Solvent dramatically affects protein structure refinement

Gaurav Chopra, Christopher M. Summa, and Michael Levitt

Department of Structural Biology, Stanford University, Stanford, CA 94305-5126; and Department of Computer Science, The University of New Orleans, New Orleans, LA 70148

Contributed by Michael Levitt, October 30, 2008 (sent for review October 11, 2008)

One of the most challenging problems in protein structure prediction is improvement of homology models (structures within 1–3 Å Cα rmsd of the native structure), also known as the protein structure refinement problem. It has been shown that improvement could be achieved using in vacuo energy minimization with molecular mechanics and statistically derived continuously differentiable hybrid knowledge-based (KB) potential functions. Globular proteins, however, fold and function in aqueous solution in vivo and in vitro. In this work, we study the role of solvent in protein structure refinement. Molecular dynamics in explicit solvent and energy minimization in both explicit and implicit solvent were performed on a set of 75 native proteins to test the various energy potentials. A more stringent test for refinement was performed on 729 near-native decoys for each native protein. We use a powerfully convergent energy minimization method to show that implicit solvent (GBSA) provides greater improvement for some proteins than the KB potential: 24 of 75 proteins showing an average improvement of >20% in Cα rmsd from the native structure with GBSA, compared to just 7 proteins with KB. Molecular dynamics in explicit solvent moved the structures further away from their native conformation than the initial, unrefined decoys. Implicit solvent gives rise to a deep, smooth potential energy attractor basin that pulls toward the native structure.

Experimental determination of protein structures is very expensive, costing U.S. $250,000 in 2000 (1) and $66,000 today (2) and can be a notoriously difficult task, especially for membrane proteins. With the continuing exponential growth of genome sequence data, there is an increasing need for methods that accurately compute the high-resolution native structure of a protein, for use in biological applications that include virtual ligand screening (3), structure-based protein function prediction (4) and structure-based drug design (5). Homology or template based modeling has been the most successful method for protein structure prediction in the critical assessment of protein structure prediction (CASP) experiments (6, 7). The power of this technique progressively increases as ever more structures are solved by world-wide structural genomics initiatives (8, 9). Nevertheless, obtaining a model with the same accuracy as a crystal structure is still an unsolved problem: structure refinement of a rough model (within 1–3 Å rmsd) to bring it closer to the native structure remains a major challenge (6, 10). Work on structure refinement has been ongoing for many decades, starting from the first Molecular Mechanics (MM) energy minimization (11, 12) and continuing to a recent study with knowledge-based (KB) statistically derived potentials (13). During this period many different potentials and a variety of simulation methodologies such as energy minimization, molecular dynamics, and replica exchange Monte Carlo have been used for structure refinement (14–20), but no method has emerged as a clear winner.

As protein molecules function in aqueous solution and crystals contain large amounts of water (21–23), it is appropriate to model the water environment for high resolution refinement of protein structures using explicit, implicit, or hybrid models. The most realistic way to include solvent effects is to immerse the protein in a periodic box of explicit water molecules and simulate the motion of the system by molecular dynamics (MD) as first done by Levitt and Sharon (24). Unfortunately, atomic motion inherent in MD introduces statistical noise that can only be removed by averaging over the many conformations generated to get a final refined structure. Solvent effects can also be included implicitly, where water is represented as a continuous medium and the effect of the solvent is represented by additional terms in the potential energy function of the protein. Because there are no explicit water molecules, energy minimization can be used in place of MD and there is no need to average over many conformations.

The many different ways to include solvent effects have been reviewed in (25, 26). Such methods extend from the early accessible surface areas (ASA) model by Lee and Richards (27), to the widely used generalized born surface area (GBSA) model (28, 29) and more recently to the screened coulomb potential implicit solvation model (30, 31). Even though implicit solvent models are less physically realistic than use of explicit solvent, their greater computational efficiency makes them an attractive choice for refinement. MD has been used for refinement but with limited success (14–17), although more recently better results was observed for refinement with spatial restraints (18–20). Such successes have been reported for a few isolated cases; as the methodologies can be computationally demanding, it is difficult to apply them broadly. Clearly, one needs to consider computationally less demanding refinement protocols involving minimization with MM and KB force fields that include implicit solvation.

