Published in final edited form as:
Proteins. 2011 January ; 79(1): 261–270. doi:10.1002/prot.22879.

Automated Minimization of Steric Clashes in Protein Structures

Srinivas Ramachandran[1,2,†], **Pradeep Kota**[1,2,†], **Feng Ding**[1], and **Nikolay V. Dokholyan**[*,1,2]

[1]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, NC 27599-7260 USA

[2]Program in Molecular and Cellular Biophysics, University of North Carolina at Chapel Hill, NC 27599-7260 USA

Abstract

Molecular modeling of proteins including homology modeling, structure determination, and knowledge-based protein design requires tools to evaluate and refine three-dimensional protein structures. Steric clash is one of the artifacts prevalent in low-resolution structures and homology models. Steric clashes arise due to the unnatural overlap of any two non-bonding atoms in a protein structure. Usually, removal of severe steric clashes in some structures is challenging since many existing refinement programs do not accept structures with severe steric clashes. Here, we present a quantitative approach of identifying steric clashes in proteins by defining clashes based on the Van der Waals repulsion energy of the clashing atoms. We also define a metric for quantitative estimation of the severity of clashes in proteins by performing statistical analysis of clashes in high-resolution protein structures. We describe a rapid, automated and robust protocol, *Chiron*, which efficiently resolves severe clashes in low-resolution structures and homology models with minimal perturbation in the protein backbone. Benchmark studies highlight the efficiency and robustness of Chiron compared to other widely used methods. We provide Chiron as an automated web server to evaluate and resolve clashes in protein structures that can be further used for more accurate protein design.

Keywords

Homology modeling; refinement; Chiron; Discrete Molecular Dynamics; Protein Design

Introduction

The role of molecular modeling of proteins in structural and molecular biology is steadily increasing due to its usefulness in generating experimentally testable hypotheses for understanding protein function. Molecular modeling techniques are also widely used for protein design[1] and fitting all-atom models to low-resolution data (e.g. electron microscopy[2] and small-angle X-ray scattering[3]). The ever-expanding database of protein 3D structures has made comparative modeling a viable option to model proteins of unknown structure.[4] However, all-atom structural modeling of proteins requires stringent quality control to ensure that the structures are physically accurate. Additionally, protein design and comparative modeling require several refinement steps to culminate in a usable structural model. Steric clash, characterized by unphysical overlap of newly positioned side-chain

[*]Correspondence to Nikolay V. Dokholyan, 3097 Genetic Medicine Bldg, Campus Box 7260, Chapel Hill, NC 27599, USA, dokh@med.unc.edu, phone: 919-843-2513, fax: 919-966-2852.

[†]Srinivas Ramachandran and Pradeep Kota contributed equally to this work

atoms with other side-chain and backbone atoms, is one of the prevalent artifacts in protein structures of low resolution.

Current state-of-the-art tools for protein quality control identify clashes qualitatively, precluding an understanding of their possible energetic effects on protein structure. For instance, WHAT_CHECK[5,6] and Molprobity[7], commonly used in protein quality control, report a steric clash based on distances between two atoms with a distance cutoff for overlap set to 0.4 Å. However, the energetic penalty of such an overlap varies widely depending on the types of atoms involved in the clash (0-10 kcal/mol). We observe that low energy clashes are present even in high-resolution structures, however the number of severe clashes is very low. Thus, in order to correctly identify severe clashes, it is important to develop a quantitative measure to evaluate the effect of clashes present in a protein, and also it is necessary to measure the extent of clashes seen in high-resolution crystal structures.

Several tools have emerged for resolution of such clashes upon identification. Steepest descent/Conjugate gradient minimization using all-atom Molecular Mechanics force fields is the most widely used method to resolve clashes in a protein structure before using the structure for further studies. However, minimization using Molecular Mechanics may not resolve severe clashes in some cases hampering subsequent Molecular Dynamics simulations. Molecular modeling tools like Rosetta are the alternate avenues for refining structures with severe clashes. These tools use knowledge-based potentials and small backbone moves to resolve clashes. However, these methods work best with smaller proteins (less than 250 residues in size).[8] Tools like MMTSB[9] and PULCHRA[10] have emerged for structure refinement and for reconstruction of all-atom representation of proteins from Cα traces, which includes removal of clashes during refinement. In the current study, we present a method for quantitative estimation and if required, resolution of clashes in a given protein structure. To accomplish the above, we developed a protocol using discrete molecular dynamics (DMD) simulations.[11,12] We also show that our protocol is more robust in comparison to other state-of-the-art tools widely used by the protein structural modeling community.

