Physical Biology

OFFICIAL JOURNAL OF
Sociedad de Biofísicos Latino Americanos

PAPER

# Statistical radii associated with amino acids to determine the contact map: fixing the structure of a type I cohesin domain in the *Clostridium thermocellum* cellulosome

To cite this article: Mateusz Chwastyk *et al* 2015 *Phys. Biol.* **12** 046002

IOP Publishing Phys. Biol. 12 (2015) 046002 doi:10.1088/1478-3975/12/4/046002

Physical Biology

RECEIVED
27 January 2015

REVISED
15 March 2015

ACCEPTED FOR PUBLICATION
10 April 2015

PUBLISHED
27 May 2015

PAPER

# Statistical radii associated with amino acids to determine the contact map: fixing the structure of a type I cohesin domain in the *Clostridium thermocellum* cellulosome

**Mateusz Chwastyk, Adolfo Poma Bernaola and Marek Cieplak**[1]

Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46, 02-668 Warsaw, Poland
[1] Author to whom any correspondence should be addressed.

**E-mail:** mc@ifpan.edu.pl

**Keywords:** cohesin, refinement, cellulosome, coarse-grained models, structure-based models

Supplementary material for this article is available online

## Abstract

We propose to improve and simplify protein refinement procedures through consideration of which pairs of amino acid residues should form native contacts. We first consider 11 330 proteins from the CATH database to determine statistical distributions of contacts associated with a given type of amino acid. The distributions are set across the distances between the $\alpha$-C atoms that are in contact. Based on this data, we determine typical radii of effective spheres that can be placed on the $\alpha$-C atoms in order to reconstruct the distribution of the contact lengths. This is done by checking for overlaps with enlarged van der Waals spheres associated with heavy atoms on other amino acids.

The resulting contacts can be used to identify non-native contacts that may arise during the time evolution of structure-based models. Here, the radii are used to guide reconstruction of nine missing side chains in a type I cohesin domain with the Protein Data Bank code 1AOH. We first identify the likely missing contacts and then sculpt the corresponding side chains by standard refinement tools to achieve consistency with the expected contact map. One ambiguity in refinement is resolved by determining all-atom conformational energies.

## 1. Introduction

Experimentally determined protein structures are often incomplete. They may come with absent fragments of their backbones or with missing atomic coordinates of the side groups. The specifics of the all-atom coordinates are of importance in many applications and, in particular, in the determination of the contact map, which defines which pairs of amino acid (denoted as AAs for short) may bind non-covalently. In structure-based coarse-grained models [1–8], the contacts play a dynamical role instead of being merely descriptive. The prediction of the positions of the side groups is a subject of many refinement methods [9–12]. Sometimes, however, these methods are not sufficient, because seemingly subtle adjustments in the orientation of the predicted side chain may give rise to different sets of the native contacts. Thus the determination of the contacts should also be taken into account in the refinement process.

Here, we propose a method to address this problem through a statistics-based determination of the optimal native contact map. The determination of the native contact map in regions involving the defective AAs may restrict the rotamer search in the refinement procedure applied to AAs in such contacts. Single-length cutoff-based criteria for what makes a native contact are too simple to provide an adequate account of the dynamics [13], because they introduce many spurious interactions at short lengths and fail to include many important couplings at longer lengths.

Another way to construct a contact map—and one which is much more relevant for the dynamics—is by studying overlaps of spheres associated with the heavy atoms of AAs. The sizes of such spheres are given by the van der Waals radii [14] multiplied by a factor of 1.24 to account for attraction [15, 16]—they identify inflection points in an associated van der Waals potential (see table 1 in the supplementary information (stacks.iop.org/pb/12/046002/mmedia)).

**Table 1.** The values of $R_{\alpha C}$ and $R_{\beta C}$ obtained by the statistical method are listed in the second and fourth column, respectively. The first column gives the name of the amino acid. The third and fifth columns provide the values of the parameter $S$ at the optimal matching. The last column provides the statistics of contacts in the full CATH structures.

| AA | $R_{\alpha C}$ | $S$ | $R_{\beta C}$ | $S$ | $N_c$ |
|---|---|---|---|---|---|
| GLY | **3.15** | 392.945 | — | — | 91300 |
| ALA | **3.35** | 590.941 | — | — | 151159 |
| SER | **3.30** | 208.813 | 2.65 | 28.1816 | 85907 |
| ASP | **3.50** | 336.216 | 3.15 | 71.8116 | 79177 |
| THR | **3.60** | 668.438 | 3.20 | 86.8286 | 96070 |
| ASN | **3.65** | 417.796 | 3.45 | 167.931 | 63208 |
| GLU | **3.65** | 665.146 | 3.55 | 285.902 | 103779 |
| LYS | **3.65** | 830.24 | 3.75 | 503.244 | 93602 |
| CYS | **3.70** | 230.31 | 3.25 | 90.2319 | 33040 |
| PRO | **3.70** | 596.032 | 3.50 | 188.403 | 65146 |
| GLN | **3.90** | 580.809 | 3.80 | 295.567 | 65047 |
| ARG | **3.95** | 1304.19 | 4.00 | 949.975 | 103114 |
| VAL | **4.00** | 1762.91 | 3.55 | 389.079 | 173342 |
| HIS | **4.00** | 604.623 | 4.00 | 347.322 | 45529 |
| ILE | **4.50** | 1833.15 | 3.95 | 679.089 | 156723 |
| MET | **4.50** | 741.289 | 4.25 | 448.451 | 55988 |
| TYR | **4.50** | 1499.52 | 4.30 | 1084.74 | 94188 |
| LEU | **4.60** | 2545.91 | 4.15 | 1189.12 | 241137 |
| PHE | **4.60** | 1931.98 | 4.55 | 1400.32 | 113959 |
| TRP | **4.70** | 824.557 | 4.75 | 637.204 | 43447 |

These enlarged spheres will be referred to as effective van der Waals spheres. Two AAs are said to be in a native contact if they have at least one overlapping pair of the effective spheres associated with the heavy atoms belonging to the two AAs. We have checked (see section 4) that this van der Waals-based criterion captures most of the specific (i.e., non-dispersive) contacts that can be obtained from the CSU server [17], which involves considerations that are more chemical in nature. In particular, the CSU approach involves checking whether one can place a molecule of water between the pair of AAs under study.

