Sub-Angstrom Accuracy in Protein Loop Reconstruction by Robotics-Inspired Conformational Sampling

Daniel J. Mandell, Evangelos A. Coutsias and Tanja Kortemme

Proteins exploit the conformational variability of loop regions to carry out diverse biological tasks including molecular recognition, signal transduction, and active site gating. New algorithms to engineer these functions by combining loop building and sequence design therefore have enormous practical applications, but require high-resolution *loop reconstruction*: the modeling of protein loop conformations given amino acid sequences. Despite significant progress in loop prediction[1-3], more accurate methods to sample and evaluate the conformational space accessible to loops are still a major bottleneck for high-resolution protein modeling. Loop reconstruction in protein design may be simplified conceptually by restricting changes to the functional loop regions, but this has been limited by both the difficulty to model purely local conformational moves and by the need for accuracy beyond what is generally achievable.

Here we address these key challenges by presenting a robotics-inspired local loop reconstruction method for peptide chains of any length, called kinematic closure (KC). Calculating the accessible configurations of objects subject to constraints, such as determining the possible positions of the interior joints of a robot arm given fixed positions for the shoulder and fingertips, has been well-studied in the field of inverse kinematics, a subfield of robotics. These techniques were first applied to proteins[4] by calculating the accessible torsion angles of tripeptides with fixed bond angles, bond lengths, and endpoints. This and other kinematics-inspired formulations[5] required numerical optimization and did not give all solutions for tripeptide closure in a single run. An analytical solution for the closure of 6 backbone torsion angles[6] could only be directly applied to tripeptides. The KC method presented here provides the key advantages of analytically determining all mechanically accessible conformations for 6 torsions of a peptide chain of any length, while simultaneously sampling the remaining torsions and backbone N-Cα-C bond angles based on a new method using polynomial resultants[7] (**Supplementary Methods** and **Supplementary Figure 2**). To enable a range of applications of KC, we couple it to the powerfully predictive Rosetta method for protein structure modeling[8]. The loop reconstruction protocol consists of a series of KC calculations (**Fig. 1a** and **Supplementary Figure 1**) comprising Monte Carlo (MC) moves in a simulated annealing protocol in Rosetta that is iterated with loop backbone minimization in a low-resolution stage, and iterated in a high-resolution all-atom stage with minimization of the loop backbone and the side-chains in the loop environment (**Fig. 1b** and **Supplementary Methods**). Before loop reconstruction, we eliminate all native side-chain information in both the loop and the protein scaffold and replace the side-chains with simultaneously optimized conformations from a rotamer library. At the beginning of each KC simulation, all native loop bond lengths, bond angles and

torsions are discarded; the bond lengths are then set to ideal values, and loop backbone N-Cα-C bond angles and torsions are sampled (**Supplementary Methods**).

We found that KC substantially improves model accuracy over the standard loop building method in Rosetta, which combines insertion of torsion segments from homologous proteins and a numerical closure technique[5], and is part of Rosetta's structure prediction protocol[9] used in the high-resolution design of a protein loop[10]. (KC also compared favorably to the state-of-the-art molecular mechanics method, as further described below). We generated 1,000 models by KC with 720 low-resolution steps followed by 720 high-resolution steps, and compared the performance to that obtained when we applied the standard Rosetta method with the same number of steps to each of 25 12-residue protein loops from a previously described benchmark set[11] (dataset 1). For each protein, we computed the root mean squared deviation (rmsd) of the backbone atoms (N, Cα, C, O) of the best scoring loop model to the crystallographic loop, after superimposing the non-loop regions of the model onto the crystal structure. Notably, the KC protocol frequently sampled regions of conformational space that were less than 1.0Å from the crystallographic loop (**Fig. 1c**), which were not sampled by the standard protocol. In the majority of cases (15/25), these conformations very close to the crystallographic loop could be identified as the best scoring models (**Fig. 1c** and **d**). Over the entire 25-loop set, KC improved the median accuracy to 0.8Å rmsd from the 2.0Å rmsd obtained using the standard method (**Fig. 2b left panel** and **Supplementary Table 1**).