Materials and Methods

Definition of steric clashes and the acceptable clash-score

We define a steric clash in a protein as any atomic overlap resulting in Van der Waals repulsion energy greater than 0.3 kcal/mol (0.5 k_BT), except i) when the atoms are bonded, ii) when the atoms form a disulfide bond or a hydrogen bond (i.e. the heavy atoms are involved in the hydrogen bond; we assign the Van der Waals radius of hydrogen to be zero), iii) when the atoms involved are backbone atoms and have separation of 2 residues (in order to accommodate the formation of tight turns). We calculate the Van der Waals repulsion energy using the non-bonded parameters from the CHARMM19 force field[13], which are identical to CNS[14] parameters except for carboxyl oxygen atoms. Since clashes are local structural artifacts, we reduce the search time and space by restricting the search to the local environment of a given atom. We determine clashes using the above definition by constructing a grid around the protein with the dimension of each cell larger than the largest Van der Waals interaction distance between any two atom pairs (~4.5 Å) and walking along the chain to check if the overlap of the atom under consideration with the heavy atoms in the same or adjacent cells leads to a clash. We then define ‘contacts’ as the number of such overlaps tested. The clash-energy of a protein is the sum of Van der Waals repulsion energy of all the clashes in the protein’s structure. In order to arrive at a descriptor independent of protein size, we define the clash-score, which is the clash-energy divided by the number of contacts. Thus, clash-score describes the clashes present in a protein-structure, but is independent of the size of the protein. To estimate the permissible Van der Waals repulsion in a given structure, we determine the clash-scores of high-resolution crystal structures (see below). The distribution of these clash-scores indicates the extent of clashes permissible in proteins as a consequence of tight packing. A clash-score that is deviant from the distribution for high-resolution structures would then point to clashes that are artifacts of model building rather than those inherent to the protein structure. Clash-score is acceptable if it is less than one standard deviation away from the mean on the higher side of the distribution of clash-score of high-resolution dataset of structures (which would include ~84% of the proteins in the dataset). From the distribution of clash scores of structures from the high-resolution dataset, we calculate the acceptable clash-score to be 0.02 kcal.mol^{-1}.contact^{-1} (Fig. 1 B).

Protein datasets

In order to understand the extent of clashes in protein structures and arrive at an acceptable clash-score, we constructed datasets of protein structures of various resolutions. We obtained two sets of protein structures from the Protein Data Bank (PDB)[15] and one set of protein structures from Swiss-model repository.[16] The sets obtained from PDB correspond to a high-resolution set (0-2.5 Å) and a low-resolution set (2.5-3.5 Å). The high-resolution set was used to arrive at the acceptable clash-score. Our high-resolution dataset comprises protein structures determined by X-ray crystallography with a reported resolution less than 2.5 Å. Other than protein chains, these structures did not contain any other biomolecules (i.e. ligands/DNA/RNA). We then split these structures into individual peptide chains and clustered them based on sequence similarity. We used individual chains because we wanted clash statistics of globular proteins and not interfaces. We considered only one representative chain from each cluster of sequences that were at least 80% similar to each other, thus creating a dataset of 4495 unique chains. We further filtered the dataset based on radius of gyration to remove non-globular peptides/proteins from the dataset. The final dataset consisted of 4311 single chains at least 25 residues long. We used Medusa[17,18] to accurately place any missing side-chain atoms in these structures.

We obtained a low-resolution dataset from PDB in order to explore if clash-score was worse in low-resolution structures compared to high-resolution structures (Results). The lower-resolution dataset contains 2942 unique protein structures determined using X-ray crystallography with a resolution between 2.5Å and 3.5Å. In addition to these two datasets, we obtained a set of 1000 homology models from the Swiss-model repository of random swiss-model entries (using the CGI-perl script provided by expasy: http://www.expasy.org/cgi-bin/get-random-entry.pl?S). We filtered these structural models based on radius of gyration resulting in a final dataset of 931 structural models.

Minimization using DMD

The DMD simulation methodology is described in detail elsewhere.[11,12] DMD is a special type of molecular dynamics (MD) algorithm, which uses square-well potentials instead of continuous potentials. Thus, in DMD we seek to solve the ballistic equations of motion instead of Newtonian equations of motion in a system of particles. We use CHARMM19 non-bonded potentials[13], EEF1 implicit solvation parameters[19] and geometry-based hydrogen bond potentials in DMD[11] to model various macromolecular interactions. The time unit of the all-atom DMD simulations is ~50 femtosecond (fs) and the temperature is maintained using Anderson's thermostat.[20] The rate of velocity rescaling (for maintaining temperature) depends on the simulation we perform (as shown in Fig. 2); in the present study we used either 200 ps^{-1} or 4 ps^{-1} as the rescaling rate.

Minimization using Rosetta protein design suite

We compared the performance of our technique to resolve clashes with that of Rosetta[1], which is used widely for protein design and refinement. We used two different protocols for minimization of protein structures using Rosetta. In the first minimization protocol, we used the "fast relax" flag with 4 fast relaxation repeat steps per cycle. We also performed "constrained relax", where all the backbone atoms were constrained to their initial coordinates. We performed 100 independent iterations for minimization using either protocol. For each of such independent computation we used a unique random seed to initiate the minimization routine.

Relaxation of clashes using molecular mechanics

We used the recently ported CHARMM[13] force field in GROMACS[21] simulation package for performing protein energy minimization.[22] In our protocol we performed an initial conjugate gradient (CG) minimization for 1000 steps, where we considered maximum force less than 200 $\text{kJmol}^{-1}\text{nm}^{-1}$ as criteria for convergence. If the clash-score of the protein had not decreased below the acceptable clash-score during CG, we performed 2 ps MD simulation at 240 K, before an equilibrium MD simulation for 2 ps at 300 K. We did not use any cut-offs in these simulations, and used OBC (Onufriev, Bashford, Case) implicit solvation.[23] We used the temperature scaling method of Ref. [24], with a time constant of 0.5 ps. We also attempted to implement a lower time constant of 0.05 ps, to mimic our DMD simulations with high rate of temperature scaling, but the results did not differ significantly between time constant of 0.5 ps and 0.05 ps in our simulations (data not shown).

Proteins used in testing minimization protocols

To test the performance of various programs used in minimizing clashes, we used several structures from low-resolution and Swiss-model datasets. We used 20 structures from our low-resolution PDB dataset with the worst clash-scores and 50 structures from our Swiss-model dataset of homology models (25 structures less than 250 residues in length and 25 structures more than 250 residues in length) to test the ability of different programs to minimize clashes.