If the coordinates of the heavy atoms beyond $\beta$-C are missing, one may ask: what is the most likely radius, $R_{\beta C}$, of an effective sphere that can be associated with the $\beta$-C of the corresponding AA to identify the native contacts correctly? Notice that the contacts may arise both from the side chain and from the backbone. One may disregard the distinction and ask a simpler question: what is the most likely radius, $R_{\alpha C}$, of an effective sphere that can be associated with the $\alpha$-C to get the right set of the native contacts? We answer these two questions by considering 11 330 fully resolved proteins that represent the CATH database [24]. Once the values of $R_{\alpha C}$ and $R_{\beta C}$ are determined, we use them to establish contacts with the defective AAs. Finally, we perform refinement within these contacts while keeping their environments frozen.

In this paper, we illustrate applications of the statistical method by considering a type I cohesin domain with the Protein Data Bank (PDB) structure code 1AOH [18]. This protein is one of the particularly stable modules found in the *Clostridium thermocellum* cellulosome [19–21]. Its large mechanical stability has been demonstrated by Valbuena *et al* [22] through single-molecule force spectroscopy. The sequential length of the chains in 1AOH is 147. The resolved structure of chain A of 1AOH has no $\alpha$-C atoms in the first four AAs (at the N-terminus) and no coordinates of the side-chain atoms beyond $\beta$-C in nine AAs: 9-LYS, 16-LYS, 47-ASN, 50-GLU, 53-GLU, 77-ARG, 110-SER, and 136-LYS, 138-GLN. In the sequentially identical chain B, no $\alpha$-Cs are missing, but there are still nine defective side groups whose list partially overlaps that for chain A.

Here, we reconstruct locations of the missing atoms in 1AOH by employing the statistical method, combined with the usage of standard software tools, to design the side group conformation to an accuracy that is sufficient for molecular dynamics studies within coarse-grained models. The tools involved are MODELLER [9], which constructs a viable side group, and the sculpting tool from the Pymol software [23] to adjust the orientation of the side group to obtain consistency with the presence of the expected contacts. We show that the procedure is essentially unique for eight missing residues in 1AOH. However, the ninth one, 9-LYS, may adopt one of two alternative orientations, and additional criteria are needed to make the choice. We settle the issue by comparing chain A to chain B and through small-scale molecular dynamics simulations. The simulations would be needed if there ware no chain for comparison.

The $R_{\alpha C}$ derived here can be valuable in molecular dynamics studies of proteins within coarse-grained structure-based models: they can be used to identify non-native contacts that may arise during time evolution of structure-based models.

## 2. The statistical method

In order to determine $R_{\alpha C}$s and $R_{\beta C}$s, we derive probability distributions of contacts arising at a distance of $l_{ij}$ between two $\alpha$-C atoms $i$ and $j$, provided at least one of the AAs is of a given type, say, arginine. The distributions are obtained by using the set of 11 330 PDB complete structures. Figure 1 gives examples

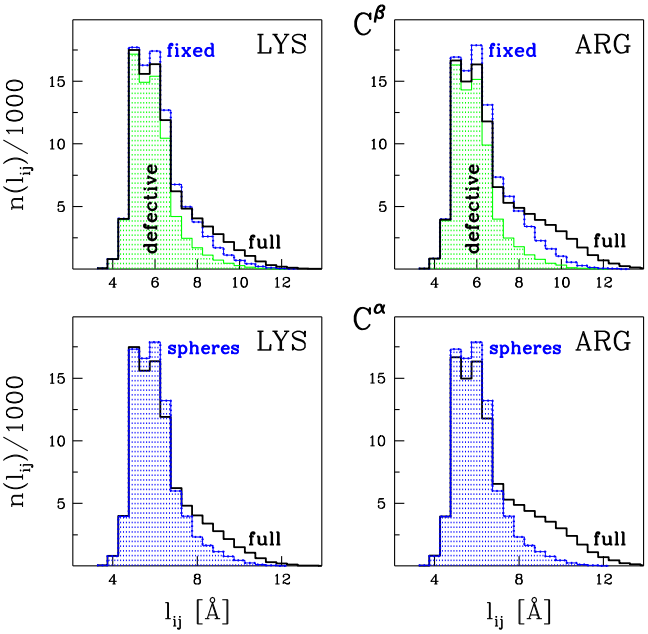


**Figure 1.** The top two panels illustrate the determination of the optimal size, $R_{\beta C}$, of a sphere associated with the $\beta$-C atoms for LYS and ARG. The thickest solid lines (in black) in the panels correspond to the undoctored distributions of contacts with the AA studied. The thin solid lines in green, delimiting the shaded areas, correspond to the distributions of contacts obtained when atoms beyond $\beta$-C on all AAs of the type studied are removed. The beaded lines in blue show the distribution obtained for the optimal selection of $R_{\beta C}$ (as listed in table 1). The bottom two panels illustrate the determination of the optimal size, $R_{\alpha C}$, of a sphere associated with the $\alpha$-C atoms for LYS and ARG. The thick solid lines are the same as in the corresponding top panels. The beaded lines in blue (denoted by 'spheres') show the optimal reconstructions of the distributions.

of such distributions for arginine and lysine (results for other AAs are presented in the SI). These are the histograms which are drawn by heavy solid lines and are denoted by the label 'full'. They are non-Gaussian, as they come with a substantial tail at large values of $l_{ij}$.

