Efficient Monte Carlo Algorithm for Simulating Reversible Aggregation of Multisite Particles

Qiang Chang and Jin Yang*
Chinese Academy of Sciences — Max Plank Society Partner Institute for Computational Biology
Shanghai Institutes for Biological Sciences, Shanghai 200031, China

We present an efficient and exact Monte Carlo algorithm to simulate reversible aggregation of particles with dedicated binding sites. This method introduces a novel data structure of dynamic bond tree to record clusters and sequences of bond formations. The algorithm achieves a constant time cost for processing cluster association and a cost between $O(\log M)$ and $O(M)$ for processing bond dissociation in clusters with M bonds. we apply the method to simulate a trivalent ligand and a bivalent receptor clustering system and obtain an average scaling of $O(M^{0.45})$ for processing bond dissociation in acyclic aggregation, compared to a linear scaling with the cluster size in standard methods.

PACS numbers: 05.10.Ln, 87.16.dr, 87.10.Rt

Reversible aggregation (or self-assembly) of particles with multiple interactive sites is of fundamental importance to diverse processes in physical and living systems including coagulation of colloidal particles, protein aggregation [1], synthesis of supramolecules in polymer science [2], and self-assembly of patchy particles such as nanoparticles [3, 4] and synthetic biomolecules [5] in material sciences [6, 7]. Reversible aggregation was traditionally studied using the generalized Smoluchowski equation [8, 9] that requires one to develop kernel functions for cluster aggregation and fragmentation to obtain the cluster size distribution. Proper kernel functions can be analytically characterized often under restrictive assumptions for particle interactions. For acyclic aggregation of multisite particles, the combination of Wertheim's thermodynamic perturbation theory [10] and Flory-Stockmayer theory [11] can predict equilibrium properties for simple systems. No analytical theory exists to treat cyclic aggregation in general. Therefore, Monte Carlo simulations are indispensable to provide new insights into the kinetics of aggregation processes with arbitrary complexity.

Reversible aggregation involves two principal types of events, bond formation and breaking. In a site-based algorithm, clusters are stored as graphs representing the connectivity between particle sites (see Fig. 1). Upon each event, an importance sampling is applied to determine a site pair to form a bond, or to determine a bond to break. To resolve information such as composition and topology of a cluster, graph traversals by depth-first (or breadth-first) search are routinely applied. For irreversible aggregation, a highly efficient algorithm using weighted union-find with path compression [12] can be applied to identify cluster membership of binding sites and amalgamate two clusters in near constant time [13], as demonstrated in simulating percolation models [14]. Unfortunately, this strategy cannot be readily adopted to simulate reversible aggregation because bond dissociation requires time-consuming reorganization of tree-based data structures involved in the algorithm. Instead, one can label each individual site on cluster connectivity graphs to track its cluster membership. Site relabeling is thus required to process each event. To process a bond formation between two clusters, because cluster sizes are known by simple bookkeeping, one can always relabel sites in the smaller cluster using the label assigned to the larger cluster to minimize the cost. However, this weighted relabeling does not work for cluster dissociation because the sizes of the two subclusters are not known *a priori*. Cluster relabeling can only be carried out by a graph traversal on one arbitrary subcluster. The average time complexity of a cluster traversal is $O(N + M)$, scaled by the total number of particles N and bonds M in the cluster, which becomes prohibitive for simulating high density systems with large connectivity graphs.

Here, we present an efficient Monte Carlo algorithm that amalgamates two clusters in $O(1)$ time and splits a cluster in time between $O(\log M)$ and $O(M)$. Unlike site-based methods, the main idea behind our algorithm is based on the observation that explicit cluster graphs are usually not required in a simulation. We use a more efficient data structure, namely *dynamic bond tree* (DBT), to track bonds and clusters without updating the actual connections between particle sites. The algorithm is numerically exact in generating observable quantities such as the cluster size distribution, average cluster size and the number of clusters. The algorithm is straightforwardly applicable to simulate aggregations that allow formation of cyclic clusters. If topologies of clusters are of interest, connections among sites can be recorded in parallel during a simulation, or alternatively, ensembles of cluster topologies can be mapped out stochastically from the corresponding DBTs by postprocessing.