Determination of the number of unsatisfied hydrogen bond donors/acceptors in the buried core of the protein

We define buried residues (buried core of the protein) as those that have a solvent accessible surface area (SASA) of zero Å^2. We calculated SASA using the method developed by LeGrand and Merz[25], but with 1024 dots on surface of each atom instead of 256 dots. We first constructed all possible hydrogen bonds in a given protein structure using the orientation-dependant hydrogen bonding potential, which is part of Medusa[17,18] and iteratively checked if each of the polar atoms in the buried residues was involved in a hydrogen bond; those that did not form hydrogen bonds were counted towards buried unsatisfied donors/acceptors.

Z-score of sidechain dihedrals (χ angles)

To assess the geometry of sidechains in the input structure and the structures obtained by various minimization techniques, we define a metric based on Dunbrack rotamer library[26]. For each residue in a given structure (except glycines and alanines), we determine the closest rotamer in the rotamer library and calculate the Z-score of the side-chain to the identified rotamer as shown below:

$$Z_j = \sum_{i=1}^{n} \frac{1}{2^{(i-1)}} \cdot \frac{\left|\chi_i - mean_\chi\right|}{sd_\chi}$$

where χ_i is the i^{th} χ angle for a given sidechain, mean_χ and sd_χ are obtained from the Dunbrack library for the rotamer that is closest to a given sidechain. We report the above Z-score, averaged over all the sidechains in a given protein structure (Supplementary Table V).

Results

Clash-score is dependent on the resolution of the protein structure

We explore the prevalence of clashes in protein structures and ask if the extent of clashes in a protein structure depends on the quality of the protein structure. Using our energetic definition of clashes (Methods), we first study the distribution of the energy of different clashes, setting the minimum clash-energy as half k_BT (0.3 kcal/mol). We find an exponential distribution of the energy of clashes in crystal structures of proteins: most of the clashes in a structure correspond to very low repulsion energy (0.3-0.6 kcal, Fig. 1 A), and the probability of finding clashes of higher energies decreases exponentially (Fig. 1 A, inset). The clash energy distribution confirms that severe clashes are very rare in protein structures. However, a lot of low-energy clashes are observed, which may be required for close packing of the protein core and also for the formation of hydrogen bonds. Even though we exclude hydrogen-bonded atoms in enumerating clashes, during the formation of a hydrogen bond, adjacent atoms are also brought very close to each other. We also observe that the probability of finding high-energy clashes is higher in low-resolution crystal structures compared to high-resolution crystal structures (Fig. 1A, inset)

In order to obtain an aggregate clash-score for a structure, we sum up the energies of all clashes and divide it by total number of ‘contacts’. Contacts include all possible local overlaps in the protein (Methods). The clash-score is thus a single number that describes the maximum permissible extent of clashes in a protein structure. In order to determine whether the clash-score of a given protein is dependent on the resolution of its crystal structure, we compute the clash-score for all the protein structures in our datasets. Since the clash-score itself is a sum of independent variables (the energy of different clashes in a protein), the clash-scores of proteins in a dataset are expected to form a normal distribution. As expected, the clash-scores of the three datasets (high-resolution, low-resolution and homology models) fit well to normal distributions (Fig. 1 B). We find that the clash score of a given protein structure is strongly dependent on its resolution (Fig. 1 B). High-resolution crystal structures feature low clash scores, while the mean clash score of lower resolution structures is significantly higher than that of high-resolution crystal structures with the homology models having the highest clash-scores in our three datasets (Fig. 1 B). These results indicate that the increased clash-scores in lower resolution structures is more the consequence of model building than being an inherent property of the protein structure. Furthermore, distribution of clash-energies in a protein show that low-resolution structures on an average feature more number of high-energy clashes compared to high-resolution structures (Fig. 1 A, inset).

Given the differences in distribution of clash-scores of different datasets, we arrived at an acceptable clash-score as the mean plus one standard deviation of the distribution of clash-scores of high-resolution structures. From the distribution of clash-scores for high and low resolution crystal structures and homology models (Fig. 1), it is clear that we require a tool that reduces the clash-score to be within the permissible value. To check if side-chain repacking alone can reduce abnormal clash-scores, we performed side-chain optimization using SCWRL[27] on 20 structures from our low-resolution PDB dataset with the worst clash-scores (Supplementary Table I). We find that just side-chain repacking is not enough to improve the clash-score in these structures, and in many cases, optimizing side-chain rotamers results in higher clash-scores compared to input structures. Thus, small perturbations in the protein-backbone may be required for resolving clashes in low-resolution structures.

Automated protocol for minimization of steric clashes

To address the need for removing severe clashes in protein structures with minimal perturbations in the protein backbone, we develop a protocol using discrete molecular dynamics (DMD) simulations[12] (Fig. 2). Severe clashes have high values of Lennard-Jones potential, the release of which results in high velocities of clashing atoms. The high velocities in the simulation result in a rapidly expanding simulation system with bonds between atoms being broken. Thus, we need to quench the high velocities arising due to clashes, which we achieve using a high heat exchange rate of the solute (protein) with the bath. In our simulations, we maintain the temperature by periodically scaling the velocities of particles according to Maxwell's distribution. By using a high heat exchange rate (or the frequency of rescaling the velocities) we ensure that the high velocities of clashing atoms are quenched. In this protocol, we rescale the velocities of all particles every 5 fs instead of ~1-5 ps commonly implemented in equilibrium MD simulations. To minimize deviation from the initial structure in these simulations, we constrain the backbone and Cβatoms to their initial positions using a harmonic potential. In some cases (like enzyme active sites), the rotameric state of certain residues is important for protein function, and those may not be involved in any clashes. In such cases, we could also constrain the side-chain atoms to their initial positions in order to maintain the initial rotameric orientations of side-chain atoms. In our protocol (Fig. 2), we first perform a short DMD simulation (10 ps) at a higher temperature (0.7 ε/k_B, roughly corresponds to 350 K; ε/k_B is the reduced unit of temperature[11]), and then compute clash-score of every snapshot in the simulation trajectory to see if the system has converged to feature a clash-score less than or equal to the acceptable clash-score (Fig. 2). We report as the minimized structure, the first snapshot in the trajectory that has an acceptable clash-score. If the clash-score is not low enough during the first simulation cycle, we perform another cycle of DMD simulation at 0.5 ε/k_B, to quench the system. The "quench" approach aids in the formation of contacts and hydrogen bonds in the protein core that might remain unsatisfied at higher temperatures. We repeat the alternating cycles of simulations at 0.7 ε/k_B and 0.5 ε/k_B until a structure with acceptable clash-score is obtained.