In this work we test the role of solvent on protein structure refinement using energy minimization and MD on an extensive test set of 75 native proteins, each with 729 near-native decoys. This follows on from previous work (13), which used in vacuo energy minimization to compare the performance of various MM potentials such as optimized potential for liquid simulations (OPLS)-AA, AMBER99, GROMOS96, and ENCAD with a statistically derived KB potential. They found that the KB potential performed best of all and refined almost all proteins toward the native structure; AMBER99 was second-best, performing better than the other MM potentials, which generally moved the decoys away from the native state. The KB potentials include the effect of solvent implicitly, in that the distribution of distances between atoms in protein crystals is affected by the water in the unit cell.

We model implicit solvent, using energy minimization with the GBSA implicit solvent model (29), as implemented in Tinker 3.9
Energy Minimization of Native Structure. Energy minimization was run on all 75 native proteins in implicit and explicit solvent. These results were compared to in vacuo energy minimization, using the KB potential (13). In all cases, we used the weighted Cα rmsd (wRMS) as a measure of deviation from the native structure (see Materials and Methods) to compensate for flexibility of loops and chain termini (Fig. 1). For all 75 native proteins, the mean wRMS value (see Fig. S1) was 0.89 ± 0.36 Å for GBSA implicit solvent, 0.38 ± 0.14 Å for KB, and 0.14 ± 0.01 Å for OPLS bulk explicit solvent. Although it may seem that bulk explicit solvent is working best as it perturbs the native structure least, this is not true. Energy minimization fails to move away from the native structure because the bulk explicit solvent acts like ice and greatly restricts movement. This example indicates why energy minimization from the native structure cannot be used to assess potentials: complete lack of convergence would appear to be perfect behavior with a wRMS value of zero. It justifies our much more extensive tests.

**Assessing Decoys Using wRMS and GDT-HA.** As the rmsd value is sensitive to large shifts of a few atoms of a molecule, we use the wRMS and global distance test (GDT) score for a more robust measure of structural similarity measure (see Materials and Methods). We computed high accuracy GDT scores (GDT-HA) (42) for
observed for KB, confirming wRMS as a good structural similarity measure. For 1nk, a good case with KB, the GDT-HA curve for KB is not significantly higher than GBSA (Fig. 2C) and both GBSA and KB improve the structure equally well for higher values of GDT-HA. Nevertheless, 1nk has a PC value of $-33.2\% \pm 19.6\%$ with KB, which is better than the value of $-19.2\% \pm 23.2\%$ for GBSA. Moreover, the minimum value of GDT-HA for 1nk is 0.7, which is >0.5 for 1pne. Thus, similar improvement with GDT-HA is seen for good cases of GBSA and KB when most residues match well (high GDT-HA_initial Value) but GBSA outperforms KB when low fraction of residues match the native with high accuracy.

**Molecular Dynamics vs. Energy Minimizations of Decoys.** As MD is computationally expensive, we selected a subset of 20 of the 75 proteins, chosen to minimize bias with a mix of PC values for KB energy minimization. The selected proteins were: 1ail, 1bkra, 1c1ka-1, 1dsl, 1ge8a02, 1gvd, 1h99a1, 1ift, 1ift, 2li, 1lwba, 1lwba, 1mt7a, 1mgta, 1int, 10bxa, 1pdo, 1qvn, 1tm, 1whi and 4euq0. MD was run for 200 ps on all 729 decoys of each of these 20 proteins (see SI Text). With a total simulation time of 2,916 ns, this represents a large increase in computational resources compared with energy minimization. The mean initial wRMS for all decoys of these 20 proteins was 1.02 ± 0.15 Å, which was not too different from the value of 1.08 ± 0.14 Å for all 75 proteins. Fig. 3 compares MD with energy minimization with KB and GBSA.

We observe that KB minimization gives most improvement over all these 20 proteins with (PC) value of $-9.6\%$ compared with +7.14% with GBSA. However, GBSA improved 9 of 20 proteins with a (PC) value of $-18.74\%$ compared with $-10.75\%$ for 19 proteins with KB (see plot in Fig. 3A). There is almost no change for OPLS explicit solvent MD at 0 ps, with <wRMS> of 1.01 Å; these structures are a result of the preparatory energy minimization in the presence of explicit solvent and move very little. Specifically, for explicit water, the (PC) value at 0 ps was $-0.32\%$ for all 20 proteins with an improvement of $-1.32\%$ for 10 proteins. Unfortunately, additional MD moves the structure far away from the native (see Fig. 3A). The (PC) value for all proteins at 100 ps was 21.30% with <wRMS> of 1.22 Å, which is much worse than the (PC) value after preparatory energy minimization of $-0.32\%$. More simulation makes the situation worse: at 200 ps the (PC) value is 29.76% with final <wRMS> of 1.31 Å. Thus, on average, decoys move away from native as the MD trajectory progresses from 100 to 200 ps (see Fig. S2). In contrast, 4 of 20 proteins improved with MD, in that the decoys moved closer to the native state. These are 1lwba, 1wi, 1bkra and 1ift with a (PC) value of $-15.41\%$ at 100 ps and $-13.05\%$ at 200 ps. Even in these “good” cases most improvement occurs in the first 100 ps and the situation deteriorates with additional MD (see Fig. S2).