In order to determine $R_{\beta C}$ we remove the coordinates of the heavy atoms beyond $\beta$-C in all AAs of a given type in the set (say, VAL). Similarly, for calculation of $R_{\alpha C}$, we remove all side chains of the studied type of AA. This step necessarily depletes the numbers of detected contacts, as shown in the top two panels of figure 1 in the context of $R_{\beta C}$ (the parts that are shaded). We now represent the missing atoms on all crippled AAs by spheres of the tentative radius $r_{\beta C}$ and centered at the $\beta$-C atoms or by spheres of the tentative radius $r_{\alpha C}$ and centered at the $\alpha$-C atoms. We check for overlaps of the spheres with the effective van der Waals spheres on the remaining types of residues (or with another tentative sphere if the contact is within the same kind of residue). We keep adjusting $r_{\beta C}$ or $r_{\alpha C}$ until the corresponding distribution of contacts matches the original distribution as closely as possible. The optimal situation for any target AA is identified by minimizing

$$S^2 = \frac{1}{N_b}\sum_{m=1}^{N_b}\left(n_m - n'_m\right)^2, \qquad (1)$$

where $N_b$ is the number of bins in the histogram and $n_m$ is the number of contacts in an $m$th bin in the original full distribution, whereas $n'_m$ is the number of observations in the same bin with the approximate rendering of the missing side-chain atoms. $R_{\alpha C}$ and $R_{\beta C}$ are defined as the optimal values of $r_{\alpha C}$ or $r_{\beta C}$, respectively. They are listed in table 1. It should be pointed out that equation (1) deals with the distributions of the numbers of contacts and not the normalized probability distributions, because each choice of $r_{\beta C}$ or $r_{\alpha C}$ comes with a different total number of contacts. The goal here is to optimize not only the shape of the distribution but also the similarity in the actual number of contacts. In principle, the function $S^2$ defined in equation (1) depends on the bin width—the basic value used here is 0.5 Å. However, reducing the width by a factor of 2 is found not to affect the radii listed in table 1. Doubling the width has only a 1% effect. If one modifies the definition of $S$ so that it involves $|n_m - n'_m|$, then the results change within 2%, which may be considered as providing the error bars on the results.

We now consider 1AOH. Chain A turns out to be more interesting to focus on, as there are somewhat different results, depending on whether one adds the structure in the missing 1–4 segment or not. We begin by adding the segment. This can be accomplished by taking it from chain B and attaching it to the original A by translation and small adjustments implemented through the use of MODELLER. There are no defects in this segment. We then replace the $\alpha$-Cs on the defective residues (i.e., with the missing side chains) by spheres with the radii listed in table 1 and check for overlaps with all other AAs. In this way, we identify the

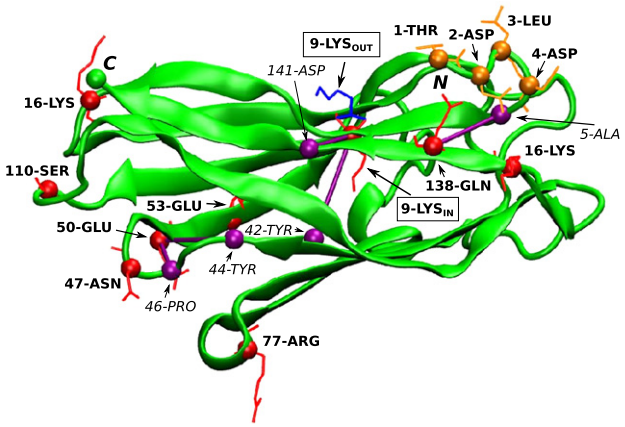


**Figure 2.** The native state of protein 1AOH after the reconstruction. Reconstructed side chains are displayed in red. The side chains on residues 1–4 are shown in orange. The two possible orientations for 9-LYS are marked in red (IN) and blue (OUT). Choice OUT is consistent with having the 9-LYS–141-ASP contact. Choice IN would be consistent with two additional predicted contacts (9-LYS–141-ASP and 9-LYS–42-TYR)–shown here as straight purple lines—and choice OUT only with one: 9-LYS–141-ASP. The remaining predicted contacts (5-ALA–138-GLN, 44-TYR–50-GLU, and 46-PRO–50-GLU) are also shown here as straight purple lines, and they are the same for both the IN and OUT states.

**Table 2.** The coordinates ($x$, $y$, and $z$) of the heavy atoms in the native state of 9-LYS in 1AOH. The subscript N refers to the native coordinates as available in the file. The subscripts IN and OUT refer to the two rotamers discussed. OUT is considered to be the most likely solution. The values highlighted in boldface are either those predicted through the statistical method combined with sculpting or those that needed small adjustments.

| 9-LYS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Atom | $x_N$ | $y_N$ | $z_N$ | $x_{IN}$ | $y_{IN}$ | $z_{IN}$ | $x_{OUT}$ | $y_{OUT}$ | $z_{OUT}$ |
| N | 24.487 | 28.756 | 46.073 | 24.487 | 28.756 | 46.073 | 24.487 | 28.756 | 46.073 |
| CA | 25.469 | 29.026 | 47.106 | **25.264** | **30.366** | **45.906** | 25.469 | 29.026 | 47.106 |
| C | 26.830 | 29.382 | 46.548 | 26.830 | 29.382 | 46.548 | 26.830 | 29.382 | 46.548 |
| O | 27.308 | 28.761 | 45.599 | 27.308 | 28.761 | 45.599 | 27.308 | 28.761 | 45.599 |
| CB | 25.578 | 27.835 | 48.031 | **24.877** | **31.288** | **46.389** | 25.578 | 27.835 | 48.031 |
| CG | | | | **23.228** | **32.013** | **46.569** | **26.584** | **27.818** | **49.118** |
| CD | | | | **22.877** | **33.337** | **47.421** | **26.712** | **29.002** | **50.032** |
| CE | | | | **22.870** | **34.684** | **48.488** | **27.301** | **28.608** | **51.383** |
| NZ | | | | **22.521** | **35.747** | **48.012** | **26.956** | **29.610** | **52.415** |

contacts which are very likely to be missing. These are: 5-ALA–138-GLN, 9-LYS–42-TYR, 9-LYS–141-ASP, 44-TYR–50-GLU, and 46-PRO–50-GLU.

In order to obtain the actual structures of the defective side chains we have used MODELLER [9] and then the Pymol sculpting tool [23] to construct the tentative side chains and then guide them toward positions that would be consistent with the predicted contact map. The idea is to sculpt the side chains so that the contacts that are likely to arise are actually built in. At this stage, we use the overlap criterion with the effective van der Waals spheres representing all heavy atoms. It is straightforward to set the corresponding rotamers in a unique way to generate these contacts, except for 9-LYS, for which there are two possibilities: either the side group points into the body of the protein or away from it. The two orientations are illustrated in figure 2 and are denoted as IN and OUT, respectively. The reconstruction corresponding to choice IN required making slight adjustments in the positioning of the $\alpha$-C and $\beta$-C atoms. No such adjustments turned out to be needed for choice OUT and for the other eight defective AAs, except for $\beta$-C at 50-GLU. The resulting refined structures are shown in tables 2 and 3. The coordinates of other atoms are listed in the PDB file for 1AOH.