We benchmark our protocol by attempting to resolve clashes in protein structures with high clash-scores. We select 20 structures from our low-resolution PDB dataset with the worst clash-scores. We reduce the clash-score of all these structures within the range of high-resolution structures (less than 0.02 $\text{kcal.mol}^{-1}.\text{contact}^{-1}$), with Cα RMSD of utmost 1 Å with respect to the corresponding initial structures (Table I, Supplementary Table II; Fig. 3 shows one of the structures). The backbone and all-atom RMSD follow similar trends, but as expected, they are slightly higher than Cα RMSD (Supplementary Table III). Some of these structures initially have severe clashes that cannot be acceptable for MD simulations (see below), but we can robustly minimize those clashes using DMD and make the structures acceptable for MD simulations. In order to ensure that stabilizing interactions in the core are not sacrificed for the sake of lowered clash energy, we examine the hydrogen bonds in the core of the protein. We find that in most of the cases, the number of unsatisfied hydrogen bond donors/acceptors decrease, and the core seldom has perturbation that causes significant loss of contacts (Table II, Supplementary Table IV). When we compare the initial and DMD-minimized structures, we notice that slight perturbations in the backbone ensure that the clashes are resolved while contacts in the core are maintained intact (Fig. 3). In order to ensure that the clash minimization does not result in distortion of side-chain geometry, we calculate the Z-score of χ angles of each side-chain (Supplementary Table V). On average, we find the Z-scores of DMD minimized structures to be close to or less than the Z-scores of the input structures. We would expect the Z-scores of DMD-minimized structures to be close to the input structures since we constrain the Cβ atoms in our simulations, which maintains the side-chains close to the initial rotameric states.

From our benchmarking simulations, we conclude that our protocol based on DMD can be used as a tool for rapid and efficient minimization of steric clashes in protein structures. Our protocol (Fig. 2) is available as a web-based application – Chiron (http://chiron.dokhlab.org). Chiron accepts protein-structures and attempts to resolve clashes with minimum backbone perturbation relative to the input structure. In the web server, any given input structure is first examined for missing atoms – missing side-chain atoms are reconstructed using Medusa[17],[18], which uses Dunbrack rotamer libraries[27]. Chiron then computes the clash-score to determine if the structure requires minimization (if the clash-score is greater than 0.02). Chiron then iteratively runs constrained DMD simulations till the clash-score is acceptable. Chiron also accepts protein-structures with ligands (designated as HETATM in the input PDB files). The user can choose the ligands to be included in determination of clash-score and for further minimization. When ligands are included in an input structure, we reconstruct side-chains in context of the ligands in order to ensure that the reconstruction does not introduce clashes with the ligands.

We now compare our approach in minimizing clashes in protein structures to other widely used simulation programs using the publicly available versions of the respective programs.

Comparison with Rosetta protein design suite

We use Rosetta, one of the widely used tools for protein design, to minimize clashes in protein structures to benchmark DMD. Rosetta uses a knowledge-based energy function for treating inter-atomic interactions in proteins.[1] We choose Rosetta for comparison with DMD since both the tools use the same bonded and non-bonded parameters derived from the CHARMM force field.[13] We use two different protocols for clash minimization using Rosetta (see Methods) and report that we are able to resolve clashes in 8 out of 10 test cases with Rosetta, but with a higher Cα RMSD relative to the starting structure when compared to DMD (Table I, gray cells). We observe that for the structures for which Rosetta is able to successfully minimize clashes, the clash-scores are much lesser than those obtained by DMD or all-atom conjugate gradient (CG) minimization (see below). The lower clash-score of the final structure obtained using Rosetta is only a consequence of the simulation protocol: while DMD and CG/MD simulations are terminated as soon as the clash-score is less than the acceptable clash-score, the Rosetta protocol has a fixed routine which provides a single structure as an output for each iteration.

Rosetta also takes much longer CPU time to minimize clashes in the given protein structure (Table I) compared to DMD. Since Rosetta performs protein design by selectively replacing fragments of the given structure by determining the best fragment from a knowledge-driven fragment database, it is imperative that the protocol must be executed multiple times for statistically significant results. We perform 100 relaxation iterations-for every chosen protein in a parallel computing environment, with a unique random seed for every individual attempt. The runtime reported in Table I is the sum of times taken for all the independent iterations on the cluster. It is logical to compare runtimes in this manner because DMD performs minimization on a single processor in the times reported in Table I.