We consider a system with a homogeneous population of particles, each of which is decorated with several symmetric surface patches (binding sites). The algorithm can be extended to a system with a heterogeneous population of particles decorated with non-identical sites which can bind to complementary sites on other particles. We assume that a single site can only sustain at most one bond. Each cluster of particles is represented as a DBT and identified by the root node. A leaf node in a DBT represents a single particle in the cluster, whereas a non-leaf node records a site-site bond. Each non-leaf node has either one or two child nodes. A node with two children indicates that the bond was formed by an association between a pair of sites that reside on two previously separate clusters represented by the two child nodes, whereas

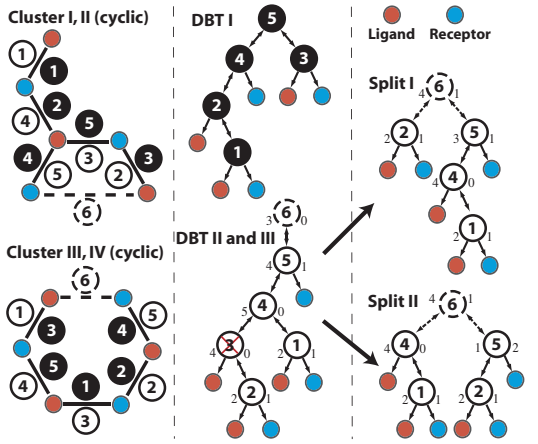

FIG. 1. Aggregation of trivalent ligands and bivalent receptors. Left panel: cluster connectivity graphs. Clusters I (branched, without bond 6), III (linear, without bond 6), II and IV (cyclic, with bond 6) has same number of ligands and receptors but different topologies (not drawn to reflect the actual structure of a cluster). Middle panel: clusters I and III represented by either DBT I or II depending on their sequences of bond formation. Sequences of bond formation are numbered (with filled and unfilled circles) in cluster graphs and in corresponding DBTs. The DBT III appends bond 6 to DBT II corresponding to cluster II or IV. Right panel: cluster dissociation (with or without bond 6) after breaking bond 3 (in DBT II or III). The number of free ligand sites (left) and the number of free receptor sites are shown as weights for subclusters represented by individual DBT nodes.

a node with a single child indicates that the bond was formed by an association between a pair of sites that reside on a same cluster represented by the child node. Therefore, a cluster is *cyclic* if and only if the corresponding DBT contains at least one non-leaf node that has a single child. Otherwise, a cluster is *acyclic*. Below, we use acyclic aggregation as an example to describe how to process bond formation and breaking using DBT structures. Processing cyclic aggregation only requires slight adaptation.

To process a bond formation, two clusters are first sampled according to their joint probability of contributing binding sites. The probability for a cluster c to contribute a binding site may be related to its number of free sites s_c, by a system-specific function $g(s_c)$, which is assigned to each cluster as a weight. [For example, consider that a cluster of a spherical volume has s_c free binding sites. Due to the effect of steric hindrance, one may assume that only free sites near the cluster surface can form a bond with a site on another cluster. In this model, assuming free sites are homogeneously distributed within the cluster volume and on the surface, one can show that $g(s_c) \sim s_c^{2/3}$ is a good approximation.] After two binding clusters are chosen, a new node z is then created as a root node of the DBT that will store the resulting cluster. The root nodes, x and y, of the DBTs of the two binding clusters become two children nodes (subclusters) of z. A weight $g(s_z)$ is assigned to z, where $s_z = s_x + s_y - 2$ and the adjustment by -2 is due to the fact that two sites are consumed to form bond z, each from one subcluster. Therefore, constructing a DBT manifests the hierarchical nature of cluster aggregation. Unlike the standard method, this procedure of merging two clusters requires neither cluster membership checking of trial binding sites nor systematical site relabeling, and thus merely takes $O(1)$ time. We point out that locating two clusters with matching sites demands searching over the entire array of clusters, which has a cost that scales linearly with the number of clusters. We will show in an example below that this cost is in most cases modest if not ignorable.

To process a bond dissociation, one first samples a bond according to its probability to dissociate. The selected bond locates to a non-leaf node x in a DBT identified by its root node z. The removal of node x will split z into two separate DBTs. If x happens to be z, its two child nodes, l and r, simply become root nodes of the two separate DBTs. Otherwise, the final two DBTs are probabilistically determined. Since the subcluster identified by x contributes a site to form the bond at the parent node of x, p, with the other child of p, after x dissociates we need to decide which of the two subclusters, represented by nodes l and r, provides the site and thus will connect to node p as a child node. We may assume that the probability of choosing either l or r is proportional to a function of the number of free sites contained in the subcluster. For instance, the number of free sites in subcluster l is $s_l - 1$. The probability of choosing l to connect to p is $g(s_l-1)/(g(s_l-1)+g(s_r-1))$. If l is selected, as a result r dissociates from cluster p. We therefore need to adjust the number of free sites in p: $s_p \leftarrow s_p - (s_r - 1)$ and recalculate the weight of p. Now we want to further decide which of r and the updated p connects to the parent node of p, and so on. This procedure iterates up to the root node z and then obtains two separate DBTs. Obviously, the total number of iterations equals the height of the DBT from node x to the root node. This procedure to break a bond in a cluster is very efficient,