When we repeat the procedure using $\beta$-C, we replace the $\beta$-Cs on the defective residues by spheres with the radii listed in table 1 but also assign the effective van der Waals spheres to the backbone atoms: N, $\alpha$-C, C, and O. Checking for overlaps with all other AAs leads to the identification of just one missing contact: 5-ALA–138-GLN. Enlargement of $R_{\beta C}$ by 10% (to account for the tails in the in the distributions of the effective radii) brings in the mechanically important 9-LYS–141-ASP contact. One can do proper refinement with these two contacts, but, clearly, the method based on the $\alpha$-C atoms appears to be much

**Table 3.** Similar to table 2 but for the remaining eight defective residues. In the case of 50-GLU, the original position of $\beta$-C is $x = 40.986$, $y = 42.661$, and $z = 41.841$.

| 1-THR | | | | 2-ASP | | | | 3-LEU | | | | 4-ASP | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Atom | $x$ | $y$ | $z$ | atom | $x$ | $y$ | $z$ | atom | $x$ | $y$ | $z$ | atom | $x$ | $y$ | $z$ |
| N | **17.295** | **24.714** | **54.855** | N | **15.444** | **24.596** | **51.961** | N | **13.836** | **22.084** | **50.484** | N | **12.144** | **22.368** | **48.270** |
| CA | **17.370** | **24.326** | **53.397** | CA | **14.126** | **24.297** | **51.469** | CA | **13.903** | **21.100** | **49.389** | CA | **11.320** | **22.756** | **47.138** |
| C | **15.967** | **23.928** | **52.967** | C | **14.223** | **23.336** | **50.275** | C | **13.051** | **21.413** | **48.148** | C | **11.885** | **23.997** | **46.448** |
| O | **15.404** | **22.985** | **53.522** | O | **14.616** | **23.729** | **49.171** | O | **13.228** | **20.814** | **47.091** | O | **11.285** | **24.528** | **45.506** |
| CB | **18.359** | **23.258** | **53.114** | CB | **13.475** | **25.635** | **51.110** | CB | **13.628** | **19.685** | **49.893** | CB | **9.892** | **23.000** | **47.607** |
| CG2 | **18.111** | **22.588** | **51.742** | CG | **12.062** | **25.508** | **50.611** | CG | **14.637** | **19.262** | **50.953** | CG | **9.243** | **21.735** | **48.145** |
| OG1 | **19.673** | **23.826** | **53.195** | OD1 | **11.503** | **24.397** | **50.522** | CD1 | **14.319** | **17.889** | **51.459** | OD1 | **9.307** | **20.677** | **47.444** |
| | | | | OD2 | **11.502** | **26.561** | **50.264** | CD2 | **16.031** | **19.320** | **50.385** | OD2 | **8.708** | **21.790** | **49.282** |

| 16-LYS | | | | 47-ASN | | | | 50-GLU | | | | 53-GLU | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| atom | $x$ | $y$ | $z$ | atom | $x$ | $y$ | $z$ | atom | $x$ | $y$ | $z$ | atom | $x$ | $y$ | $z$ |
| N | 45.376 | 33.602 | 50.616 | N | 37.812 | 46.450 | 47.480 | N | 39.749 | 41.629 | 43.674 | N | 37.759 | 37.453 | 36.359 |
| CA | 46.324 | 34.480 | 51.250 | CA | 39.037 | 46.908 | 48.118 | CA | 39.646 | 42.212 | 42.340 | CA | 36.890 | 36.839 | 35.357 |
| C | 46.407 | 35.693 | 50.359 | C | 39.843 | 45.689 | 48.595 | C | 39.104 | 41.097 | 41.458 | C | 35.845 | 35.998 | 36.057 |
| O | 46.141 | 35.580 | 49.164 | O | 41.014 | 45.809 | 48.912 | O | 39.741 | 40.062 | 41.287 | O | 36.000 | 35.654 | 37.222 |
| CB | 47.689 | 33.796 | 51.306 | CB | 38.689 | 47.830 | 49.299 | **CB** | **39.034** | **43.344** | **42.977** | CB | 37.728 | 35.920 | 34.420 |
| CG | **47.660** | **32.653** | **52.323** | CG | **38.099** | **49.143** | **48.781** | CG | **38.342** | **44.055** | **42.549** | CG | **37.517** | **35.070** | **34.940** |
| CD | **47.237** | **31.353** | **51.637** | ND2 | **38.811** | **49.712** | **47.813** | CD | **37.501** | **44.561** | **41.415** | CD | **36.657** | **34.483** | **35.982** |
| CE | **46.920** | **30.286** | **52.687** | OD1 | **37.065** | **49.610** | **49.232** | OE1 | **37.441** | **45.720** | **41.010** | OE1 | **36.199** | **33.527** | **36.033** |
| NZ | **46.793** | **28.961** | **52.040** | | | | | OE2 | **36.872** | **43.690** | **40.957** | OE2 | **36.461** | **35.207** | **37.126** |

**Table 3.** (Continued.)

| 77-ARG | | | | 110-SER | | | | 136-LYS | | | | 138-GLN | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| atom | *x* | *y* | *z* | atom | *x* | *y* | *z* | atom | *x* | *y* | *z* | atom | *x* | *y* | *z* |
| N | 29.821 | 50.270 | 42.439 | N | 47.060 | 42.382 | 47.691 | N | 11.048 | 28.776 | 42.968 | N | 16.598 | 30.011 | 47.009 |
| CA | 28.621 | 49.897 | 43.191 | CA | 48.217 | 42.171 | 48.554 | CA | 11.389 | 29.327 | 44.290 | CA | 17.124 | 30.310 | 48.321 |
| C | 28.808 | 48.583 | 43.976 | C | 48.139 | 42.686 | 49.988 | C | 12.904 | 29.138 | 44.412 | C | 18.602 | 30.525 | 48.022 |
| O | 27.894 | 48.121 | 44.659 | O | 48.734 | 42.085 | 50.884 | O | 13.466 | 28.147 | 43.923 | O | 19.293 | 29.612 | 47.520 |
| CB | 27.389 | 49.832 | 42.285 | CB | 49.459 | 42.649 | 47.892 | CB | 10.638 | 28.556 | 45.422 | CB | 16.888 | 29.125 | 49.286 |
| CG | **27.142** | **51.197** | **41.639** | OG | **49.702** | **41.895** | **46.713** | CG | **10.887** | **29.234** | **46.771** | CG | **16.504** | **28.472** | **49.693** |
| CD | **26.908** | **52.253** | **42.720** | | | | | CD | **10.165** | **30.582** | **46.820** | CD | **16.322** | **27.008** | **49.702** |
| NE | **26.885** | **53.603** | **42.113** | | | | | CE | **10.533** | **31.327** | **48.105** | NE2 | **16.975** | **26.443** | **50.766** |
| CZ | **27.048** | **53.842** | **40.795** | | | | | NZ | **9.727** | **32.562** | **48.222** | OE1 | **15.743** | **26.192** | **49.030** |
| NH1 | **27.244** | **52.838** | **39.961** | | | | | | | | | | | | |
| NH2 | **27.012** | **55.109** | **40.329** | | | | | | | | | | | | |

