets of Different Resolutions

	Median of Matches [*Z* Scores]		
	3.8–4.0 Å	4.0–4.5 Å	4.5–5.0 Å
	(331 Chains)	(207 Chains)	(69 Chains)
ALA	0.00	0.00	0.00
CYS	0.45	0.53	0.30
ASP	0.39	0.40	0.34
GLU	0.27	0.28	0.26
PHE	1.14	0.90	0.52
GLY	0.96	0.60	0.41
HIS	0.85	0.80	0.52
ILE	0.20	0.04	−0.06
LYS	0.10	0.09	−0.01
LEU	0.20	0.01	−0.11
MET	0.60	0.49	0.31
ASN	0.40	0.43	0.25
PRO	0.37	0.15	0.03
GLN	0.39	0.40	0.23
ARG	0.37	0.39	0.30
SER	0.00	0.00	0.00
THR	0.19	0.16	0.16
VAL	−0.01	0.00	−0.21
TRP	1.45	1.30	0.76
TYR	1.26	1.09	0.73
ALL	0.24	0.15	0.06
Number of positions	117,050	77,946	18,746

Table 4. The Support Vector Machine Coefficients $C_{aa,i}$ and $Bias_{aa}$ of the Classifiers: TRP/non-TRP, and MET/non-MET

i	Mutation	*aa=TRP*	*aa=MET*
1	ALA	0	0
2	CYS	−0.27	−0.28
3	ASP	−0.06	0.17
4	GLU	0.22	0.19
5	PHE	0.39	0.13
6	GLY	−0.11	−0.08
7	HIS	0.14	−0.31
8	ILE	0.19	0.18
9	LYS	0.04	0.15
10	LEU	−0.08	0.33
11	MET	0.02	0.34
12	ASN	0.17	0
13	PRO	−0.04	−0.01
14	GLN	0.04	0.38[a]
15	ARG	0.01	−0.15
16	SER	−0.18	−0.02
17	THR	−0.08	−0.33
18	VAL	−0.01	0.31
19	TRP	0.76[a]	−0.02
20	TYR	−0.04	−0.28
Bias		−0.97	−0.44

$Bias_{aa}$, a vector of biases for each amino acid type; $C_{aa,i}$, the 20 × 20 matrix of coefficients.
[a]Most influential coefficient for a specific classifier.

Formally, a classifier for a certain amino acid type, *aa*, is a weighted sum of the 20 *Z*-scores at a certain position, *pos*:

$$CLASS_{pos,aa} = \sum_{i=1}^{20} C_{aa,i} \cdot ZS_{pos,i} + Bias_{aa} \quad where: \|C_{aa,1:20}\| = 1 \qquad \text{(Equation 1)}$$

where $C_{aa,i}$ is the 20-by-20 matrix of coefficients and $Bias_{aa}$ is a vector of biases for each amino acid type, both determined by the SVM as best separating the data. The higher the value of $CLASS_{pos,aa}$, the higher the probability that position *pos* is of amino acid type *aa*.

We use the SVM classifiers to generate a new set of N-by-20 matrices by replacing every $ZScore_{pos,aa}$ with $CLASS_{pos,aa}$. We demonstrate that the SVM matrices perform better than the *Z*-score matrices in inferring the amino acid preference at each position. To that end, we consider the simplest prediction scheme wherein the index of the maximal value in each row is predicted to be the amino acid type of the corresponding position. For chains with resolutions of 3.8–4.0 Å, this scheme correctly predicts the amino acid type for 15% of the positions with the SVM matrices, but only for 11% of the positions with *Z*-score matrices. For chains with resolutions of 4.0–4.5 Å, the scheme correctly predicts the amino acid type for 13% and 9% of the positions with the SVM and *Z*-score matrices, respectively. For chains with resolutions of 4.5–5.0 Å, the scheme correctly predicts the amino acid type for 9% and 7% of the positions with the SVM and *Z*-score matrices, respectively. We conclude that both matrix forms predict the amino acid type much better than random (1/20 or 5% correct), yet the values derived by SVM provide a stronger signal.