To further quantify the improvements in conformational sampling by KC, we examined the sources of error for the cases where the best scoring structure was ≥1.0Å rmsd from the crystallographic loop using the KC protocol (10 of 25 loops) and the standard protocol (18 of 25 loops). In cases where the best scoring ≥1Å rmsd model scored worse than the crystallographic loop (subjected to a short relaxation protocol), the error results at least in part from poor conformational sampling, since the scoring function could have correctly identified conformations near the crystallographic loop had they been sampled (see **Supplementary Discussion** for details). There were 16 cases where insufficient sampling led to ≥1.0Å rmsd reconstructions using the standard protocol, versus 5 cases using KC (**Supplementary Table 4**). Sampling errors cannot be considered entirely independently from scoring errors, since the scoring function guides the simulation trajectories, but since the same scoring function is used for both methods, these results suggest that KC increases accuracy by improved conformational sampling. Other potential sources of error are detailed in **Supplementary Tables 6** and **7** and illustrated in **Supplementary Figures 3** and **4**. We also note that the dataset (as dataset 2 below)

was filtered for cases with ligands or ions contacting the loop[2], which were not modeled by the methods.

To assess the performance of the Rosetta KC loop reconstruction protocol with respect to other methods, we compared the KC protocol to the state-of-the-art molecular mechanics method[2]. For direct comparison, the Rosetta KC and standard protocols were applied to the published 20 12-residue starting structures with perturbed loops and side-chain environments used to assess the molecular mechanics method[2] (dataset 2), rather than starting from randomized loop conformations as in **Figure 1b**. A representative set of KC reconstructions from this dataset is shown in **Figure 2a**. The Rosetta KC protocol improved median accuracy to 0.9Å from 1.2Å using the molecular mechanics method and from 2.0Å using the standard Rosetta method (**Fig. 2b middle panel** and **Supplementary Tables 2** and **5**).

Notably, both the Rosetta and the molecular mechanics[2] methods perform reconstructions without knowledge of the native side-chain conformations surrounding the loop (**Supplementary Methods**), which makes prediction substantially more challenging, but broadens the range of applications to designing new loop conformations that may interact differently with neighboring side-chains (**Supplementary Discussion**). We note that applications to comparative modeling may be even more challenging, as the loop endpoints and surrounding backbones can also be substantially perturbed, which we do not consider here. This does not preclude the application of KC in high-resolution refinement and comparative modeling, as shown by a successful example of using our Rosetta KC method in the most recent CASP experiment (Vatsan Raman, Rhiju Das & David Baker, personal communication). Although our implementation of Rosetta KC facilitates efficient bond angle sampling while guaranteeing exact loop closure, bond angle sampling only had a minor role for loop reconstruction accuracy (**Supplementary Table S8**).

Functional loops in antibodies and signaling proteins in complex with their partners exhibit conformational plasticity against a relatively structured core. To assess the ability of KC to model such functional loops, we applied the method to interface loops from 4 proteins (Rac, Ras, CDC42, and ubiquitin) crystallized with 18 different partners (dataset 3). KC reconstructed the loops to 0.8Å median rmsd to the crystallographic loops across the set (**Fig. 2b right panel** and **Supplementary Table 3**). Notably, the KC protocol produced high-accuracy reconstructions of the same loop in the GTPase Rac crystallized in different conformations when bound to different partners (**Fig. 2c).** Moreover, the KC prediction accuracy is higher than simply taking the loop from a crystal structure of the same protein bound to another partner (**Supplementary Table 3** and **Supplementary Discussion**). This result highlights the potential of our method in refinement applications (predicting a conformation closer than the template structure) and also for

modeling loop changes in important conformational switch proteins. Given the diverse biological roles played by protein switches like Rac, sub-angstrom loop reconstructions by the local, analytic sampling protocol described here could be coupled with the successful Rosetta design method[12,13] to enable the modeling and engineering of these versatile systems. Further, reconstruction and design could be harnessed to reshape or elongate interface loops precisely matching a particular binding partner, creating highly specific complexes for use as protein biosensors or biotherapeutics.

The described state-of-the-art loop reconstruction method is available free-of-charge as a module of the academic release version of the Rosetta program for protein modeling and design, at http://www.rosettacommons.org. (Will be released upon publication)

This work was supported by grants from the NIH to T.K. (PN2-EY016525) and E.A.C (R01-GM08171), the UC Lab research program (T.K.), and a PhRMA Foundation Predoctoral Fellowship (D.J.M.). We would like to thank David Baker and Andrej Sali for critical reading of the manuscript and valuable comments, and Ben Sellers and Matt Jacobson for sharing data and for helpful discussions.