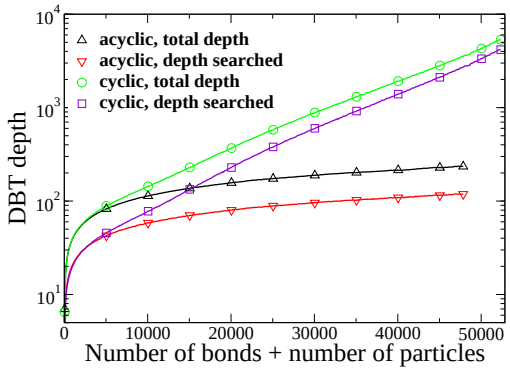

FIG. 2. Size of a cluster ($N+M$) vs. the depth of the corresponding DBT (total or searched), for acyclic or cyclic aggregations. Simulations use 15000 receptors and 10000 ligands, with fixed rate constants $k_+ = 6.67 \times 10^7 s^{-1}$ and $k_{++} = 100k_+$. The dissociation rate constant $k_{\rm off}$ was varied for simulations to generate clusters of different sizes. Cluster sizes and the DBT depths were obtained by averaging within bins at an equal size of 100.

which has a cost between $O(\log M)$ when a DBT is well balanced and $O(M)$ when a DBT forms a linear cascade due to sequential attachment of single particles. The latter scenario is pathological and unlikely to persist because dynamic bond association and dissociation prevent formation of stable linear DBTs for any cluster such that on average DBTs are more or less flattened during a simulation.

As we will show, this algorithm provides a substantial speedup for processing bond dissociations in high density clusters. For cyclic aggregation, the procedure is largely the same as described above, except that whenever an intracluster site pair forms a bond, a node is created with only one subtree that corresponds to the same cluster contributing both binding sites (see Fig. 1). Figure 1 illustrates how to process cluster aggregation using DBTs for an example system of trivalent ligands binding to bivalent receptors (TLBR). Especially, we note that an equivalent class of DBTs exists for each cluster with a distinct connectivity, and vice versa. For instance, Fig. 1 shows that cluster I may be represented as DBT I or II depending on the sequence of bond formation. The stochasticity in breaking a bond in a cluster can also result in diverse fragmentation of the cluster.

To demonstrate our algorithm, we specialize to the TLBR system that is representative to aggregation of a mixture of heterogeneous particles with multiple complementary binding sites. In this system, extracellular ligands binding to cell-surface receptors and subsequent receptor crosslinking by receptor-bound ligands on the cell surface can induce receptor aggregation. We consider the system well mixed and apply the law of mass action to account for the rates of bond association and dissociation. Here, we simply assume that the probability for a cluster to contribute a receptor (ligand) site is proportional to the number of free receptor (ligand) sites in the cluster, i.e., $g(s_c) \equiv s_c$. A bond can be formed only between a ligand site and a receptor site. Each bond has an equal probability to break. The system involves three rate processes: (1) free ligands precipitating to bind cell surface receptors with a rate constant k_+, (2) receptor crosslinking by ligands already bound to receptors with a rate constant k_{++} and (3) ligand-receptor bond dissociation with a rate constant $k_{\rm off}$. At the start of a simulation, all ligand and receptor sites are free with no bond formed. For each iteration, one first determines the waiting time for the next event, then selects one process that fires the next event and finally updates the configuration of the system [15].

Figure 2 shows that the DBT depth in acyclic aggregation has a very slow growth with the increase in the cluster size, which fit to a monomial function of $M^{0.45}$. Cyclic aggregation exhibits a steeper growth of the DBT depth against the cluster size because intracluster bonds increase the DBT depth on top of an acyclic aggregate with the same number of receptors and ligands. The depth of the deepest DBTs on average is only one-tenth of the cluster size (about 4,000 vs. 50,000). For cyclic aggregation, we assume each site pairs has an equal probability to form a bond. Note that this assumption overestimates the probability of intracluster bond formation because geometric constraints may prohibit interactions between certain intracluster site pairs [16]. One can use a parameter $\phi \in [0, 1]$ to characterize the average probability of a given pair of intracluster sites to form bond. Upon each association event, a trial intracluster ligand-receptor site pair is accepted to form a bond with a probability ϕ. Here by setting $\phi = 1$, we intend to present a worst-case scenario for cyclic aggregations in terms of the DBT depth for clusters and expect that the performance of simulating any cyclic aggregation model of the TLBR system will lie between this extreme model and the acyclic aggregation ($\phi = 0$).