Table 4 lists the $C_{aa,i}$ coefficients for the classifiers of amino acids TRP and MET. The classifier for TRP has large positive weights of 0.76 and 0.39 for TRP and PHE mutations, respectively. This is not surprising given the strong *Z*-score signal of TRP and the fact that both TRP and PHE are aromatic. However, for other amino acids, the dependencies between the *Z* scores are more complex. For example, the largest weight for the MET classifier is assigned to *Z*-scores of GLN rather than MET. In fact, for 8 of the 20 classifiers, the largest weight is given to the *Z*-score of a mutation that is different from the amino acid being classified. We elaborate on the implications of these non-trivial relations in the discussion. Table S2 lists the coefficients of all 20 classifiers.

Sequence Identification and Threading

So far we have calculated the amino acid preferences of single positions along the backbone. We now use these preferences to resolve the sequence-to-backbone assignment at low resolutions. Our goal is to find the optimal alignment of the sequence to these preferences. Applications of sequence assignment to find such an alignment (Zou and Jones, 1996; Holton et al., 2000; Terwilliger, 2003; Cohen et al., 2004; Cowtan, 2008) use summation of the positional preferences, and we follow this same procedure. We also note that this task is very similar to the threading done in comparative modeling, which is well-studied in protein structure predication (Dunbrack, 2006).

```
STRUCTURE: ---------KRFEVKKWNAVALWAWDIVVDNCAICRNHIMDLCI
SEQUENCE:  MGTNSGAGKKRFEVKKWNAVALWAWDIVVDNCAICRNHIMDLCI

STRUCTURE: ECQ----------CTVAWGVCNHAFHFHCISRWLKTRQVCPLDN
SEQUENCE:  ECQANQASATSEECTVAWGVCNHAFHFHCISRWLKTRQVCPLDN

STRUCTURE: REWEFQKYGH
SEQUENCE:  REWEFQKYGH
```

Figure 3. An Example of Successful Automatic Threading of the Sequence of RBX1 Ubiquitin-Ligase onto Chain D of PDB: 4A0C (resolution 3.8 Å)

The gaps ("-") in the "STRUCTURE" rows indicate regions of the sequence that are unstructured in the crystallographic structure. Note that the alignment is successful at the edges of the unstructured regions.

Our threading application is a slight variant of the Needleman-Wunch algorithm, which finds the optimal global alignment between the structural positions along the backbone and a given amino acid sequence. The score of aligning a certain amino acid type against a structural position is the value of the SVM matrix at the corresponding column and row. Unlike the standard Needleman-Wunch algorithm, we do not allow gaps inserted into the threaded amino acid sequence, assuming that it is perfectly known. Likewise, we do not allow gaps inserted between structural positions along the backbone, unless a chain break was reported by the crystallographer. Within chain breaks, we actually want to enforce a gap and do it by a gap-opening reward of 20.0 that adds to the scoring function.

Figure 3 shows an example of successful threading by our threading algorithm. The sequence belonging to chain D of the 4A0C crystal structure was threaded onto the backbone at exactly the same register as reported in the original PDB deposit (Fischer et al., 2011). Note that this threading is not trivial, as the number of structured positions is smaller than the full length of the sequence. Therefore, there are many possible alignments, yet the CC-value signal that we compute is strong enough to determine the correct one. Next, we checked whether the sequence itself can be automatically inferred from the crystallographic data. To that end, we applied our threading application to all the sequences in the UniProt database (UniProt Consortium, 2017) that are longer than 89 residues (the number of structured positions in 4A0C_D). A total of 515,232 sequences were threaded and none of these gave a threading score that was higher than that of the correct sequence.

The threading of all the 331 chains in the 3.8–4.0-Å resolution range gave very encouraging results: our application threaded 83% of the chains exactly as reported in the PDB deposits. For an additional 9% of the chains, the correct threading was as reported in the PDB for more than 85% of the residues. In these cases, the mismatches in threading usually occurred at the termini of the structure. Only 8% of the chains performed poorly and threaded correctly only 40% of the residues on average. The threading of all the relevant UniProt sequences on each chain also gave very good results: for 84% of the chains, the best-scoring sequence was either the correct sequence or a very close homologous (sequence identity of more than 98%). The correct sequence was in the top 0.1%, 1%, and 10% of the best-scoring sequences for 89%, 94%, and 97% of the chains, respectively.