**Table 4.** Additional contacts created in the refined structure of chain A through the statistical method. The 4-ASP–136-LYS contact is not listed, as it comes from the overlap with the backbone part of 136-LYS. It is hydrophilic-hydrophilic. The superscripts IN denote contacts which appear only for the IN conformation of 9-LYS. The remaining 9-LYS–141-ASP is common for the IN and OUT conformations. However, it is specific only in the OUT state.

| No. | $i$ | $j$ | $l_{ij}$ | Nature of the contact |
|---|---|---|---|---|
| 1 | 5-ALA | **138-GLN** | 6.6867 | hydrophilic–hydrophilic |
| 2 | **9-LYS** $^{IN}$ | 42-TYR | 9.2569 | hydrophobic–hydrophobic |
| 3 | **9-LYS** $^{IN}$ | 121-PHE | 7.9983 | not specific |
| 4 | **9-LYS** $^{IN}$ | 139-PHE | 5.5637 | hydrophobic–hydrophobic |
| 5 | **9-LYS** | 141-ASP | 6.2004 | hydrophilic–hydrophilic |
| 6 | 44-TYR | **50-GLU** | 9.4219 | hydrophilic–hydrophilic |
| 7 | 46-PRO | **50-GLU** | 5.6474 | hydrophobic–hydrophobic |
| 8 | **50-GLU** | 75-PRO | 8.4961 | not specific |
| 9 | **53-GLU** | 103-THR | 8.2496 | not specific |

more effective than the one with the $\beta$-C atoms. This is despite the fact that coordinates of the backbone heavy atoms and of the $\beta$-C atoms are provided by the PDB structure file. Furthermore, consideration of $\beta$-C does not resolve the ambiguity about the IN- and OUT-solutions for 9-LYS.

Both orientations IN and OUT come with the 9-LYS–141-ASP contact, which turns out to be crucial for the emergence of large mechanostability, but the IN choice generates three additional contacts, which affects the mechanostability further. In chain B the structure of 9-LYS happens to be set, and it is in the OUT state, so the OUT choice for chain A should be correct.

Table 4 lists the derived additional contacts for chain A of protein 1AOH. Six of them have been obtained through the statistical method. Three contacts would exist only in the IN state, and they come as a result of refinement. The nature of the contacts is established by using the CSU server.

## 3. Tests of the optimality of the refined side chains

For 9-LYS we have found two solutions. Now, we demonstrate that orientation OUT is favored by consideration of energies calculated within an all-atom model. Our molecular dynamics simulations were conducted by using version 2.9 of the NAMD simulation package [26] with the CHARMM22 force field [27, 28] for conformations IN and OUT as the starting states. Our simulations were conducted in a water box of size 85 Å in length, 80 Å in width, and 81 Å in height. It contained more than 50 000 TIP3P molecules of water [29]. The system was made charge neutral with the use of eight $Na^+$ ions. The total number of atoms in the system was 52 274 and 52 259 for conformations IN and OUT, respectively. The periodic boundary conditions were employed.

The first stage of the simulations involves energy minimization of the system for 100 000 conjugate gradient steps of 1 fs each. The second stage implemented heating of the system up to the temperature of 310 K. The heating process was done in three steps of 0.5 ns each. The first of them was done at temperature 110 K in the NVT ensemble. The second was done in the same ensemble but at 210 K. The last equilibration step was conducted in the NPT ensemble at 310 K. Each new stage used the coordinates and velocities of the atoms obtained from the previous stage. The temperature was controlled by the standard Langevin algorithm and the pressure by the Langevin piston pressure control algorithm.

When the system is prepared in the OUT state (from the very beginning of the procedure) it persists in it for at least 0.5 ns. On the other hand, when it is prepared in the IN state, it stays there for about 0.2 ns, and then it converts to the OUT state, which confirms the preference for the OUT structure. The same conclusion is reach through the consideration of energy as monitored during the 0.2 ns during which the system remains in the state it was prepared in. Specifically, we monitored the conformational potential energy ($E_{conf}$) of the system and the root mean square fluctuations (rmsf) in the positions of the side-chain atoms. We used the MDenergy plugin from the VMD package [30] for this purpose. The conformational energy is a sum of $E_b$ and $E_{nb}$, where the former is a collection of the bonded terms (i.e., bonds, angles, dihedrals, and improper dihedrals) and the latter of the non-bonded either electrostatic or van der Waals terms. Table 5 lists the average values of these energies, and it also provides the energy of the interactions of the protein with water, $E_{pw}$. It is observed that 9-$LYS_{OUT}$ leads to a lower conformational energy than 9-$LYS_{IN}$ and also to lower binding with water, even though 9-$LYS_{OUT}$ is more exposed. We also find that the RMSF values for 9-LYS are higher in the OUT state than in the IN state —movements in the OUT state are less restricted.

It is interesting to observe that the energy balance changes when the 1–4 segment is not included and the 4-ASP–136-LYS disappears. The IN state is now not observed to be unstable, and it actually comes with a lower conformational energy than the OUT state (table 5).