**Movement of Decoys on GBSA Potential Energy Surface.** The basic assumption for all of the refinement methodologies described here is that the native structure occurs at a minimum of the potential energy and is surrounded by a smooth attractor basin. How does solvent influence the downhill energy path from a decoy to the native state? Fig. 4 shows the near-native potential surface for GBSA energy minimization. For good cases with GBSA (1kpta, 1hn4, and 100xa), the native state is located at a well-defined minimum of the potential energy surface (pink disk in Fig. 4). Energy minimization moves toward the native state and results in a low wRMS value; there is a clear attractor basin, which is most apparent for 1kpta, where the final energy minimized decoys cluster together and come very close to the native state. For these good cases, GBSA solvation gives rise to a well-defined energy basin with the native structure at its minimum; initial decoys on hills surrounding the basin move toward the native state upon minimization. The topography of the energy surface limits the improvement possible for a particular method. This is shown for the bad cases with GBSA (1ge8a01, 1h99a1, and 1pdo), where the native struct-

---

**Fig. 2.** Comparison of the GDT-HA scores for energy minimization with KB (solid) and GBSA (dashed). (A) The change in GDT-HA between final minimized structure and the initial decoy GDT-HA_final - GDT-HA_initial is plotted against GDT-HA_initial for all 729 decoys of each of the 6 proteins that show large negative percentage change (PC) for the KB and GBSA potentials. Values of GDT-HA_initial - GDT-HA_initial are averaged over GDT-HA_initial bins of width 0.1. A higher value of GDT-HA_final - GDT-HA_initial indicates an improvement in the decoy (gets closer to the native structure). Overall (GDT-HA_final - GDT-HA_initial) is higher for GBSA indicating that this method improves the decoy structure more than KB. (B) Showing 1pne, a good case with GBSA, that is less good with KB. The native structure of 1pne is shown in gray, the best decoy minimized with KB in red, and the best decoy minimized with GBSA in purple. GBSA gives a better match to the native structure than does KB. (C) Showing 1nk, a good case with KB. The lowest GDT-HA_initial value is better (~0.7) than for 1pne (~0.5). For 1nk, GBSA and KB do equally well.

---

Chopra et al.
tecture is not in the major energy basin containing all energy minimized decoys for 1ge8a01 (Fig. 4). For 1h99a1 and 1pdo, the native is located in a valley surrounded by one or more hills separating it from the major energy basin. For these bad cases, the nature of the energy surface causes large shifts from the native structure upon minimization. Often energy minimization moves from the initial decoys to a region far from the native state but with similar energy values as the native state. For bad cases, the region in which minimized decoys cluster is much larger than it is in the good cases, which often show a compact cluster.

Discussion

Improvement upon Minimization with GBSA Implicit Solvent. Energy minimization in GBSA implicit solvent yields the largest negative PC and performs refinement better than both KB energy minimization and MD in explicit solvent. Moreover, GBSA energy minimized decoys correctly identify the native basin by clustering together, as shown by the good cases in Fig. 4. An X-ray structure is not perfectly accurate and it is useful to estimate when a near-native structure cannot be distinguished from the actual native state. Two studies (43, 44) have estimated the “accuracy limit” of the crystal structures, i.e., the maximum coordinate deviation after superposition of structures of the same protein determined experimentally in different crystal packing arrangements. These were estimated to be between 0.8 Å cRMS (44) and 0.95 Å (43) rmsd for heavy atoms. Thus, we can assume that a model within 0.8 Å rmsd is indistinguishable from the native state. For a very favorable case with GBSA minimization (1kpta), the PC value is $-55.1\% \pm 24.6\%$, (wRMS$_{\text{final}}$) is 1.15 Å and (wRMS$_{\text{final}}$) is 0.51 Å, which is indistinguishable from the native state. A value of (wRMS$_{\text{final}}$) <0.8 Å was observed for 18 proteins by GBSA energy minimization and 12 proteins by KB energy minimization. Moreover, GBSA outperforms KB when tested on the near-native decoys, which were energy minimized, using the ENCAD potential (40) to remove any bad contacts due to the decoy generation procedure (see SI Text, and Figs. S3 and S4). Clearly minimization with GBSA implicit solvent is an excellent refinement method.