We also compared the threading performance for the three resolution ranges: 3.8–4.0 Å, 4.0–4.5 Å, and 4.5–5.0 Å. Respectively, our application threaded 83%, 81%, and 64% of the chains in each range exactly as reported in the PDB deposits. For 92%, 89%, 73% of the chains, the threading was as reported in the PDB for more than 85% of the residues. When threading all the relevant UniProt sequences, for 84%, 81%, and 45% of the chains in each range, the best-scoring sequence was the correct one. We conclude that very good threading results are obtained up to resolution of 4.5 Å, with a sharp performance drop at lower resolution (the 4.5–5.0 Å range). This is similar to the trend we observed for the SVM and *Z*-scores. Tables S4, S5, and S6 summarize the individual threading results for each chain in the three resolution ranges, respectively.

Predicting the Threading Accuracy

We showed that the threading application works perfectly for the great majority of the cases. Yet, its usefulness will greatly increase if we could be warned about those cases for which threading is less successful. To this end, we note that the threading score of our application is proportional to the chain length (Figure 4). We can therefore define for each chain a Threading Quality Index by dividing the total threading score by the chain length, i.e., the threading score per position. We see in Figure 4 that all the chains for which our threading application was less successful (red circles) are either short (<100 residues) or have very low Threading Quality Indices.

Figure 5 shows how the accuracy of threading improves with increase of the Threading Quality Index. We see that above a Threading Quality Index of 0.25 the threading accuracy is almost always better than 85% of the residues, and in most cases it is 100%. This is a stringent threshold, yet 72% of the chains in our 3.8–4.0-Å set are above it. We observe very similar results for threading in the 4.0–4.5-Å resolution range, where 60% of the chains have Threading Quality Indices above 0.25. Also for these chains, the threading accuracy was higher than 85% in all cases but one.

To further test the limits of our threading application, we reran eight of the PDB entries with backbone models that had 0.9–2.7 Å root-mean-square deviation from the deposited structures (Table S7). Six of these models were from homologous templates. Another two were the initial backbone models used by the crystallographers who deposited the structures. We were able to thread six of these models correctly, with more than 85% of the residues aligning with the PDB structures. Encouragingly, the two models that performed less well can be identified by their low threading scores. Both had Threading Quality Indices well below 0.25. They also performed more poorly in ranking the UniProt sequences. We can therefore state that

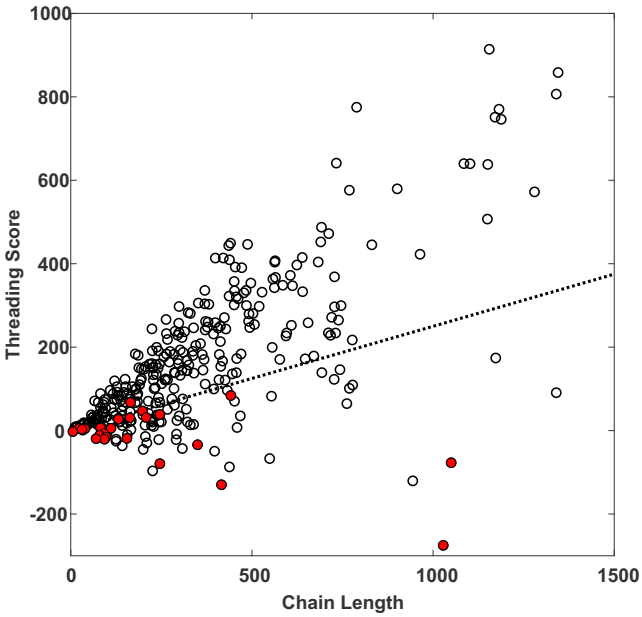


Figure 4. The Threading Scores of Each Chain in the 3.8–4.0-Å Resolution Range (Circles, 331 Chains) Are Proportional to the Chain Lengths

Full red circles mark chains for which less than 85% of the residues were threaded as in the PDB-deposited structures. These chains are either very short or show particularly weak CC-value signals. The line has a slope of 0.25 Score Units per position and 72% of the chains are above the line. All red circles are below this line except for one case. The line exemplifies our definition of the Threading Quality Index as threading score/chain length.