We now assess the impact of the corrections in the structure file on mechanostability. When one stretches

**Table 5.** Average values of the energies for chain A in 1AOH for two orientations of the 9-LYS side chain. The energies are in the units of kcal mol$^{-1}$. The upper half of the table refers to the situation in which the *N*-terminal segment 1–4 is included and the monitoring run is 0.2 ns long. The lower half refers to the situation in which the segment is not included and the run is 0.4 ns long: the IN state appears to be stable. The error bars are the standard deviations obtained over the duration of the runs.

| side chain | $\langle E_{nb} \rangle$ | $\langle E_{pw} \rangle$ | $\langle E_{conf} \rangle$ | *N*-terminal segment |
|---|---|---|---|---|
| 9-LYS$_{IN}$ | −2064.63 ± 58.21 | −5476.77 ± 94.99 | 301.17 ± 64.32 | with 1–4 |
| 9-LYS$_{OUT}$ | −2319.66 ± 76.74 | −5344.85 ± 140.83 | 37.76 ± 75.63 | with 1–4 |
| 9-LYS$_{IN}$ | −2079.03 ± 57.98 | −5227.18 ± 120.21 | 217.30 ± 60.53 | without 1–4 |
| 9-LYS$_{OUT}$ | −2047.98 ± 55.15 | −5215.46 ± 116.81 | 240.70 ± 63.36 | without 1–4 |

cohesin corresponding to 1AOH experimentally, one obtains a characteristic force for mechanical unravelling, $F_{max}$, which is close to 480 ± 14 pN at the constant pulling velocity, $v_p$, of 400 nm s$^{-1}$ [22]. Previously, we have stretched 1AOH within a structure-based coarse-grained model [8, 22, 25] constructed by fixing the structure file by introducing a 7.5 Å cutoff in $l_{ij}$s if at least one of the AA involved had missing side chains. This procedure has resulted in the introduction of several tens of extra contacts. In [22], we did not include the first four missing residues—we took the structure file of chain A as it was. The model incorporated an overdamped molecular dynamics involving only the $\alpha$-C atoms. The stretching simulations for the reconstructed structures have been accomplished at several values of $v_p$. By extrapolating to the experimental speed, $F_{max}$ has turned out to be equal to about 3.7 $\epsilon$/Å, where $\epsilon$ is the energy parameter—it is the depth of the potential well associated with a native contact. By using the calibration of $\epsilon$/Å being approximately equal to 110 pN [25], the theoretical extrapolated result is 410 pN. At $v_p$ of 0.005 Å/$\tau$ ∼ 500 000 nm s$^{-1}$ it is 4.2 ± 0.2 $\epsilon$/Å, i.e., about 460 pN. The error bars are determined based on 100 trajectories.

We now repeat these simulations for the case with the reconstructed side chains and when the 1–4 residues are included and, therefore, with one of the mechanically important contacts 4-ASP–136-LYS. The simulations were performed by using a stochastic dynamics in order to mimic random kicks by molecules of the implicit solvent. The dynamics of the individual $\alpha$-Cs was governed by a Langevin equation: $m\ddot{r} = -\gamma\dot{r} + F_c + \Gamma$, where $F_c$ is the total force on an atom due to molecular potentials and $\Gamma$ is a Gaussian noise term with dispersion $\sqrt{2\gamma k_B T}$ ($k_B$ is the Boltzmann constant). The friction coefficient $\gamma$ is taken to be equal to $2m/\tau$, where $\tau$ is the characteristic time scale in the model. It is expected to be of order 1 ns, as it is associated with the diffusive (instead of ballistic) coverage of molecular distances. The molecular potentials involve [8] the native contacts of the Lennard-Jones form, repulsive non-native contacts, and the local backbone stiffness represented by a chirality potential [7] that effectively takes into account the dihedral terms. The equation of motion is solved by a fifth-order predictor–corrector scheme. In the stretching simulation, the native contacts are considered to be broken only if the distance $r_{ij}$ between AAs $i$ and $j$ involved in the process is larger than $1.5\sigma_{ij}$ (close to the inflection point of the Lennard-Jones potential) for the last time.

The inset of figure 3 shows typical force–displacement ($F - d$) patterns for the two choices of positioning of the side group on 9-LYS at $v_p = 0.005$ Å/$\tau$. (The two traces are obtained for the same string of random numbers in the Langevin noise that controls thermal fluctuations.) After averaging over 50 trajectories, the correct orientation OUT yields 4.2 $\epsilon$/Å (the black line, also repeated in the main part of the figure), whereas orientation IN yields 4.4 $\epsilon$/Å (the blue dotted line). If the missing side chains are not reconstructed (the green line), we get 3.7 $\epsilon$/Å. Thermal fluctuations are of order 0.1 $\epsilon$/Å at the temperature $T = 0.3\,\epsilon/k_B$ used here.

If the 1–4 segment is removed, the IN-solution yields $F_{max}$ of 4.2 $\epsilon$/Å (as shown by the magenta line in the main part of figure 3), and the OUT-solution yields 4.0 $\epsilon$/Å. We conclude that one can mimic the mechanical effect of contact 4-ASP–136-LYS in the *truncated* A chain by adopting the IN-orientation and thus incorporating the three extra contacts that come with this choice.

## 4. Properties of the contact map—comparisons to CSU

Figure 4 compares the contact maps in the fully fixed native structure of 1AOH (orientation OUT) obtained by using the overlap criterion (denoted as OV) to that derived through the CSU approach. The overlap-based method introduces 382 contacts, including those derived by the statistical method, and the CSU procedure adds 25 specific contacts. In the CSU approach, the contacts are considered specific if they correspond to the hydrogen bonds, hydrophobic couplings, and stacking interactions. The ionic bridges count as non-specific, but they are typically captured by the OV-based approach (which is not concerned with the sizes of the water molecules). The CSU method gives rise to 270 specific contacts. Of these, 5-ALA–138-GLN, 9-LYS–42-TYR, and 9-LYS–141-ASP are predicted by our statistical method, whereas 44-TYR–50-GLU and 46-PRO–50-GLU are absent.