Simulation results verified that our algorithm is statistically identical to the method using graph traversals in obtaining the average cluster size and the number of clusters (see Fig. 3(a)). To simulate a system with high density clusters by the acyclic

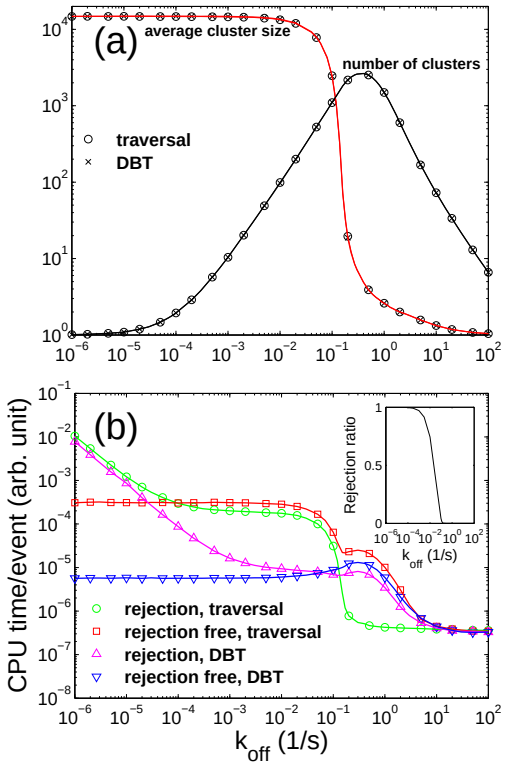

FIG. 3. (a) Average cluster size and the number of clusters, by methods using DBTs (cross) or graph traversals (circle). The cluster size is measured by the number of receptors in a cluster (not including free receptors). The average cluster size is given by: $\sum_{n=1}^{N_R} n^2 x_n/(N_R - F_R)$, where x_n is the number of clusters of size n, N_R is the total number of receptors and F_R is the number of free receptors. The results were obtained by averaging 5000 samples (each sample was separated by 100 events) at the equilibrium. (b) Performance of four schemes for simulating acyclic aggregation of the TLBR system: rejection or rejection-free sampling with or without employing DBTs. Inset: rejection ratio [the ratio of the number of effective events to the number of all events] in different phase regimes. The mean CPU time per event was obtained by averaging after a system equilibrated. Parameters are identical to the ones indicated in Fig. 2. All simulations were run on a same platform.

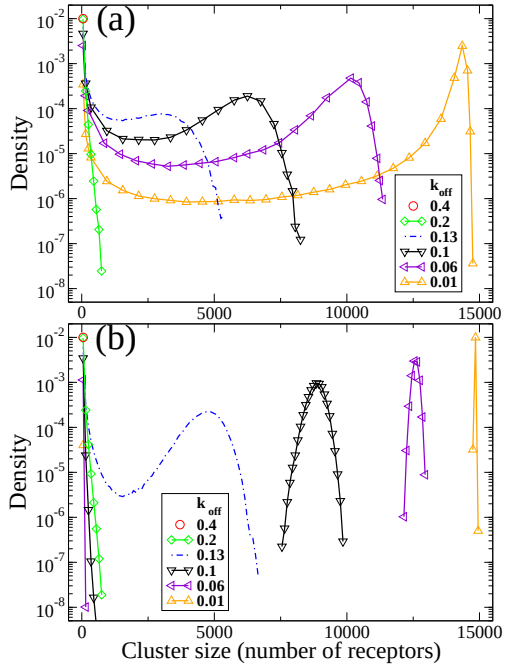

FIG. 4. Cluster size distributions of (a) acyclic and (b) cyclic aggregations under varied $k_{\rm off}$. Other parameters used in simulations are identical to the ones indicated in Fig. 2. Density is calculated as the probability to find a receptor in a cluster of a given size and averaged within bins of size 100. The system was sampled $100,000$ times after the system equilibrated. Each data point was separated by 100 events to eliminate the correlation in time.

aggregation, a rejection-free sampling [17] of binding sites is often required to further overcome the bottleneck caused by high rejection ratio in the rejection sampling that excludes intracluster site pairs from binding. Fig. 3(b) shows the performance comparison between different methods that use graph traversals or DBTs with or without rejection-free sampling. The combination of DBTs and the rejection-free sampling has a superior performance over other approaches, especially when rejected samples become dominant in the rejection sampling (the rejection ratio > 0.9). Except only for the method using rejection sampling with graph traversals, a hump (about $k_{\rm off} = 0.4s^{-1}$) in each curve in Fig. 3(b) reflected a small performance penalty due to sampling over a maximal number of clusters for binding clusters near the phase transition boundary. Such effect disappears for high density systems, because the number of clusters drops drastically when the system is at high-density regime as the average cluster size reaches the maximum (see Fig. 3(a)).