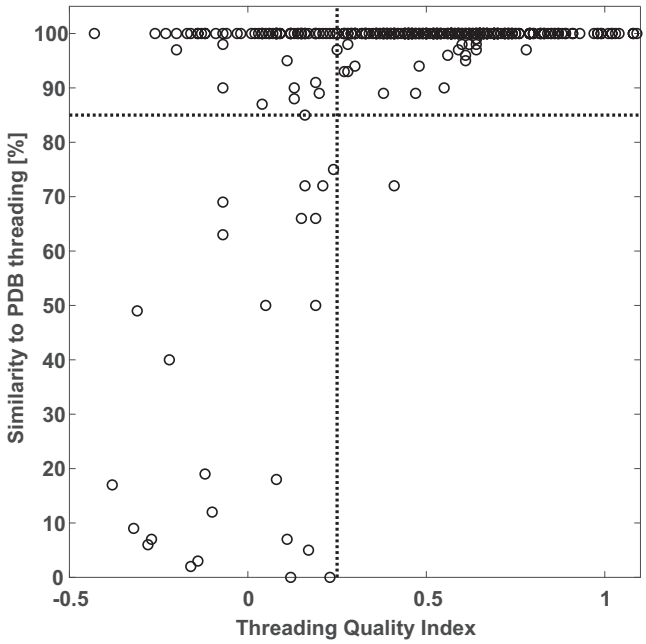


Figure 5. The Similarity (as Percentage of the Chain Length) between the Result of Our Threading Algorithm and the Threading Reported in the PDB Structures

This similarity is plotted for each chain as a function of the Threading Quality Index. When the Threading Quality Index is above 0.25 (dashed vertical line), the fit to the PDB threading is higher than 85% (in all except one case) and in most cases it is 100%.

crystallographers will get a reliable confidence estimation for the sequence assignment based on these measures.

Application on the CCT Chaperonin

The performance of our approach on the CCT chaperonin is featured here as an especially challenging case. The asymmetric unit of this crystallographic dataset (Dekker et al., 2011) comprises two particles with a total of 32 chains and 17,000 residues. Eight different genes make up the CCT, each occurring twice in each particle and four times in the entire asymmetric unit. The difficulty in assigning the sequences of the CCT arises from the high sequence and structural similarity between the subunits. We tested our application by threading all the sequences in UniProt onto each chain in the structure without imposing symmetry, i.e., we mutated a position only in one chain without altering its other three corresponding occurrences. Despite this very strict criterion, our application performed perfectly on all the chains with Threading Quality Indices of 0.11–0.26. The correct CCT gene was always the best-scoring UniProt sequence of more than 100,000. Next in rank were UniProt sequences of the correct CCT gene from other organisms (sequence identity of ~60%) followed by sequences of the other seven CCT genes from various organisms (sequence identity of ~30%). We believe that this success paves the way to apply our approach on other crystallographic structures with several homologous chains in the asymmetric units. Such cases are not unique to CCT and occur in other molecular systems, for example the AAA-ATPase ring module of the proteasome (Bohn et al., 2010). We also note that this method is much simpler than the exhaustive enumeration of all arrangements that was used previously to find the native CCT subunit order (Kalisman et al., 2013).

DISCUSSION

Figure 6 summarizes our automatic approach for sequence assignment. A software package that executes the complete workflow and guides the user via an easy-to-use graphic user interface is available for download from our website. We encourage crystallographers working with low-resolution datasets to simply use it on their initial all-ALA backbone model. If the resulting threading has a Quality Index above 0.25 and is high ranking against UniProt sequences, then it is very likely to be correct. This initial sequence assignment can then be further validated by manual or automated inspection. For example, tools used for quality assessment in the comparative modeling field (Kryshtafovych et al., 2018) can be used to independently identify small regions in which the initial alignment is incorrect. For cases with poorer scores, the threading may still be correct, but should be considered with caution. Rerunning our application at later stages of backbone refinement is then advisable.