We also attempt to implement the “classic relax” protocol designed to perform more extensive relaxation employing small and shear moves of the protein backbone along with minimization and repacking of the side chain atoms. Since the classic relax protocol is extensive, we would expect it to take longer than the fast or constrained relax runs. Minimization of a 71 amino acid protein (1CTX) takes almost one day of CPU time, with no significant improvement in the final output compared to fast and constrained relaxation routines (data not shown). Hence, we do not perform classic relaxation of the remaining test cases. Overall, although we can use Rosetta to minimize the number of clashes in a given protein in 8 out of 10 cases tested, the downside of this approach is higher RMSD relative to the initial structure and longer computation time required for minimization compared to the other protocols tested. From the results obtained for the first ten PDB structures (Table I) and also from literature [8], we observe that Rosetta is not ideal for minimization of proteins longer than 250 residues. Hence, we do not use Rosetta to minimize the ten additional structures we consider for benchmarking DMD and CG/MD simulations (Supplementary Table II, IV).

Comparison with CG minimization

We compare our results to those obtained using conjugate gradient (CG) minimization using all-atom forcefield (CHARMM). We perform CG minimization of the test set of proteins (Table I, Supplementary Table II; see methods for the protocol). We perform subsequent molecular dynamics (MD) simulations for those proteins whose clash scores are not reduced below the acceptable clash score using CG minimization alone. Most of the protein structures attain acceptable clash-scores with CG alone, which leads to much shorter time required for minimization compared to DMD or Rosetta. The quality of the structure is also maintained in these cases. Thus, CG is an efficient way to remove clashes in most protein structures. If the structures require subsequent MD simulations for complete minimization, the time taken is comparable to DMD.

However, in four out of twenty structures considered, CG does not converge because of unreasonable Van der Waals repulsion energy leading to high velocity, which in turn causes some of the bonds to break. We show these clashes before and after DMD minimization for one such structure (PDB ID 1GFF, Fig. 4) that does not converge due to severe clashes leading to infinite forces in the simulation system. We are able to perform DMD simulations with these structures because we employ soft potentials for non-bonded interactions coupled with frequent rescaling of velocities. Soft-core potentials are an available option in MD forcefields too, but have been implemented for free energy perturbation simulations, but not for use in CG or MD to resolve severe steric clashes. Thus, existing minimization protocol using CG/MD method is not robust when applied to structures with severe clashes, which can be resolved using DMD. Furthermore, in all the four cases in which CG/MD does not converge, the DMD minimized structures are acceptable for subsequent MD simulations.

Robustness of DMD in resolving clashes

Minimization of steric clashes with minimal backbone perturbation need not imply a final structure that has formed proper contacts in its buried core. In order to estimate the quality of the structures generated by the different relaxation techniques we use, we enumerate the number of unsatisfied hydrogen bond donors/acceptors in the buried core before and after minimization, which informs us if all polar contacts in the core are well formed. Since a very small number of unsatisfied hydrogen bond donors/acceptors are seen in the core of high-resolution crystal structures, lesser the number of unsatisfied hydrogen bond donors/acceptors in the core, better is the quality of the structure, provided the steric overlaps are absent. In our benchmark set, we find that the numbers of buried unsatisfied hydrogen bond donors/acceptors of the test set of proteins are either maintained or decreased in 19/20, 6/10 and 15/16 cases for DMD, Rosetta and CG/MD respectively (Table II, Supplementary Table IV; CG/MD does not converge in 4/20 cases). If we were to categorize these unsatisfied partners as backbone or sidechain atoms, we find that all the methods are much better at forming new backbone hydrogen bonds, but lose many hydrogen bonds involving side chain atoms. This effect could be due to removal of clashes involving side chain, while at the same time stabilizing secondary structural elements in the protein. We conclude from the analysis of unsatisfied hydrogen bond donors/acceptors that all protocols we test (DMD, Rosetta and CG/MD) maintain the hydrogen bonds in the core of the protein, and thus provide final structures that can be compared on the basis of backbone RMSD to initial structures.

In order to test whether severe steric clashes are the cause for failure of CG during minimization of four of the twenty structures we use for benchmarking, we use the structure generated by DMD upon minimization to run a 2 ps MD simulation using CHARMM in GROMACS. These simulations proceed normally, further establishing the robustness of DMD in minimizing steric clashes. The reason DMD succeeds in minimizing steric clashes in cases where CG fails is that DMD automatically defines soft potentials between clashing atoms, thus accommodating even structures with severe clashes for minimization.

To further test the different protocols, we perform minimization simulations on 25 structures of proteins smaller than 250 residues with the worst clash scores from our swiss-model dataset (see methods). We observe a trend (Supplementary Table VI) similar to that of the low resolution crystal structures: minimization using Rosetta takes longer with the minimized structure having higher RMSD to the starting structure, while DMD and CG/MD are comparable in terms of RMSD, while CG minimization alone is faster. However, there are three structures (Q03EE3, A6L8G0, P49048) where CG/MD does not converge due to severe clashes in the starting structures. Thus, simulations on homology models reiterate that DMD protocol is robust in accepting structures with severe clashes. To evaluate the robustness of these protocols in minimizing larger structures, we perform DMD and CG/MD simulations on 25 structures that are more than 250 residues long with worst clash-scores (Supplementary Table VII). We notice that CG/MD does not converge on a majority of these larger structures, mainly because of the poor quality of these structures. Most of these structures feature bond-lengths that are more than ten standard deviations away from mean, causing CG to fail, and hence we do not use these simulations for arriving at any conclusions.

