

Supporting Information

Diverse Fragment Clustering and Water Exclusion Identify Protein Hot Spots

John L. Kulp III,[†] John L. Kulp Jr.,[‡] David L. Pompliano,[‡] and Frank Guarnieri^{‡□§*}

[†] *Chemistry Division, Naval Research Laboratory, Washington, DC 20375-5342.*

[‡] *BioLeap, Inc., 238 West Delaware Ave., Pennington, NJ 08534.*

[□] *Department of Physiology and Biophysics, Virginia Commonwealth University, Richmond, VA 23298.*

[§] *Department of Biomedical Engineering, Boston University, Boston, MA 02218.*

E-mail: frankguarnieri@yahoo.com

METHODS

Simulated Annealing of Chemical Potential.

The Monte Carlo method used to compute the fragment distributions employed in this study follows the Grand Canonical Monte Carlo (GC/MC) scheme formulated by Adams.¹ The protein structure is placed in a periodic simulation cell large enough to accommodate three layers of solvent. In concept, a chemical potential is imposed between an ideal gas reservoir of fragments and the simulation cell. The system will adjust to this chemical potential until equilibrium is reached where the average number of fragments becomes stable. The details of this method are described elsewhere.^{2,3}

Guarnieri⁴ developed a method, followed here, where the chemical potential is slowly “annealed” from high positive values to low negative values in a series of steps. For each change in chemical potential, GC/MC is run until equilibrium is reached. This method has several benefits: (1) starting at a high positive value of chemical potential, the simulation cell becomes densely packed, forcing fragments into all interstices of the protein and more efficiently sampling multi-fragment configurations—particularly critical for water modeling; (2) as the chemical potential is annealed to lower values, the system goes through a phase transition where the region of the simulation cell away from the protein surface is voided—the chemical potential at which this happens distinguishes strongly bound fragments from transient weakly interacting fragments; and (3) as the chemical potential is lowered beyond the phase transition, fragment binding sites become isolated. In this latter regime, fragments have a negligible interaction with other fragments, and the chemical potential—the average free energy per molecule—can then be used to characterize the free energy of the fragment binding in each individual site. Importantly, a fragment binding metric is achieved that incorporates the entropic contribution to the binding free energy, in contrast to methods that only encompass the interaction energy. Our experience has shown that not including configurational entropy is a key source of inaccuracy in the relative

ranking of binding sites by affinity. Clark et al.⁵ have shown that the lowest chemical potential at which a fragment occupies a given site is predictive of fragment binding affinity as determined by isothermal calorimetry.

There are numerous practical issues in implementing the basic GC/MC method that have to do with making the sampling process efficient enough to be useful. Mezei⁶⁻⁸ demonstrated that the fragment insertion success rate at high fragment densities is very low, and can be made far more efficient by tracking cavities available for inserting fragments and biasing the choice of insertion locations to those cavities. However, this method is moderately complex to implement and is sensitive to grid assignments, requiring grid shifting methods. Our code used in this study uses an alternative strategy; it achieves insertion efficiency by dramatically lowering the computational cost of failed insertions by truncating the expensive energy calculation using numeric overflow detection. That is, the energy calculation is aborted as soon as a steric clash is detected. Another efficiency issue is how many Monte Carlo steps should be used at each chemical potential value to achieve equilibration. Typically, a fixed number of steps, between three and ten million, are taken, but this is never optimal for any particular fragment. We take steps until several equilibration criteria are met (e.g. samples per fragment, rate of change of fragment number, etc.), and these criteria can be adjusted as a function of the imposed chemical potential. This results in an optimal run length for each type of fragment, in the range of half a million to five million steps. Further, the number of steps is dynamically adjusted as the number of fragments in the system decreases as the chemical potential is lowered through the simulation, thus achieving a consistent amount of sampling per fragment. Another efficiency technique commonly used is the application of a cutoff distance to limit the number of fragments or protein residues included in fragment energy calculations. The typical method of using the same cutoff distance for all types of fragments has the problem of a cutoff being too short (leaving out important energy contributions) or too long (including insignificant contributions at added cost) for a given fragment type. We use a rigorous energy-based cutoff, which accommodates a wide variety of fragments by pre-computing an optimal cutoff distance for each fragment type based on the maximum interaction energy between any pair of fragment types or residues. The cutoff energy contribution used in this study was 0.1 kT. Different cutoff distances are used for atom-atom models and less expensive multipole models for fragments or residues.

The annealing schedule used in these studies specifies a parameter $B = \mu_{\text{ex}}/kT + \ln \langle N \rangle$ where $\langle N \rangle$ is the average number of fragments in the system and μ_{ex} is the excess chemical potential associated with protein interactions. The schedule starts at B of +100.0, decrements B in an exponentially collapsing sequence of values down to zero, then an exponentially expanding sequence from zero down to -100, or until there are no fragments left in the system. The simulation is then restarted from a checkpoint file for the value of B just prior to the phase transition. B is then decremented by 0.5 until no fragments have greater than 50% occupancy remain. We define a cluster as a set of fragments of different types that are located within 2.5 Å of each other—the maximum allowed distance from the center of mass of one fragment to another in a cluster. The value of B at which occupancy drops below 50% is used to calculate an excess chemical potential (free energy per molecule) for ranking sites by affinity. These values are reported in Table S1 for the fragments used in this study.

