BioGeometry NEWS
February 2004
biogeometry.cs.duke.edu

Research

Simplified Model for Conformational Flexibility in Small Molecule Docking by Jeffrey J. Headd and Johannes Rudolph

Introduction: The ability to predict protein-protein interactions from the structures of individual proteins would aid greatly in understanding many biological questions. Though bound docking has been done with some success, its biological significance is limited since proteins do not dock as rigid objects. Thus, a methodology for unbound docking would be far more useful in studying biological systems. Towards developing a successful method for unbound docking, we have performed a model study by docking small molecules to a conformationally perturbed protein. Our aim is to understand conformational variables in proteins that may direct future approaches to unbound protein-protein docking.

Protein Preparation: For this study, H-Ras and its substate GTP were used as a model system (PDB ID: 1CLU). Starting from the crystallized conformation of the GTP-bound form of H-Ras, the protein was allowed to move in solution for 10 picoseconds. This was accomplished by placing H-Ras in a solvation sphere of water and running a ten picosecond molecular dynamics calculation with a femtosecond time step using the Tinker MD package. Positional printouts (PDB) were recorded every picosecond, which were then used in docking calculations. RMSD values for each pose are displayed in Table 1. RMSDs for the whole protein vary from 0.9 to 1.17 Å and for the active site residues vary from 0.77 to 1.18 Å. Visual inspection of these ten poses demonstrated that a wide variety of sidechain conformations were sampled in the MD simulation, providing a diverse test bed for simulating unbound docking (data not shown).

Docking: The ten poses collected from the MD simulation were used in docking calculations with the rigid crystal structure conformation of GTP. Of the ten, seven bound correctly, with three failures. Docking scores for each pose are listed in Table 1. The algorithm works by exhaustively searching the six-dimensional space of translations and rotations without knowledge of the binding site. The docking score is generated by counting proximal pairs of spheres (atoms) while not allowing more than five overlaps. The poses are ranked according to highest score (minimize bump while maximizing sphere proximity). A docking calculation is considered to be a success when the highest scoring position of the small molecule has an RMSD of less than 2.0 Å with respect to the position in the crystal structure. This simple scoring system allows for fast exploration of the complete set of possible docking positions in order to arrive at the correct solution. One docking calculation completes in an average of about 20 minutes, allowing this study to be completed in a few hours time.

Results: Visual inspection of the docked configurations using VMD revealed that three residues in H-Ras, Gly15, Asp30, and Lys117, were directly linked to the outcome of each docking calculation. Figure 1 shows GTP (red) in the binding pocket of H-Ras (silver) with the three residues of interest in green. These three residues form a coordinated triangle that surrounds the correctly bound GTP. This triangular pocket is depicted in Figure 2, with GTP (red) surrounded by line depictions of three different conformations of Gly15, Asp30, and Lys117. The green depictions are where these three residues were in the crystal structure. In cases where the docking calculation failed, these three residues were closer together, effectively pinching off the active site

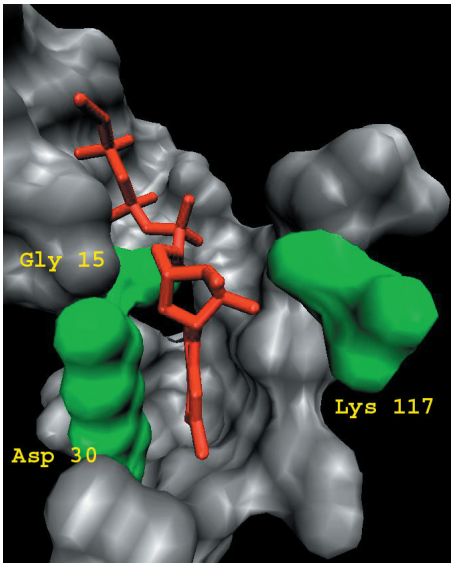

Figure 1

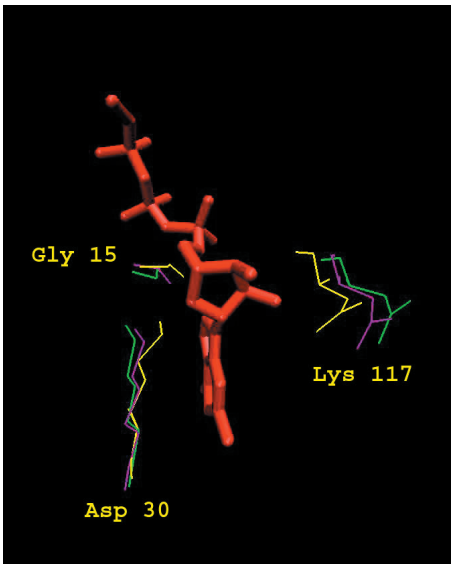

Figure 2

(see yellow depiction of the three side chains in Figure 2). In cases where the docking calculation succeeded, these three residues were either close to the crystal conformation (purple depiction in Figure 2) or further apart relative to each other, widening the binding pocket. Quantitatively, there is a correlation between the RMSD for these three residues and the success of the docking algorithm. RMSDs for these residues ranged from 0.64 to 1.29 Å, with two of the three failed conformations having values greater than 1.10 Å. The third failure, Pose 9, has an unexpectedly low RMSD of