Our results clearly indicate that Chiron is able to resolve unnatural clashes in low-resolution crystal structures and homology models efficiently with minimum deviation of the protein backbone from the initial structure. However, attempting to refine homology models using simulations may drive the models away from the native state while featuring an RMSD of 1 Å with respect to the input structure. Hence it is important to ensure that the minimization protocol we employ does not result in a structure that is farther away from the native state compared to the initial structure. To verify whether Chiron drives a given structure away from its native state upon minimization, we consider five homology models from CASP8 predictions featuring the worst clash-scores, to compare against their native structures that are now available in the PDB. We perform refinement and analyses similar to those described above on these five models and observe that Chiron is able to resolve clashes from the homology models within 1 Å of the initial structure and yet not drift away from the native structure (Supplementary Tables VIII and IX). We also observe that the other methods we tested present similar trend with respect to the Cα-RMSD from the native state after refinement (Supplementary Tables VIII and IX). We can conclude from this analysis that clash-refinement using Chiron does not drive the structures further away from the native state.

Discussion

The purpose of the current study is to quantitatively characterize the energetic effect of steric clashes on protein structures. Our quantitative definition of clashes and systematic analysis on protein-structures of varying quality reveal that indeed steric clashes are present in all structures including high-resolution crystal structures. Based on our findings, we developed a precise metric using clash-scores of high-resolution structures to quantify the quality of a structure with respect to clashes. To resolve severe clashes robustly, we propose a new dynamics based approach that uses high temperature, high heat exchange simulations to quench high velocities arising due to clashes and. As a part of our protocol, we have successfully used soft-core potentials for clash minimization, which have so far been implemented only for free energy calculations and for enhanced sampling in traditional MD simulations.

As evidenced by our analyses, steric clashes are common structural artifacts of homology models and low-resolution crystal structures. Our elucidation of clashes in a protein structure and the distribution of clash-scores in different proteins enable one to easily determine if a given protein features abnormal steric overlaps. Thus, this study provides an important quality control parameter for protein structure that is based on data from high-resolution crystal structures. Once such abnormality has been identified, one would like to refine their structure to ensure acceptable clash-scores. Though all protein modeling programs have some techniques to resolve steric clashes, we demonstrate that they either take too long or do not converge for structures with severe clashes.

To address the lack of tools to minimize clashes in proteins, we have developed a robust and automated methodology using DMD. Implementing soft-core potentials for clashing atoms in combination with high heat-exchange coefficients, we are able to resolve clashes in proteins with severe clashes. In the process, we also ensure minimum perturbation to the protein backbone. With respect to clash minimization, DMD outperforms Rosetta in terms of RMSD and time taken for minimization. While the CG/MD protocol is faster in minimizing many structures, severe clashes cause the currently available CG/MD protocol to fail, establishing the robustness of DMD. Given these advantages over available programs, we believe that our methodology, available as an easy-to-use web server (http://chiron.dokhlab.org) provides an accessible platform to refine protein structures. We acknowledge that the currently available programs may upon customization be efficient in minimization of clashes. On the other hand, since DMD (and hence Chiron) is transferrable to proteins with any extent of clashes, the users are not required to customize Chiron for their specific structural models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the American Heart Association [grant number 09PRE2090068 to S. R.]; the University of North Carolina Research Council [to F.D.]; and the National Institutes of Health [grant numbers R01GM080742, ARRA supplements GM080742-03S1, GM066940-06S1 to N. V. D]. The authors acknowledge members of Brian Kuhlman’s lab for discussions regarding Rosetta protocols.

References

1. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science. 2003; 302(5649):1364–1368. [PubMed: 14631033]
2. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. Protein structure fitting and refinement guided by cryo-EM density. Structures. 2008; 16(2):295–307.
3. Forster F, Webb B, Krukenberg KA, Tsuruta H, Agard DA, Sali A. Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. J Mol Biol. 2008; 382(4): 1089–1106. [PubMed: 18694757]
4. Schwede T, Sali A, Honig B, Levitt M, Berman HM, Jones D, Brenner SE, Burley SK, Das R, Dokholyan NV, Dunbrack RL Jr. Fidelis K, Fiser A, Godzik A, Huang YJ, Humblet C, Jacobson MP, Joachimiak A, Krystek SR Jr. Kortemme T, Kryshtafovych A, Montelione GT, Moult J, Murray D, Sanchez R, Sosnick TR, Standley DM, Stouch T, Vajda S, Vasquez M, Westbrook JD, Wilson IA. Outcome of a workshop on applications of protein models in biomedical research. Structure. 2009; 17(2):151–159. [PubMed: 19217386]
5. Hooft RWW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature. 1996; 381:272–272. [PubMed: 8692262]
6. Vriend G, Sander C. Quality control of protein models: directional atomic contact analysis. J Appl Cryst. 1993; 26:47–60.
7. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB I, Snoeyink J, Richardson JS, Richardson DC. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nuc Acids Res. 2007; 35:W375–W383.
8. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, Meiler J. Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry. 49(14):2987–2998. [PubMed: 20235548]
9. Feig M, Karanicolas J, Brooks CL 3rd. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graph Model. 2004; 22(5):377–395. [PubMed: 15099834]
10. Rotkiewicz P, Skolnick J. Fast procedure for reconstruction of full-atom protein models from reduced representations. J Comput Chem. 2008; 29(9):1460–1465. [PubMed: 18196502]
11. Ding F, Tsao D, Nie H, Dokholyan NV. Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure. 2008; 16(7):1010–1018. [PubMed: 18611374]
12. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Discrete molecular dynamics studies of the folding of a protein-like model. Fold Des. 1998; 3(6):577–587. [PubMed: 9889167]
13. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem. 1983; 4:187–217.
14. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr. 1998; 54(Pt 5):905–921. [PubMed: 9757107]
15. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235–242. [PubMed: 10592235]
16. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T. The SWISS-MODEL Repository and associated resources. Nucleic Acids Res. 2009; 37:D387–392. Database issue. [PubMed: 18931379]
17. Ding F, Dokholyan NV. Emergence of protein fold families through rational design. PLoS Computional Biology. 2006; 2(7):e85.
18. Yin S, Ding F, Dokholyan NV. Modeling backbone flexibility improves protein stability estimation. Structure. 2007; 15(12):1567–1576. [PubMed: 18073107]
19. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins. 1999; 35(2): 133–152. [PubMed: 10223287]

20. Anderson HC. Molecular dynamics simulations at constant pressure and/or temperature. Journal of Chemical Physics. 1980; 72:10.

21. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. Journal of Computational Chemistry. 2005; 26(16):1701–1718. [PubMed: 16211538]

22. Bjelkmar P, Larsson P, Cuendet MA, Hess B, Lindahl E. Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models. J Chem Theory Comput. 2010; 6:459–466.

23. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins. 2004; 55(2):383–394. [PubMed: 15048829]

24. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys. 2007; 126(1):014101. [PubMed: 17212484]

25. Scott MLG, Kenneth MM Jr. Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables. Journal of Computational Chemistry. 1993; 14(3):349–352.

26. Dunbrack RL Jr. Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 1997; 6(8):1661–1681. [PubMed: 9260279]

27. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009; 77(4):778–795. [PubMed: 19603484]

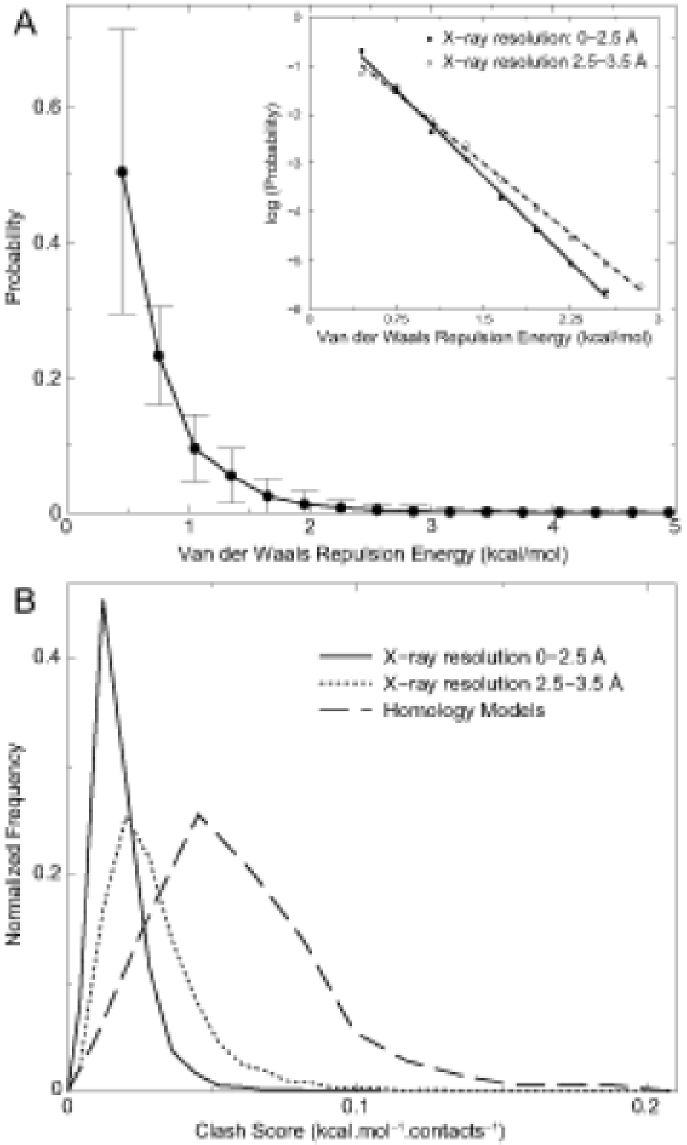

Figure 1. Clash score is dependent on the resolution of the structure

Distribution of the energy of different clashes in a structure shows that most of the clashes are low in energy, with only a few clashes having high repulsive energy (**A**). The probability of the energy of clashes has an exponential distribution (inset). Comparing the distribution of the clash score of structures of high-resolution (between 0 and 2.5 Å), low-resolution (between 2.5 and 3.5 Å) and homology models reveals that higher the resolution of the structure, lower the clash score (**B**). The Gaussian fits for the different distributions are also shown as solid line (high-resolution), dotted line (low-resolution) and dashed line (homology models).

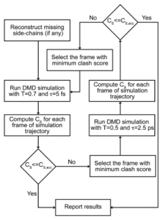
No
C_S<=$C_{S,acc}$
Yes
Reconstruct missing side-chains (if any)
Select the frame with minimum clash score
Run DMD simulation with T=0.7 and τ=5 fs
Compute C_S for each frame of simulation trajectory
Compute C_S for each frame of simulation trajectory
Run DMD simulation with T=0.5 and τ=2.5 ps
C_S<=$C_{S,acc}$
No
Select the frame with minimum clash score
Yes
Report results

Figure 2. Flow chart for clash minimization using DMD

A typical clash minimization cycle is represented as a decision-making sequence of events with alternating cycles of DMD simulations with different parameters. Results are reported upon successful minimization of clashes. C_S refers to clash-score, $C_{S,acc}$ refers to the acceptable clash-score, T refers to simulation temperature in ε/k_B and τ refers to the heat exchange coefficient.

