Supplementary Information for

TopoLink: Evaluation of structural models using chemical cross-linking distance constraints

Állan J. R. Ferrari[1], Milan A. Clasen[2], Louise Kurt[2], Paulo C. Carvalho[2], Fabio C. Gozzo[1], Leandro Martínez [1,3,*]

[1]Institute of Chemistry, University of Campinas, [2]Carlos Chagas Institute, Fiocruz, Brazil. and [3]Center for Computational Engineering & Science, University of Campinas, Campinas, SP, Brazil

*leandro@iqm.unicamp.br

Computing solvent accessible topological distances

The computationally intensive part of the calculation is the evaluation of topological distances. The topological distances are obtained, here, by the solution of an optimization problem consisting on the minimization of a linker length subject subject to not overlapping with the protein atoms. A sketch of the model used for the definition of the optimization problem is shown in Figure 1. This strategy to compute surface-accessible distances retains the physical character of the linker in terms of atoms and, as we will discuss, is adaptive to specific structural properties of different types of residues.

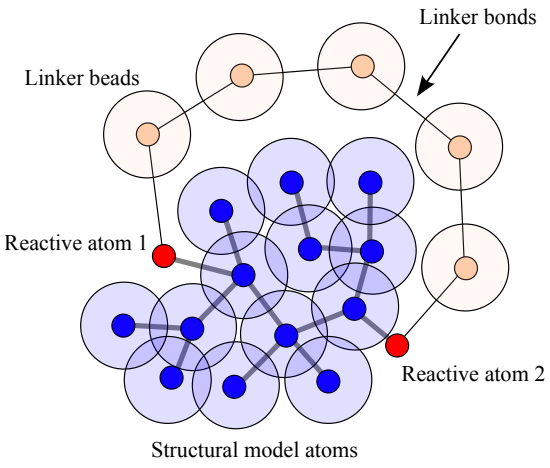

Figure 1: Representation of the model used for topological distance determination. The linker is represented by a series of beads attached on extremities to the reactive atoms. The beads are sequentially bound by harmonic bonds, and repel the protein atoms at overlapping distances.

.

The *linker* is computationally defined as a series of beads in space, that is, a series of

points with associated radii. The beads are attached to the reactive atoms defined by the `linktype` keyword, and attached to each other, by distance-dependent harmonic potentials. We use the following functional form for the bond "energy" between beads:

$$U_{ij}^{\text{bond}} = \begin{cases} k_1 d^2 & \text{if } d \leq d_0 \\ k_1 d^2 + k_2 (d - d_0)^2 & \text{if } d > d_0 \end{cases} \tag{1}$$

where i and j are the indexes of consecutive beads of the linker, d is the distance between beads, d_0 is a distance parameter of the order of a covalent bond length, and k_1 and k_2 are constants, with k_2 greater than k_1.

The definition of this bond energy has the following justification: The distance d_0 is such that $(n_b+1) \times d_0$, where n_b is the number of beads, is the maximum length of the linker (there are $n_b + 1$ consecutive bonds between n_b beads plus the two reference atoms). Therefore, if the length is greater than $n_b + 1 \times d_0$, there must be a penalty for the distention of the linker, and this penalty might have physical meaning. For instance, if we think of the linked beads as a simplified representation of the actual linker molecule, the penalty for distension of the linker must be of the order of the penalty of the distention of the molecule. That is, with d_0 being a distance of the order of a covalent bond, k_1 must be of the order of the force constant of a covalent bond. Stretching the linker more than its maximum length will be similar to distending the covalent bonds of an actual molecule which has assumed its maximum length. Typically, this penalization is high. The functional form of the bonds could be, therefore, simply $k_2(d - d_0)^2$ for all distances. However, in this case, the linker length would always be close to the maximum linker length, and at best one would find linkers with zero penalty, without obtaining an actual knowledge of the minimum topological distance. Therefore, there must be a component of energy that promotes the shortening of the linker. This is guaranteed by, first, eliminating the k_2 penalty for pairwise distances smaller than d_0, and using in this case a softer penalty (with constant k_1) which is smaller as the distance between beads decreases. The constant k_1 must be soft, because large penalties tighten the linker and perturb the optimization by frequently causing undesirable overlaps between beads and model atoms (see below). Finally, the term $k_1 d^2$ persists at distances greater than d_0 to preserve the smoothness of the function. The calculation of this energy depends linearly on the number of atoms of the linker.

The linker beads must not overlap with the model atoms. Here, this restriction is incorporated into the objective function. The overlap energy between a linker bead and an atom is defined as

$$U_{ik}^{\text{overlap}} = \begin{cases} k_{\text{overlap}} (d_{\min} - d)^2 & \text{if } d \leq d_{\min} \\ 0 & \text{if } d > d_{\min} \end{cases} \tag{2}$$

where i is the index of the linker bead, k is the index of the structure atom, k_{overlap} is a constant, d is the current distance between the bead and the atom, and $d_{\min}$ is the distance

tolerance for overlaps. This overlap function is quadratic for distances shorter than the tolerance for overlap and null for distances greater than the overlap distance. Therefore, it is only affected by short-range interactions of the beads of the linker with structure atoms, and is exactly zero when no overlaps exist. Despite that it depends on the calculation of distances to all atoms of the protein, the fact that it is only very short-ranged allows the effective use of linked cell methods [1], with which the number of distances computed is small and proportional to number of atoms of the linker only. This is the same overlap function used in the packing program Packmol [3, 4] and is evaluated with the same efficient strategies reported in [4].