Our approach shares many technical aspects with published sequence assignment tools that are widely used. These include the mutation of each position independently, the use of *Z* scores, and the threading of the sequence based on these

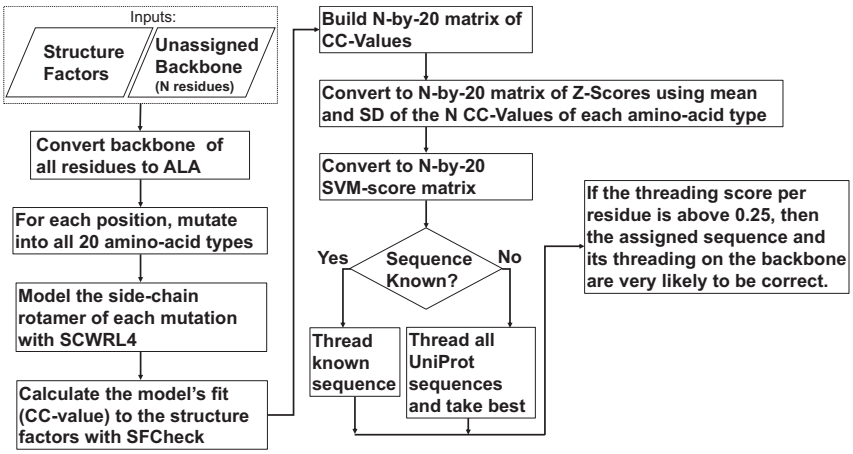


Figure 6. Workflow that Finds the Sequence Assignment to a Backbone with *N* Structural Positions

scores. We believe that the key to our success at lower resolutions is our reliance on reciprocal-space measures rather than real space correlations. In particular, these measures are not affected by the inaccuracy of the initial phases, which are more severe the lower the resolution is. Correlation coefficient in reciprocal space is a widely used measure in crystallography for global applications such as molecular replacement (Navaza, 1994) or heavy-atom localization (Weeks et al., 2003). Here, we show that a local and more subtle signal is also available in this measure. Accordingly, we suggest that our approach could be expanded to other local modeling tasks, such as completion of small loops or determination of ligand orientations.

The success of our approach on most chains raises the intriguing question of why some chains fail so drastically? Analysis of the cases for which the Threading Quality Index is below 0.15 shows that they include most of the chains that are shorter than 100 residues. However, chain length can only be partially blamed, since beside the short-chains there are much longer chains that also fail threading. Furthermore, we note that the Threading Quality Index is highly correlative with the initial *Z* scores calculated for each chain (Figure S1). The latter score is not concerned with threading at all and indicates that failure is due to issues arising from the way we process the crystallographic data. It appears that most chains on which threading failed originated from a very small set of ten PDB entries. We have indications that SFCheck has a numerical instability that degrades the accuracy of the CC value when run on several of the inputs, some of which are the ill-performing chains. We are currently working on a new program to calculate CC values using graphics processing units. We hope that it will resolve these numerical issues as well as speed up the runtime considerably.

We briefly discuss another intriguing issue concerning the correlations between the *Z*-scores revealed by the trained weights of the SVM classifiers (Table 3). For example, the MET/no-MET classifier does not give the *Z*-score of the MET mutation more weight than the CYS, HIS, LEU, GLN, THR, and VAL mutations. These couplings raise the possibility that the 20 natural amino acids may not be the best basis set to extract sequence information from the crystallographic data. Previously, Holton et al. (2000) suggested the use of statistical moments of the electron density as features to infer the residue types. These, and possibly other features, may be eventually added to our SVM classifiers and increase their discriminative power. Preliminary tests have shown that neural net machine learning can provide an even better classifier; this too may be the subject of a future study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Compilation of PDB Datasets
  - Training of Support Vector Machine (SVM) Classifiers
  - Global Alignment Algorithm for Threading
- QUANTIFICATION AND STATISTICAL ANALYSIS
- SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes one figure and seven tables and can be found with this article online at https://doi.org/10.1016/j.str.2018.08.011.