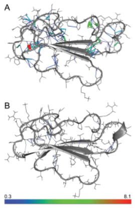

Figure 3. Steric clashes in a representative protein before and after DMD minimization

PDB ID 1CTX is shown before (A) and after (B) DMD minimization. Clashes are represented as cylinders connecting clashing atoms. The thickness of the cylinder and the color correspond to the severity of the clash. Clashes with highest energy are colored red, while the ones with lowest energy are colored blue. The color bar at the bottom indicates the range of clash energies in kcal/mol. Similarly, clashes with highest energy have the largest cylinder radius. It can be observed that DMD minimization removes all the severe clashes seen in the PDB structure.

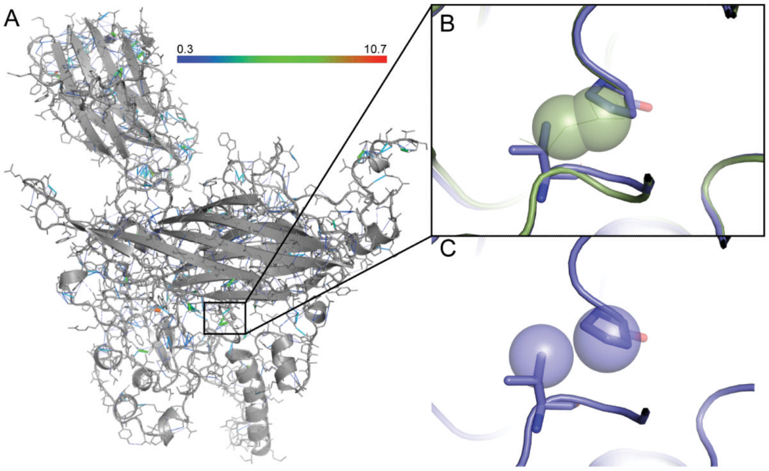

Figure 4. An example of a severe clash that leads to failure of CG minimization

PDB ID 1GFF does not converge in CG simulations due to infinite force at P138. We observe severe clashes between CG1 of V20 from a bound peptide and CD, CG atoms of P138 (**A**). One of these clashes is represented as spheres (**B**); the structure from PDB is colored green, while the structure obtained after DMD minimization is colored blue. DMD minimization completely removes these clashes (**C**).

Table I

Performance of Chiron compared to Rosetta and CHARMM

PDB ID	Size[1]	R[2]	Initial Clash Score[3]	Chiron			Rosetta			CG/MD		
				Final Clash Score[3]	D[4]	Total Time [5]	Final Clash Score[3]	D[4]	Total Time[5]	Final Clash Score[3]	D[4]	Total Time[5]
1CTX	71	2.80	16	1.98	0.48	3 m	3.50	1.35	2h 57m	1.78	0.41	2 m
1PFC	111	3.13	20	1.89	0.58	10 m	1.58	1.16	3h 28m	2.00	0.28	1 m
2ABX	148	2.50	17	1.99	0.49	7 m	3.40	2.79	7h 19m	1.92	0.50	5 m
1PY4	388	2.90	13	1.91	0.28	15 m	1.39	1.54	1d 15h	1.98	0.17	3 m
1MCW	430	3.50	14	1.99	0.51	27 m	1.85	1.47	1d 22h	1.98	0.29	1 m
1CN1	474	3.20	15	1.99	0.34	19 m	1.82	0.94	1d 23h	2.00	0.29	3 m
2ZIX	529	3.50	13	1.99	0.33	21 m	1.68	1.97	2d 9h	1.88	0.23	3 m
1TMF	806	3.50	22	1.98	0.34	35 m	1.60	2.35	6d 5h	Not converged		
1R24	846	3.10	13	1.96	0.28	37 m	1.57	1.19	6d 9h	1.88	0.18	9 m
4GPD	1332	2.80	13	1.98	0.31	66 m	1.81	1.20	13d16h	Not converged		

[1] Number of amino acids in the protein

[2] Resolution of the crystal structure in Å

[3] Normalized clash-score x 10^2

[4] Cα RMSD in Å

[5] Total user time taken to complete the simulations on a single processor. d - days ; h - hours ; m - minutes

Table II

Number of unsatisfied hydrogen bonds before and after minimization

PDB ID	Size[1]	Number of Unsatisfied Hydrogen Bonding Partners											
		Initial Structure			Chiron			Rosetta			CG/MD		
		BB[2]	SC[3]	Total[4]	BB[2]	SC[3]	Total[4]	BB[2]	SC[3]	Total[4]	BB[2]	SC[3]	Total[4]
1CTX	71	4	3	7	4	1	5	4	2	6	2	0	2
1PFC	111	4	1	5	0	1	1	4	4	8	1	1	2
2ABX	148	8	0	8	5	3	8	9	7	16	4	2	6
1PY4	388	27	9	36	10	11	21	19	15	34	10	10	20
1MCW	430	32	17	49	31	10	41	14	14	28	24	12	36
1CN1	474	88	26	114	61	33	94	54	15	69	54	18	72
2ZIX	529	6	5	11	12	9	21	29	5	34	15	6	21
1TMF	806	77	33	110	70	31	101	93	29	122	Not Converged		
1R24	846	82	28	110	44	27	71	33	28	61	26	27	53
4GPD	1332	193	55	248	175	62	237	135	61	196	Not Converged		

[1] Number of amino acids in the protein, Polar atoms belonging to the backbone

[2] or side-chains

[3] do not form hydrogen bonds and belong to residues that are buried

[4] total number of unsatisfied hydrogen bonding partners in the protein