The equilibrium model by Goldstein and Perelson [18] and recent Monte Carlo simulations [19] showed that the branched (acyclic) TLBR system exhibits phase separation of sol and sol-gel phases. Here, we use our algorithm to explore the effect of cyclic aggregation on the cluster size distribution. Fig. 4 shows that below the phase transition ($k_{\rm off} > 0.01s^{-1}$) the equilibrium cluster size distributions of cyclic aggregation (see Fig. 4(b)) are more segregated with narrowed right peaks that are shifted toward higher cluster sizes, indicating an earlier onset of phase separation at a higher $k_{\rm off}$ (i.e., lower ligand-receptor affinity), compared to that in the acyclic aggregation (see Fig. 4(a)).

In conclusion, we presented an efficient kinetic Monte Carlo algorithm for simulating reversible aggregations of multisite particles, especially for systems with a large number of particles that nucleate into high density clusters. The algorithm records clusters and processes bond formation and breaking using dynamic bond trees, which avoids costly operations on connectivity graphs. As a result, the substantial gain in computation enables fast simulation of aggregation involving large number of particles. The algorithm is quite general and provides a fast mean to evaluate aggregation of patchy particles as well as basic physical models such as reversible site or bond percolation under various conditions.

We thank William Hlavacek and John Pearson for helpful discussions. This work was supported by National Science Foundation of China through grant 30870477 (JY).

* jinyang2004@gmail.com

[1] G. B. Fields, D. O. V. Alonso, D. Stigter, and K. A. Dill, J. Phys. Chem., **96**, 3974 (1992).
[2] P. Cordier, F. Tournilhac, C. Soulié-Ziakovic, and L. Leibler, Nature, **451**, 977 (2008).
[3] C. A. Mirkin, R. L. Letsinger, R. C. Mucic, and J. J. Storhoff, Nature, **382**, 607 (1996).
[4] T. M. Hermans, M. A. C. Broeren, N. Gomopoulos, P. Van Der Schoot, M. H. P. Van Genderen, N. A. J. M. Sommerdijk, G. Fytas, and E. W. Meijer, Nature Nanotechnol., **4**, 721 (2009).
[5] B. Bilgicer, D. T. Moustakas, and G. M. Whitesides, J. Am. Chem. Soc., **129**, 3722 (2007).
[6] S. C. Glotzer and M. J. Solomon, Nature Mater., **6**, 557 (2007).
[7] A. B. Pawar and I. Kretzschmar, Macromol. Rapid Commun., **31**, 150 (2010).
[8] F. Family, P. Meakin, and J. M. Deutch, Phys. Rev. Lett., **57**, 727 (1986).
[9] G. Odriozola, A. Schmitt, A. Moncho-Jordá, J. Callejas-Fernández, R. Martínez-García, R. Leone, and R. Hidalgo-Álvarez, Phys. Rev. E, **65**, 31405 (2002).
[10] M. S. Wertheim, J. Stat. Phys., **35**, 19 (1984); **35**, 35 (1984); **42**, 459 (1986); **42**, 477 (1986).
[11] P. J. Flory, *Principles of polymer chemistry* (Cornell University Press, 1953).
[12] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. (The MIT Press, 2001).
[13] R. E. Tarjan, ACM, **22**, 215 (1975).
[14] M. E. J. Newman and R. M. Ziff, Phys. Rev. Lett., **85**, 4104 (2000).
[15] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, J. Comput. Phys., **17**, 10 (1975).
[16] M. I. Monine, R. G. Posner, P. B. Savage, J. R. Faeder, and W. S. Hlavacek, Biophys. J., **98**, 48 (2010).
[17] J. Yang and W. S. Hlavacek, Arxiv preprint: 0812.4619 (2008).
[18] B. Goldstein and A. S. Perelson, Biophys. J., **45**, 1109 (1984).
[19] J. Yang, M. I. Monine, J. R. Faeder, and W. S. Hlavacek, Phys. Rev. E, **78**, 31910 (2008).