ACKNOWLEDGMENTS

Z.B.-A. and N.K. were supported by the Israel Science Foundation grant number 1768/15. M.L., who is the Robert W. and Vivian K. Cahill Professor of Cancer Research, was supported by NIH award GM1157490. Computation was carried out on the BioX3 cluster supported by NIH S10 Shared Instrumentation Grant 1S10RR02664701). We thank Drs. Losic and Rass for sharing with us their initial crystallographic backbone models.

AUTHOR CONTRIBUTIONS

Conceptualization, M.L. and N.K.; Methodology, M.L. and N.K.; Software, Z.B.-A. and N.K.; Investigation, Z.B.-A., M.L., and N.K.; Data Curation, Z.B.-A. and N.K.; Writing – Original Draft, M.L. and N.K.; Writing – Review & Editing, Z.B.-A., N.K., and M.L.; Visualization, Z.B.-A. and N.K.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 6, 2017
Revised: April 18, 2018
Accepted: August 23, 2018
Published: October 4, 2018

REFERENCES

Aller, S.G., Yu, J., Ward, A., Weng, Y., Chittaboina, S., Zhuo, R., Harrell, P.M., Trinh, Y.T., Zhang, Q., Urbatsch, I.L., and Chang, G. (2009). Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. Science *323*, 1718–1722.

Bae, B., Davis, E., Brown, D., Campbell, E.A., Wigneshweraraj, S., and Darst, S.A. (2013). Phage T7 Gp2 inhibition of *Escherichia coli* RNA polymerase involves misappropriation of σ70 domain 1.1. Proc. Natl. Acad. Sci. USA *110*, 19772–19777.

Bohn, S., Beck, F., Sakata, E., Walzthoeni, T., Beck, M., Aebersold, R., Förster, F., Baumeister, W., and Nickell, S. (2010). Structure of the 26S proteasome from *Schizosaccharomyces pombe* at subnanometer resolution. Proc. Natl. Acad. Sci. USA *107*, 20992–20997.

Cohen, S.X., Morris, R.J., Fernandez, F.J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V.S., Kleywegt, G.J., and Perrakis, A. (2004). Towards complete validated models in the next generation of ARP/wARP. Acta Crystallogr. D Biol. Crystallogr. *60*, 2222–2229.

Cowtan, K. (2008). Fitting molecular fragments into electron density. Acta Crystallogr. DD Biol. Crystallogr. *64*, 83–89.

Dekker, C., Roe, S.M., McCormack, E.A., Beuron, F., Pearl, L.H., and Willison, K.R. (2011). The crystal structure of yeast CCT reveals intrinsic asymmetry of eukaryotic cytosolic chaperonins. EMBO J. *30*, 3078–3090.

Dunbrack, R.L., Jr. (2006). Sequence comparison and protein structure prediction. Curr. Opin. Struct. Biol. *16*, 374–384.

Fischer, E.S., Scrima, A., Böhm, K., Matsumoto, S., Lingaraju, G.M., Faty, M., Yasuda, T., Cavadini, S., Wakasugi, M., Hanaoka, F., et al. (2011). The molecular basis of CRL4DDB2/CSA ubiquitin ligase architecture, targeting, and activation. Cell *147*, 1024–1039.

Fleishman, S.J., Unger, V.M., Yeager, M., and Ben-Tal, N. (2004). A Calpha model for the transmembrane alpha helices of gap junction intercellular channels. Mol. Cell *15*, 879–888.

Forneris, F., Ricklin, D., Wu, J., Tzekou, A., Wallace, R.S., Lambris, J.D., and Gros, P. (2010). Structures of C3b in complex with factors B and D give insight into complement convertase formation. Science *330*, 1816–1820.

Holton, T., Ioerger, T.R., Christopher, J.A., and Sacchettini, J.C. (2000). Determining protein structure from electron-density maps using pattern matching. Acta Crystallogr. D Biol. Crystallogr. *56*, 722–734.

Kalisman, N., Schröder, G.F., and Levitt, M. (2013). The crystal structures of the eukaryotic chaperonin CCT reveal its functional partitioning. Structure *21*, 540–549.

Karakas, E., and Furukawa, H. (2014). Crystal structure of a heterotetrameric NMDA receptor ion channel. Science *344*, 992–997.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795.

Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. (2018). Assessment of model accuracy estimations in CASP12. Proteins *86S1*, 345–360.

Levitt, M. (1983). Molecular dynamics of native protein: I. computer simulation of trajectories. J. Mol. Biol. *168*, 595–620.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. *226*, 507–533.

Levitt, M., Hirshberg, M., Sharon, R., and Daggett, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput. Phys. Commun. *91*, 215–231.

Lu, M., Symersky, J., Radchenko, M., Koide, A., Guo, Y., Nie, R., and Koide, S. (2013). Structures of a Na+-coupled, substrate-bound MATE multidrug transporter. Proc. Natl. Acad. Sci. USA *110*, 2099–2104.

Maeda, S., Nakagawa, S., Suga, M., Yamashita, E., Oshima, A., Fujiyoshi, Y., and Tsukihara, T. (2009). Structure of the connexin 26 gap junction channel at 3.5 A resolution. Nature *458*, 597–602.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J. Appl. Crystallogr. *40*, 658–674.

Muñoz, I.G., Yébenes, H., Zhou, M., Mesa, P., Serna, M., Park, A.Y., Bragado-Nilsson, E., Beloso, A., de Cárcer, G., Malumbres, M., et al. (2011). Crystal structure of the open conformation of the mammalian chaperonin CCT in complex with tubulin. Nat. Struct. Mol. Biol. *18*, 14–19.

Navaza, J. (1994). AMoRe: an automated package for molecular replacement. Acta Crystallogr. A *50*, 157–163.

Needleman, S.B., and Wunch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

Özkan, E., Chia, P.H., Wang, R.R., Goriatcheva, N., Borek, D., Otwinowski, Z., Walz, T., Shen, K., and Garcia, K.C. (2014). Extracellular architecture of the SYG-1/SYG-2 adhesion complex instructs synaptogenesis. Cell *156*, 482–494.

Shaya, D., Findeisen, F., Abderemane-Ali, F., Arrigoni, C., Wong, S., Nurva, S.R., Loussouarn, G., and Minor, D.L., Jr. (2014). Structure of a prokaryotic sodium channel pore reveals essential gating elements and an outer ion binding site common to eukaryotic channels. J. Mol. Biol. *426*, 467–483.

Terwilliger, T.C. (2003). Automated side-chain model building and sequence assignment by template matching. Acta Crystallogr. D Biol. Crystallogr. *59*, 45–49.

UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45*, D158–D169.

Vaguine, A.A., Richelle, J., and Wodak, S.J. (1999). SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr. D Biol. Crystallogr. *55*, 191–205.

Wang, C., Chung, B.C., Yan, H., Wang, H.G., Lee, S.Y., and Pitt, G.S. (2014). Structural analyses of Ca^{2+}/CaM interaction with NaV channel C-termini reveal mechanisms of calcium-dependent regulation. Nat. Commun. *5*, 4896.

Weeks, C.M., Adams, P.D., Berendzen, J., Brunger, A.T., Dodson, E.J., Grosse-Kunstleve, R.W., Schneider, T.R., Sheldrick, G.M., Terwilliger, T.C., Turkenburg, M.G., et al. (2003). Automatic solution of heavy-atom substructures. Methods Enzymol. *374*, 37–83.

Wei, Q., and Dunbrack, R.L., Jr. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One *8*, e67863.

Zou, J.Y., and Jones, T.A. (1996). Towards the automatic interpretation of macromolecular electron-density maps: qualitative and quantitative matching of protein sequence to map. Acta Crystallogr. D Biol. Crystallogr. *52*, 833–841.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
171 PDB entries in the 3.8-4.0 Angs range	Protein Data Bank	PDB identifiers are listed in Table S4
113 PDB entries in the 4.0-4.5 Angs range	Protein Data Bank	PDB identifiers are listed in Table S5
28 PDB entries in the 4.5-5.0 Angs range	Protein Data Bank	PDB identifiers are listed in Table S6
Software and Algorithms		
Automatic Sequence Threader	This manuscript	http://biolchem.huji.ac.il/nirka/software.html
SCWRL4	Krivov et al., 2009	http://dunbrack.fccc.edu/scwrl4/
SFCheck	Vaguine et al., 1999	http://www.ccp4.ac.uk/html/sfcheck.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Nir Kalisman (nirka@mail.huji.ac.il).