References

1. Jacobson, M.P. et al. *Proteins* **55**, 351--367 (2004).
2. Sellers, B.D., Zhu, K., Zhao, S., Friesner, R.A., and Jacobson, M.P. *Proteins* **72**, 959--971 (2008).
3. Felts, A.K. et al. *Journal of Chemical Theory and Computation* **4**, 855--868 (2008).
4. Go, N. and Scheraga, H.A. *Macromolecules* **3**, 178--187 (1970).
5. Canutescu, A.A. and Dunbrack, R.L.J. *Protein Sci* **12**, 963--972 (2003).
6. Wedemeyer, W.J. and Scheraga, H.A. *Journal of Computational Chemistry* **20**, 819--844 (1999).
7. Coutsias, E.A., Seok, C., Wester, M.J., and Dill, K.A. *International Journal of Quantum Chemistry* **106**, 176--189 (2005).
8. Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., and Baker, D. *Science* **310**, 638--642 (2005).
9. Qian, B. et al. *Nature* **450**, 259--264 (2007).
10. Hu, X., Wang, H., Ke, H., and Kuhlman, B. *Proc Natl Acad Sci U S A* **104**, 17668--17673 (2007).
11. Wang, C., Bradley, P., and Baker, D. *J Mol Biol* **373**, 503--519 (2007 Oct 19).
12. Kuhlman, B. et al. *Science* **302**, 1364--1368 (2003).
13. Kortemme, T. et al. *Nat Struct Mol Biol* **11**, 371--379 (2004).