The complete energy function, is therefore,

$$U = U_{A,1}^{\text{bond}} + U_{n_b,B}^{\text{bond}} + \sum_{i=1}^{n_b-1} U_{i,i+1}^{\text{bond}} + \sum_{i=1}^{n_b} \sum_{j=1}^{N^*} U_{i,j}^{\text{overlap}} \tag{3}$$

where the first two terms are bond terms related to the attachment of the first atom of the linker with the first reactive atom (atom A), and to the attachment of the last atom of the linker to the second reactive atom (atom B). The third term is the sum over all consecutive atoms of the linker of the bond energy, and the last term is the sum over all atoms of the linker and the atoms of the structure of the volume exclusion term. The last sum is indicated as N^*, where N is the number of atoms of the structure, because the overlap function is not computed for the side-chain atoms of the reactive residues.

To obtain the minimum topological path between two reactive sites, we solve the optimization problem

$$\min U(\vec{x}_{\mathsf{linker}})$$

where U is a function of the position of the linker beads. Analytical derivatives are computed and a Conjugate-Gradient-Newton method of local optimization, as described in [2], is used. Multiple initial random conformations of the linker are used to obtain the linker of shortest length. In the current implementation, if the same best path is obtained three times, the global minimizer is assumed to be found.

To validate the present strategy, we have compared the topological paths obtained with TopoLink with the paths obtained with Xwalk [5]. Xwalk uses a completely different approach, consisting of an breadth-first search of paths defined by a sequence of points of three-dimensional grid defined to not overlap with the structure. Figure 2 displays the topological distances obtained by Xwalk as a function of the distances obtained by TopoLink, with default parameters for both methods, for crosslinks of at most 30Å between Lysine residues, on the surface of a barnase dimer (PDB id. 1BRS) which is provided as the test case for Xwalk. The methods coincide in 26 of 30 crosslinks found by Topolink. XWalk suggested one path that was not suggested by Topolink, but this path does not appear to physically justified as it crosses the protein core. The exact paths and lengths are dependent, on both methods,

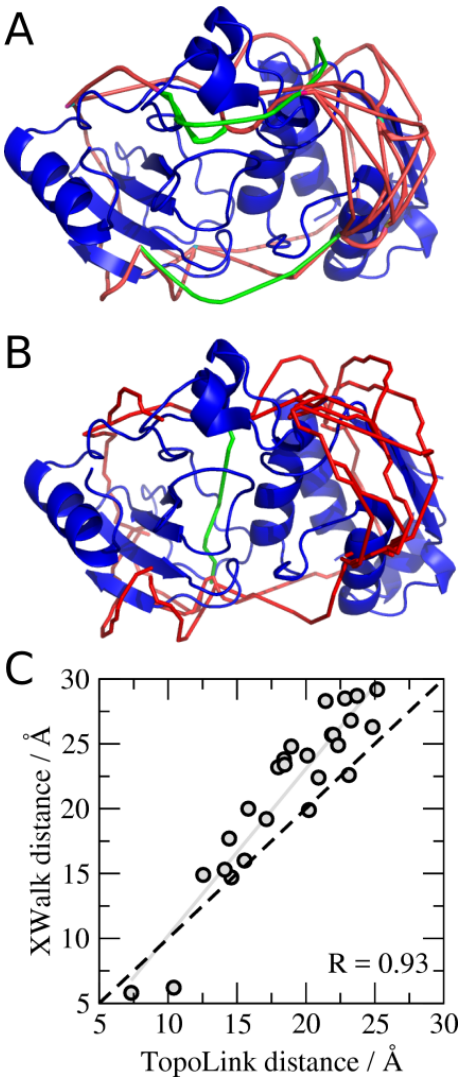

Figure 2: Topological distances obtained with (A) TopoLink [5] and (B) Xwalk for a barnase dimer (PDB ID. 1BRS). In this case, a linker of 30Å was considered for linking pairs of Lysine residues. 30 links were found by TopoLink, and 27 were found by Xwalk, from which 26 are similar to those found by TopoLink. The links which do not coincide are shown in green, and seem to be physically consistent in the TopoLink search, while the extra link found by XWalk crosses the protein core. (C) The paths found by TopoLink are most times shorter than those found by XWalk, particularly for longer links. These differences are dependent on methodological parameters, which were kept to default values in both cases.

on some input parameters that can in principle be adjusted by the user (excluded volume sizes and convergence criteria, for instance). We have carefully tuned the parameters in TopoLink such that in the large database comparisons we report the crosslinks found appeared to be physically consistent. With the final tuned parameters, TopoLink appears to be consistently more reliable than Xwalk to find accessible paths with physical accuracy. At the same time, the reasonable degree of coincidence of the methods shows that both strategies are reasonable enough to be used for model analysis and experimental design.

References

[1] Griebel, M., Knapek, S., Zumbush, G. (2007) Numerical Simulations in Molecular Dynamics; Springer, Berlin-Heidelberg.

[2] Birgin, E. G., Martínez, J. M. (2014) Solving Unconstrained Subproblems. *Practical Augmented Lagrangian Methods for Constrained Optimization*. 1st edn. SIAM, Philadelphia.

[3] Martínez, J. M., Martínez L. (2003) Packing optimization for automated generation of complex system's initial configurations for molecular dynamics and docking. Journal of Computational Chemistry. $J.\ Comp.\ Chem.$, **24**, 819-825.

[4] Martínez, L., Andrade, R., Birgin, E. G., Martínez, J. M. (2009) Packmol: A package for building initial configurations for molecular dynamics simulations. $J.\ Comp.\ Chem.$, **30**, 2157-2164.

[5] Kahraman, A., Malstöm, L., Aebersol, R. (2011) Xwalk: computing and visualizing distances in cross-linking experiments, $Bioinformatics$, **27**, 2163-2164