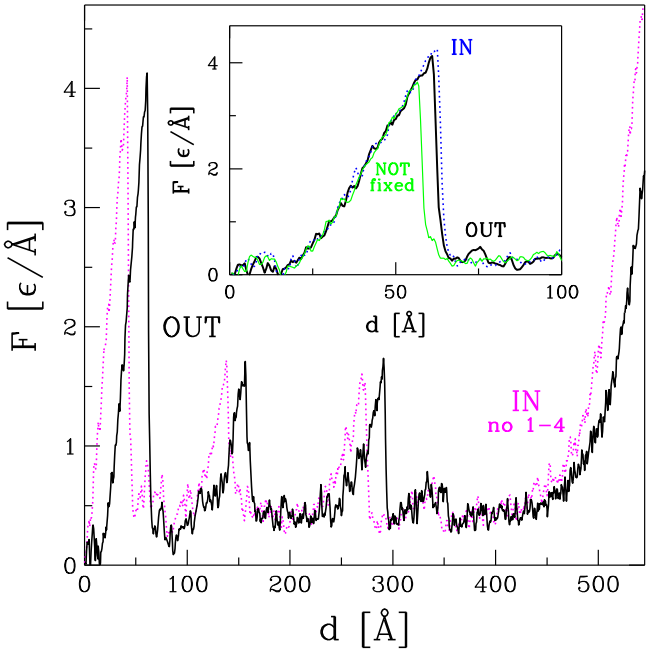


**Figure 3.** Typical force–displacement traces for chain A. Solid black line: an example of an $F - d$ trace for stretching the corrected 1AOH structure at constant speed. The side group of 9-LYS is in the OUT position. The first peak is the largest, and its height defines $F_{max}$. Dotted magenta line: the $F - d$ pattern when 9-LYS is in the IN position but the 1–4 amino acids are missing—as a result the force peak is moved to the left. The inset compares the solid trace of the main figure (the OUT situation) to the one corresponding to position IN in the immediate vicinity of the major force peak (blue dotted line) and to the situation in which the missing side groups are not fixed (the green solid line).

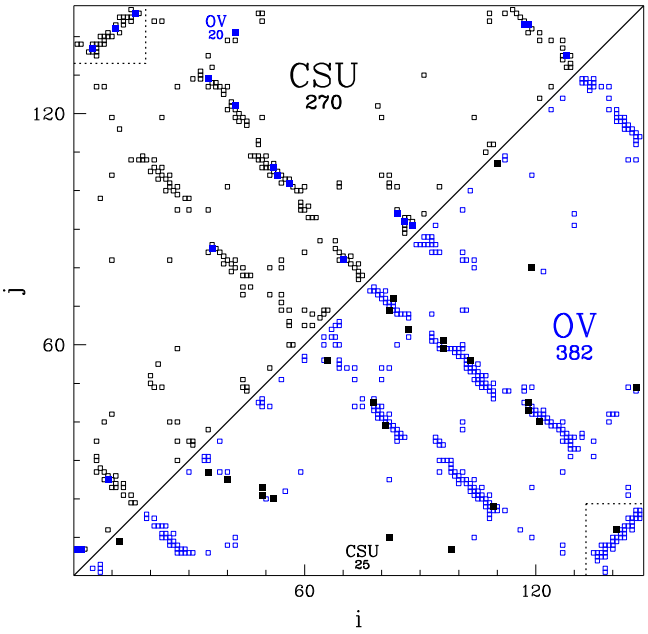


**Figure 4.** The native contact map for 1AOH. Contact maps are expected to be symmetric. (There are exceptions that arise when using the CSU.) In the upper part, the open squares show the 270 contacts obtained using the CSU-based approach and the blue solid squares—20 contacts that are additionally predicted by the overlap (OV) criterion. In the lower part, the blue open squares show the 382 contacts predicted by the overlap criterion and the black squares—25 contacts that are additionally predicted by CSU. The dotted rectangles indicate contacts that are responsible for the large value of $F_{max}$.

The overlap criterion adds 20 contacts to those found through CSU. Overall, both contact maps look broadly similar. We think that molecular dynamics simulations within the structure-based coarse-grained models should employ either the overlap contact map combined with the extra specific CSU contacts or the CSU maps with the extra contacts obtained through overlap. The contacts responsible for the large mechanostability are marked off by the dotted rectangle in figure 4.

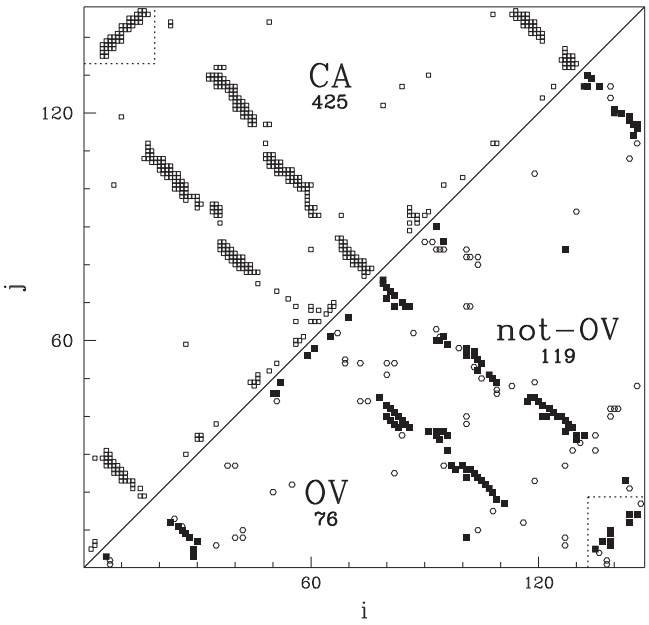


**Figure 5.** The upper part shows 425 contacts obtained by replacing all AAs by spheres of radius $R_{\alpha C}$ and centered on the $\alpha$-C atoms. The contact map is for the fully fixed chain A of 1AOH. The black squares in the lower part show those contacts in the CA map that are not present in the underlying OV map. The open hexagons in the lower part show those contacts in the OV map that are not present in the CA map. The dotted rectangles indicate contacts that are responsible for the large value of $F_{\max}$.

It is interesting to ask what kind of contact map arises as a result of replacing all AAs, not only the defective ones, by the spheres of radius $R_{\alpha C}$. Figure 5 shows such a map—it is denoted as CA. It is seen that it captures the pattern of the overlap-based contact map of figure 4, but it has 43 more contacts. This constitutes about 11% of the contacts generated through the overlaps of the effective van der Waals spheres associated with the heavy atoms of the repaired structure. The CA map comprises 425 contacts, out of which 119 are not in the original OV map, and 76 of the underlying OV map are missing in the CA map.