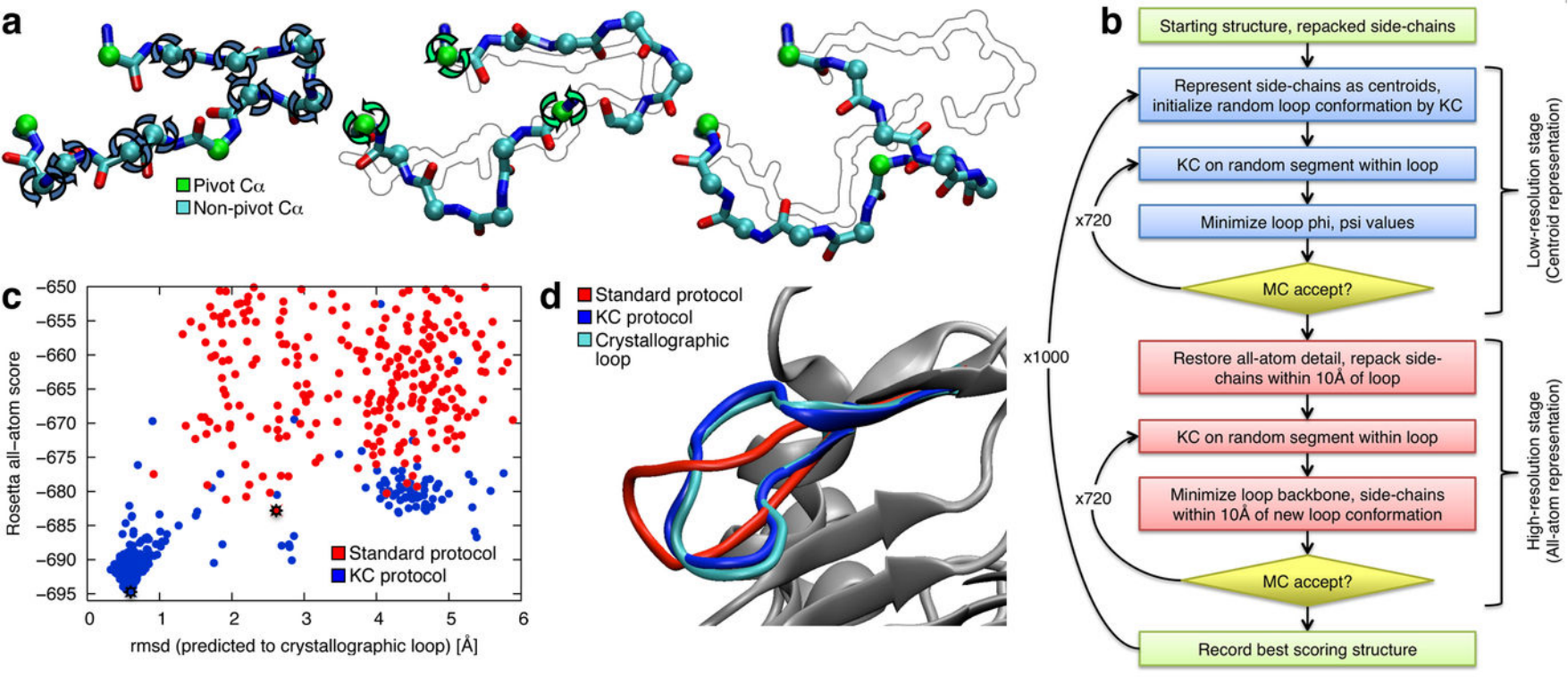

Figure 1 | Loop reconstruction with kinematic closure (KC). (**a**) The KC move. 3 Cα atoms of an *N* residue chain are designated as pivots (green spheres, arrows), and the remaining *N* - 3 are non-pivot Cα atoms (cyan spheres, arrows). The chain is opened by randomly assigning new values to the non-pivot torsions according to the Ramachandran map for each residue type. KC then finds all values of the pivot torsions that close the loop, if any exist, keeping the endpoints fixed. The previous state is shown in outline. In a 12-residue loop (shown) 24 torsions are adjusted (6 pivots torsions and 18 non-pivot torsions). (**b**) The KC loop reconstruction Monte Carlo (MC) protocol in Rosetta. "Repacking" is simultaneous optimization of side-chain conformations by Metropolis MC simulated annealing. (**c**) Performance of the Rosetta KC and standard protocols on a 12-residue loop from Serratia protease (pdb 1srp). Only KC densely samples regions <1.0Å from the crystallographic loop. Asterisks mark the lowest scoring reconstructions from the two methods. The standard protocol required ~280 CPU-hours per protein, KC required ~320. (**d**) Improved sampling by KC increases reconstruction accuracy (0.6Å versus 2.6Å using the standard protocol). The lowest scoring reconstructions from (c) are shown.

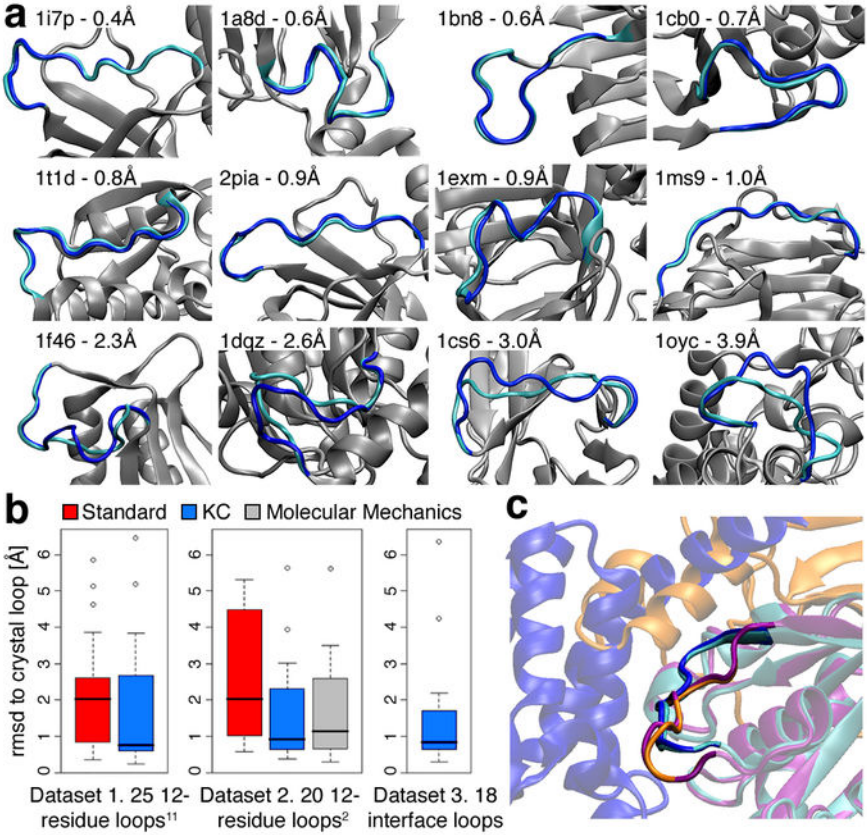

Figure 2 | Performance of the KC loop reconstruction protocol. (**a**) Representative set of 12-residue loop reconstructions (blue) on dataset 2. Pdb codes and rmsd to the crystallographic loop (cyan) are shown. (**b**) Boxplot comparison of the standard Rosetta and KC Rosetta protocols on dataset 1 (left panel), both Rosetta protocols with the molecular mechanics method on dataset 2 (middle panel), and the KC Rosetta protocol on dataset 3 (right panel). Boxes span the interquartile range (IQR, 25th-75th percentiles), black lines represent the median, whiskers extend to furthest values within 0.8 times the IQR, and open circles are outliers. The perturbed starting loop conformations from reference[2] are used across all methods in the middle panel. (**c**) KC reconstruction of conformational changes in the Rac switch I loop when bound to Pseudomonas aeruginosa ExoS toxin (blue reconstruction on cyan crystal structure, blue partner, pdb 1he1) and Rho GDI (orange reconstruction on purple crystal structure, orange partner, pdb 1hh4).