Fragment Clustering and Water Exclusion.

The procedure for fragment clustering involves a number of different parameters: (1) the number of different fragment types used, Table S2; (2) the lowest chemical potential energy of each fragment type, Table S1; (3) the highest chemical potential energy of waters used to exclude fragments, Table S1; (4) a radius for fragment cluster assignment, 2.5 Å; and (5) a radius for water exclusion, 1 Å. We initially load fragment data into a fragment data browsing tool—the BioLeap Fragment-based Design tool (BFD)—for a wide range of free energies. Using the visualization controls, an excess chemical potential level is selected that saturates the protein structure with all eight organic fragments. We then lower the excess chemical potential individually for each fragment type—recapitulating the simulations as fragments leave—until each organic fragment occupies multiple sites around the protein surface. Next, we retain fragments in sites where clusters form; a cluster is defined as a set of fragment types that are located within 2.5 Å of each other—the maximum allowed distance from the center of mass of one fragment to another in a cluster. Next, μ_{ex} for waters is selected below the value where the phase transition occurs—the bulk volume of the simulation box is voided and only surface-bound waters remain. These are tightly-bound waters. Fragments are eliminated that have a heavy atom that is within 1 Å of the oxygen of a bound water molecule. The μ_{ex} of water is lowered until at least some clusters survive elimination. Finally, there will often be more than one cluster site remaining. Additional fragment types are added to increase the chemical diversity (typically 5-6 total) and clusters recalculated, and eliminated by water exclusion until only one site remains. The chemical diversity parameter characterizes the number of fragment types required to identify the hot spot. For example we could have identified the HEWL hot spot with three to five fragments and did not need all eight, Table S2. This procedure is agnostic to the final outcome and is systematically applied to minimize the number of clusters. Occasionally there will be a minimum of 2-3 sites identified, as with sites at interfaces between homo-dimers or -trimers, or high affinity allosteric sites. Sites with the highest chemical diversity and affinity, not excluded by tightly-bound waters, are consistently identified by this method. A high degree of correlation has been found between these sites and active sites on enzymes, protein-protein interface hot spots, and functional sites on receptors. This is one trajectory of our current work.

Rendering.

All figures were rendered with the program PyMOL.⁹

Table S1. Excess chemical potential (kcal/mol) for each fragment and water used in cluster analysis.


















pdb									
ID	ACM	ANT	BEN	FUR	IBU	MNO	PYD	THL	Water
1IR8	-27.29	-23.11	-12.63	-13.84	-8.60	-14.57	-19.03	-12.75	-16.00
1IR9	-27.30	-17.87	-11.47	-12.10	-8.61	-19.82	-16.71	-18.58	-19.50
1LSY	-25.53	-19.60	-7.96	-9.17	-6.85	-16.31	-12.61	-12.74	-11.32
1LSZ	-24.95	-19.02	-11.45	-12.08	-2.76	-14.56	-17.27	-12.74	-11.90
1XEI	-29.03	-24.85	-14.36	-12.67	-6.85	-19.80	-20.19	-15.07	-15.54
1XEJ	-25.54	-26.60	-14.37	-16.16	-5.10	-23.88	-21.35	-16.82	-20.65
1XEK	-21.46	-27.18	-10.87	-14.42		-22.14	-19.03	-17.40	-21.23
2LYO	-28.45	-22.53	-10.29	-15.59	-7.44	-18.65	-16.12	-13.33	-15.41

Table S2. All fragments clustered in high energy binding site for all structures except iso-butane in 1XEK; although iso-butane bound in close proximity to binding site, it was eliminated due to water exclusion.

pdb	Chemical								
ID	diversity	ACM	ANT	BEN	FUR	IBU	MNO	PYD	THL
1IR8	4	✓	✓	✓	✓	✓	✓	✓	✓
1IR9	5	✓	✓	✓	✓	✓	✓	✓	✓
1LSY	5	✓	✓	✓	✓	✓	✓	✓	✓
1LSZ	5	✓	✓	✓	✓	✓	✓	✓	✓
1XEI	4*	✓	✓	✓	✓	✓	✓	✓	✓
1XEJ	3	✓	✓	✓	✓	✓	✓	✓	✓
1XEK	4	✓	✓	✓	✓		✓	✓	✓
2LYO	4	✓	✓	✓	✓	✓	✓	✓	✓

*Another site shows clustering; other cluster has water between cluster and protein but eliminated with chemical diversity set to 6.

- (1) Adams, D. J. *Mol. Phys.* **1975**, 29, 307.
- (2) Frenkel, D.; Smit, B. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications*; Academic Press: New York; Vol. 1.
- (3) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1989.
- (4) Guarnieri, F.; Mezei, M. *J. Am. Chem. Soc.* **1996**, 118, 8493.
- (5) Clark, M.; Guarnieri, F.; Shkurko, I.; Wiseman, J. *J. Chem. Inf. Model* **2006**, 46, 231.
- (6) Jedlovsky, P.; Mezei, M. *J. Am. Chem. Soc.* **2000**, 122, 5125.
- (7) Jedlovsky, P.; Mezei, M. *J. Chem. Phys.* **1999**, 111, 10770.
- (8) Mezei, M. *Mol. Phys.* **1980**, 40, 901.
- (9) The PyMOL Molecular Graphics System, Version 1.4, Schrödinger, LLC.