METHOD DETAILS

Compilation of PDB Datasets

For this study we compiled three sets of PDB entries in the following resolution ranges: 3.8 - 4.0Å, 4.0 - 4.5Å, and 4.5 - 5.0Å. Each set was downloaded from the PDB website with the following search criteria: (i) solved by X-ray crystallography in the relevant resolution range; (ii) less than 70% sequence identity to selected set; (iii) deposited during the years 2010-2017. We further discarded entries for which: (i) the entry contained large number of nucleic acid nucleotides (mostly ribosome structures); (ii) the entry contained a chain shorter than 18 residues; (iii) the entry contained a chain with an all-UNK annotation. Overall, 173, 113, and 28 PDB entries were compiled for the three resolution ranges, respectively. In the asymmetric unit of each PDB entry we identified all repeating chains with identical sequences. These chains were treated identically, i.e. an amino-acid mutation modeled onto a position in one chain was also modeled onto all the corresponding positions of the identical chains before any reciprocal-space calculation. If a certain position did not appear in all the identical chains, it was discarded from the analysis. Overall, we had 331, 207, and 69 chains (or groups of identical chains) for the three resolution ranges, respectively.

Training of Support Vector Machine (SVM) Classifiers

We trained 20 SVM two-state classifiers - one for each amino-acid type. Each classifier was trained to determine if a certain structural position is either a certain amino-acid or anything except that amino acid (e.g. TRP/non-TRP). This is a highly imbalanced classification problem with the positive class being on average only 5% of the cases. Wei and Dunbrack (2013) have shown that classifiers trained on balanced sets perform better when applied to general test data. Accordingly, we generated a balanced set of structural positions for the training and testing of each classifier from a subset of 211 chains in the 3.8-4.0Å resolution range. That set included all the positions matching the classified amino-acid type and an equal number of random positions that were not the amino-acid. Half of the positions in the set were used for training and the other half for testing. We trained the classifiers in MATLAB using the 'fitcsvm' function with the default settings: a linear classifier with a constant penalty cost for misclassifications.

Global Alignment Algorithm for Threading

Our threading application is based on the Needleman-Wunch algorithm for optimal global sequence alignment (Needleman and Wunch, 1970). We prohibit gap opening in the amino-acid sequence by assigning a gap-opening penalty of minus infinity. Likewise, gaps in the structural position sequence are prohibited by a gap-opening penalty of minus infinity, except at chain-breaks as they are reported for the crystallographic backbone. At chain-breaks we enforce gap opening by assigning gap-opening reward of +20.0 that is added to the scoring function. The use of a positive gap-opening rewards requires a small variant on the Needleman-Wunch algorithm: one must check that a gap is not re-opened (because of the positive sign) inside another gap. This is done during the building of the score matrix, by remembering if the optimal path that led to a matrix cell is currently 'inside a gap'. The gap-extension penalty is set to -0.25. After the global alignment has been found, the gap rewards (i.e. 20.0 * number_of_gaps – 0.25 * length_of_gaps) are subtracted from the threading score. This allows comparison of scores between PDB structures with different number of structural gaps. We used the UniProt compilation of 548,890 protein sequences to test the threading algorithm against unrelated sequences.

QUANTIFICATION AND STATISTICAL ANALYSIS

The findings are based on analysis of data from three sets of PDB entries in three resolution ranges: 3.8-4.0Å - 331 chains from 171 PDB entries ; 4.0-4.5Å - 207 chains from 113 PDB entries ; 4.0-4.5Å - 69 chains from 28 PDB entries. Where appropriate, statistical details are given in the table legends.

SOFTWARE AVAILABILITY

The application with a GUI of the complete threading process can be downloaded at http://biolchem.huji.ac.il/nirka/software.html. The application runs on UNIX platforms with an average running time of 12 hours per chain. The downloadable package also includes the SFCheck executable for calculating CC-Values (it has been modified to include more significant digits in the output). The package does not include SCWRL4, which can be downloaded upon request from the Dunbrack Lab and installed separately.