Along with the good examples that show large improvement with GBSA, there are some very bad cases where decoys move far away from the native state upon minimization. What are the reasons for these bad cases? Is it a failure of the minimization method or of the potential function? Because GBSA minimization can cause large movements toward the native state, we are convinced that Tinker 3.9 uses a very good minimization protocol. Performance seems to depend on the accuracy with which GBSA implicit solvent mimics the solvent present in the crystals. For the good cases in Fig. 4 we see that the energy basin is smooth and has the native state at the minimum. For the bad cases in Fig. 4 we see a flat energy landscape where no correlation exists between the energy and the distance to
the native state. Such bad cases seem to occur when the protein chain has long loops and chain termini with high B-factors (1ge801, 1h99a1 and 1pdo in Fig. 1). We believe that a more accurate representation of the energy surface is needed for these bad cases with GBSA. It is also possible that crystal contacts affect the position of these chain segments. It seems clear that GBSA implicit solvation removes the ruggedness around the native state by giving rise to a smooth energy basin in good cases or a flat energy landscapes in bad cases.

**Molecular Dynamics in Explicit Solvent Moves Away from Native.** For the subset of 20 of 75 proteins simulated using MD in explicit solvent, only 4 proteins were refined (PC < 0) by MD compared to 9 with GBSA and 19 with KB energy minimization. For all 20 proteins, the mean PC value increased from −0.32% to 21.30% from 0 ps to 100 ps of MD; worsening to 29.76% was observed from 100 ps to 200 ps (Fig. 3A). Additional MD simulations of 10 ns in explicit solvent also moves the decoys away from the native structure (see SI Text, Fig. S5, and Table S1). This does not prove that MD tends to deform the native state; with sufficient averaging over long runs the native state might be reached, as shown for few isolated cases (14–17) where the refined structures were ~2 Å rmsd from native. MD is a very popular method used to describe pathways of folding and unfolding. This technique is fundamentally different from energy minimization, in that MD introduces random noise and can get stuck in a minimum, which could be far from the native state. Given infinite time, it could find the native state provided it was in a sufficiently deep energy basin. We conclude that use of MD with periodic boxes of explicit water is not a good refinement method in that it is out-performed by both Tinker GBSA implicit solvent and Encad KB minimization.

It has recently been shown that MD does not help discriminate the native state from its decoys, in that the correlation between energy and rmsd vanishes after MD simulation with AMBER and a GB potential (45). Other recent work used replica exchange sampling with CHARMM22/GBSW potential to refine homology modeling targets with spatial restraints; it worked well but it is unclear whether this would have happened without the restraints (20). The use of accurate potentials with the native state as a global minimum is a necessary but not sufficient condition for extensive sampling methods to work for refinement. MD simulation is strongly dependent on random thermal perturbations involved in heating and we expect it to move all over a flat energy surface (Fig. 5). In addition, it seems that the potential surface used in MD is very rugged as can be seen by the energy minimized decoys (pink dots) in Fig. 5 (a plethora of local minima around the initial decoys). It seems that we need better sampling or significantly more computer time for MD to move out of these local minima.

**Nature of Energy Surfaces for Various Potentials.** Our 2D visualization of the energy surface for a representative set of decoys gives...
Materials and Methods

Weighted Cα rmsd and High Accuracy GDT. Normally protein structural deviation from the native state is measured by the rmsd of coordinates after optimum rigid body superposition of all atoms. This measure can overcome the problem of flat landscapes. It might also indicate parts of the protein that are not well-defined in the absence of the crystal lattice.


Acknowledgments. We thank Nir Kalisman for help with GDT-HA score calculation and the reviewers for a very careful review. This work was supported by National Institutes of Health Grant GM63817 (to G.C., C.M.S., and M.L.). Simulations were done on the BioX2 supercluster and supported by National Science Foundation Grant CNS-0619926.