## 5. Conclusions

We have shown that considerations of contact-making in a protein can influence the refinement process in a meaningful and constructive way. Any arising ambiguities coming from the existence of possible but less optimal solutions can be resolved by performing small-scale, all-atom molecular dynamics simulations. For chain A of the cohesin domain of type I discussed here, this procedure yields results that are consistent with the structure of chain B (for which the repaired contact map turns out to be almost the same as the repaired contact map for chain A) and also compatible with the experimental results on mechanostability.

We have focused on one procedure that defines the contacts and found that the overlaps of spheres associated with the $\alpha$-C atoms placed on the defective atoms work better than those associated with the $\beta$-C atoms.

The modifications in the contact map CA relative to OV are substantial and lead to a 19% increase in $F_{\max}$: from 4.20 $\pm$ 0.15 $\epsilon$/Å to 5.00 $\pm$ 0.17 $\epsilon$/Å. We observe that for each of the two contact maps, and between the maps, all $F - d$ patterns look similar, indicating the presence of just one pathway. A sample of 25 trajectories is shown in figure 6 in the SI. Despite the difference in $F_{\max}$, it appears that using the $\alpha$-C-centered spheres may still be a sensible strategy to apply in situations where structural information is missing. One example is multi-protein complexes, such as cellulosomes, where only some substructures are resolved and interactions between them are unknown. Another is large conformational changes, arising either during stretching or folding, which generate non-native attractive contacts as a result of time evolution.

It would be interesting to generalize our statistical method to other definitions of contacts, such as those based on the CSU approach or on introducing a length cutoff. The overlap-based method used here is a compromise between the simplicity of use and the possibility of accounting for the physical differences between the AAs.

It should be noted that, generally, the absence of AA fragments in crystallographical structures may indicate experimental difficulties or the presence of segments with no unique stable conformations, such as disordered loops and flexible linkers. 1AOH is an example of the former, as evidenced by making comparisons between chains A and B. However, our method can also find applications of the latter, as it can

pin down possible fixed choices while leaving the contactless regions free to move.

## Acknowledgments

We appreciate illuminating discussions with A Gomez-Sicila, M Gunnoo, B Różycki, M Sikora, and D Thompson. The computer resources were financed by the European Regional Development Fund under the Operational Programme Innovative Economy Nano-Fun POIG.02.02.00-00-025/09. This work has been supported in part by the EU 7th Framework Programme under the project NMP4-SL-2013-604530 (CellulosemePlus) and by the ERA NET grant ERA-IB (EIB.12.022) (FiberFuel). It was also co-financed by the Polish Ministry of Science and Education from the resources granted for the years 2014-2017 in support of international scientific projects.

## References

[1] Hoang T X and Cieplak M 2000 *J. Chem. Phys.* **113** 8319
[2] Clementi C, Nymeyer H and Onuchic J N 2000 *J. Mol. Biol.* **298** 937
[3] Karanicolas J and Brooks C L III 2002 *Protein Sci.* **11** 2351
[4] Cieplak M, Hoang T X and Robbins M O 2002 *Proteins: Struct. Funct. Bio.* **49** 114
[5] Takada F, Koga N and Takada S 2003 *Proc. Natl Acad. Sci. USA* **100** 11367
[6] Levy Y, Wolynes P G and Onuchic J 2004 *Proc. Natl Acad. Sci. USA* **101** 511
[7] Sułkowska J I and Cieplak M 2008 *Biophys. J.* **95** 3174
[8] Sułkowska J I and Cieplak M 2007 *J. Phys.: Cond. Mat.* **19** 283201
[9] Sali A and Blundell T L 1993 *J. Mol. Biol.* **234** 779
[10] 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 and Adams P D 2012 *Acta Cryst.* D **68** 352
[11] Canutescu A A, Shelenkov A A and Dunbrak R L 2003 *Prot. Sci.* **12** 2001
[12] Summers N L and Karplus M 1989 *J. Mol. Biol.* **210** 785
[13] Cieplak M and Hoang T X 2003 *Biophys. J.* **84** 475
[14] Bondi A 1964 *J. Phys. Chem.* **68** 441
[15] Tsai J, Taylor R, Chothia C and Gerstein M 1999 *J. Mol. Biol.* **290** 253
[16] Settanni G, Hoang T X, Micheletti C and Maritan A 2002 *Biophys. J.* **83** 3533
[17] Sobolev V, Sorokine A, Prilusky J, Abola E E and Edelman M 1999 *It Bioinformatics* **15** 327
[18] Tavares G A, Beguin P and Alzari P M 1997 *J. Mol. Biol.* **273** 701
[19] Beguin P and Lemaire M 1996 *Crit. Rev. Biochem. Mol. Biol.* **31** 201
[20] Bayer E A, Chanzy H, Lamed R and Shoham Y 1998 *Curr. Opin. Struct. Biol.* **8** 548
[21] Bayer E A, Belaich J P, Shoham Y and Lamed R 2004 *Annu. Rev. Microbiol.* **58** 521
[22] Valbuena A, Oroz J, Hervas R, Vera A M, Rodriguez D, Menendez M, Sułkowska J I, Cieplak M and Carrion-Vazquez M 2009 *Proc. Natl Acad. Sci. USA.* **106** 13791
[23] *The PyMOL Molecular Graphics System*, Version 1.5.0.4 Schroedinger, LLC (http://pymol.org)
[24] Sillitoe I *et al* 2013 *Nucl. Acids Res.* **41** D490
[25] Sikora M, Sułkowska J I and Cieplak M 2009 *PloS Comp. Biol.* **5** e1000547
[26] Phillips J C, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel R D, Kale L and Schulten K 2005 *J. Comp. Chem.* **26** 1781
[27] MacKerell A D *et al* 1998 *J. Phys. Chem.* B **102** 3586
[28] MacKerell A D Jr, Feig M and Brooks C L III 2004 *J. Comp. Chem.* **25** 1400
[29] Jorgensen W L, Chandrashekar J, Madhura J D, Impey R W and Klein M L 1983 *J. Chem. Phys.* **79** 926
[30] Humphrey W, Dalke A and Schulten K 1996 *J. Mol. Graphics* **14** 33