0.96 Å. Upon further inspection, Pose 9 failed because a chance orientation of Glu63 created an alternate binding pocket that, in conjunction with a fairly pinched off true binding pocket, allowed for a false positive score at an alternate site. This site may be favored based on geometry alone, but is biologically not viable due to the close packing of similar charges between GTP and H-Ras (data not shown). Of the successful dockings, six had RMSD values for the three crucial residues less than 1.0 Å. Pose 10 is the exception, with an RMSD of 1.09 Å. Visual inspection of this pose demonstrated that the three key residues have moved in parallel to the length of the pocket, neither widening nor pinching off the pocket (data not shown). Thus, RMSD measurements of binding site residues are not necessarily predictive of docking.

Protein	Whole Protein RMSD	Active Site RMSD	Gly 15, Asp 30, and Lys 117	High Score	Docking Outcome
Pose 1	0.899	0.77	0.858	266	Success
Pose 2	0.966	0.948	0.69	282	Success
Pose 3	0.979	0.898	0.635	273	Success
Pose 4	1.053	1.059	0.796	268	Success
Pose 5	1.029	1.098	1.139	244	Failure
Pose 6	1.098	1.033	0.984	260	Success
Pose 7	1.098	1.086	0.902	247	Success
Pose 8	1.122	1.178	1.293	239	Failure
Pose 9	1.174	1.096	0.967	252	Failure
Pose 10	1.170	1.133	1.093	250	Success

Table 1
The table shows RMSD changes during molecular dynamics and their effect on rigid body docking attempts.

Discussion: As seen in Figure 2, the variation in structure that determines success vs. failure for small molecule docking is not all that substantial. Thus, the algorithm is clearly sensitive to even minor changes in the protein structural orientation. With the observed failures resulting from either an obscuring of the active site or the creation of new false positive sites, it is clear that there is no simple solution to unbound docking. Still, the successful docking of seven different conformations demonstrates that a geometry based docking methodology may be useful for predicting the docking sites of unbound conformations. In regard to protein docking, it is likely that fewer false positive sites will occur given the increased size and complexity of the interface of protein docking pairs compared to small molecule sites.

Future Work: To address the failures of the docking methodology by shape alone, it may be prudent to consider a simplified energy function to reward correct placement of charges and to select against obvious charge clashes. Though more computationally intensive, a simplified function applied in conjunction with the currently implemented bump evaluation will not largely change the complexity of the algorithm, and may lead to higher success in small molecule docking, including solving the unbound docking problem for both small molecules and protein-protein interactions. An energy function addition may also help in cases of pairs with more substantial conformational shifts in docking, where correct energetics can help direct the binding pair into the correct conformation.

Research News

Computing High-Stringency COGs Using Turan Graphs by Jack Snoeyink

One of the outcomes of the case studies in Jack Snoeyink's "Applied Optimization in Computational Biology" seminar was a new algorithm by Craig Falls (UNC CS) for computing COGs (clusters of orthologous groups), a problem introduced by Bradford Powell (UNC Genetics). All three have co-authored a submission to ISBM. COGs were described by Tatusov et.. al., 1997, as a technique for searching several complete genomes to identify groups of genes that are likely orthologs---genes in the different organisms that are related by evolution. The online COG database is a useful, and growing, resource for the study of the proteins encoded by these genes.

The original clustering procedure was based upon finding overlapping sets of three mutually best-fit genes (BeTs), and was run on 7 genomes from 5 *clades*--families of related organisms. As more genomes are added, however, it is natural to ask for overlapping sets of $m>3$ mutually best-fit genes. To find highly-conserved orthologs in the herpesvirus, Montague and Hutchinson (2000) defined COGs of "higher stringency" as sets of overlapping m-cliques.

The number of m-cliques grows exponentially as m grows, making straightforward computation of COGs exponentially more difficult as the number of genomes grows. The number of groups, however, remains quite tractable. We observe that the chief obstruction in practice is related to Turan graphs from extremal graph theory. By recognizing maximal Turan-type subgraphs, we are able to compute COGs of stringency up to the number of organisms in practical examples. Bradford's original data set required 3 hours and 58 minutes by clique enumeration, but only 11 seconds with the algorithm we have developed.

The BioGeometry project is funded by the National Science Foundation under grant CCR-00-86013. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.