Introduction to rare events and collective variables

Giovanni Bussi

May 22, 2017

1 Motivation

Molecular dynamics (MD) simulations can be used to produce trajectories for systems composed of hundreds to millions of atoms. However, it is very difficult to analyze all these coordinates without simplifying them by means of some form of dimensional reduction. The variables obtained from a dimensional reduction are usually complex collective functions of the microscopic coordinates, and are thus called *collective variables*.

Collective variables (CVs) are however not used just to *analyze* simulations, but can also be used to *bias* simulation. This is particularly interesting in the case of rare event sampling. The timestep for MD is typically in between 10^{-13} and 10^{-12} seconds. However, many processes happen on timescales that are many orders of magnitude larger (e.g. milliseconds or seconds). Sometime these long time scales can be rationalized in terms of the slow dynamics of a limited number of CVs subject to *free energy barriers*. In these cases, knowledge of these CVs can be exploited to accelerate dynamics.

In the next sections we will learn more about what is a CV and how it can be used to *analyze* and *bias* a molecular dynamics simulation.

2 Collective variables

Collective variables (CV) provide a coarse graining of the coordinates of a system. Let's consider some example:

- A nucleoside, where isomerization around the glycosidic bond can be described by the torsional angle χ.
- A protein that is folding, where the progression along the folding trajectory might be described using the number of native contacts.
- A ion translocating across a membrane, where the translocation might be described using the z projection of its Cartesian coordinates, or better of their distance from the center of the membrane.

- The transfer of a proton between two atoms, that can be described by the distances between the hydrogen and the two atoms.
- The association of two ions in solution, that can be described by the distance between the two ions and perhaps the number of water molecules interacting with each ion.

In these cases one might want to describe a process that in reality involves a large number of atoms looking only at a few degrees of freedom of the system (e.g. a torsional angle or a distance between two atoms). Notice that the state of the system is fully determined by the microscopic coordinates and velocities q and v. A CV is just an arbitrary function of q and, optionally (but almost never!), of v:

$$s = s(q, v)$$

In the following we will use everywhere s as if a single CV was defined, but a process might require multiple variables to be described. All the equations are easily generalized to the case where s is actually a vector.

2.1 Distances, angles, torsions, and centers

Common CVs are distances, angles, and dihedral angles between atoms or groups of atoms. Angles are not necessarily between atoms that are chemically bound. CVs might depend on the position of the center of a molecule rather than that of a specific atom. For instance, if you want to analize a small molecule translocating across a membrane, you might use the z projection of its center.

2.2 Positions

Atomic positions might be directly used as CVs. Notice however that MD simulations are virtually always subject to a translationally invariant potential energy function. Thus, the position of an atom is usually not capable to describe any process. Think about a ion translocating across a membrane. For a given z component of the ion coordinates, the ion might be inside the membrane. However, if the membrane moves, for the same component the ion might be outside the membrane. It is thus much better to use the z component of the distance between the coordinates of the ion and the center of the membrane.

2.3 Coordination numbers

Notice that collective variables can in principle be discontinuous functions of positions. This is fine if you only want to analyze an MD trajectory. However, if you want to bias a trajectory you will need collective variables that are continuous functions of positions. This is why variables such as the "number of neighbors of atom i" are written as

$$s_i = \sum_j f(d_{ij})$$

Here d_{ij} is the distance between atom i and atom j, and $f(x)$ is a so-called "switching function" that goes from 1 to 0 as x goes from 0 to infinity. A typical switching function is

$$f(r) = \frac{1}{1 + (r/r_0)^6}$$

Here r_0 here is the distance for which the value of the switching function is 1/2. Replacing the exponent 6 with a larger (smaller) number can make the switch more (less) steep.

Notice that for historical reasons you often find the function above written as$f(r) = \frac{1-(r/r_0)^6}{1-(r/r_0)^{12}}$, which is actually identical.

2.4 Number of native contacts

Imagine that you know the list of a number of contacts that are formed when a protein is folded and are not formed when it is not folded (native contacts). You can analyze a folding trajectory by counting the number of formed contacts with a collective variable such as

$$s = \sum_{ij \in NC} f(d_{ij})$$

Here NC is the list of pairs of atoms corresponding to native contacts, and f is a switching function like the one mentioned above.

2.5 Root mean square deviation

Image that you want to know if the structure of a protein correspond to the native structure or not. The best way to do it is probably to look at number of native contacts. However, a very common alternative is to compute the root mean square deviation (RMSD) between present coordinates and coordinates in the reference structure. RMSD is defined as

$$s = \sqrt{\frac{1}{N} \sum_i \left(r_i - r_i^{(ref)} \right)^2}$$

Here r_i are the coordinates of atom i and $r_i^{(ref)}$ are the coordinates of the corresponding atom in the reference structure. N is the number of atoms. Using the formula above as is would lead to the same problems that you would have using directly positions as CVs. RMSD is often computed after choosing the reference frame (i.e. after applying a rototranslation) that minimizes the RMSD itself. Sometime it is computed after choosing a translation that minimizes the RMSD. Also notice that typically not all the atoms are used for RMSD calculations. It is also possible to choose a rototranslation that minimizes the RMSD of a group of atom and them compute the RMSD using a different group of atoms. This is common for instance to track the position of a ligand in the reference frame of a protein.

2.6 Gyration radius

Imagine that you have a molecule and you want to compute its "radius". A possible definition is the gyration radius, which is defined as

$$s = \sqrt{\frac{1}{N} \sum_i (r_i - r_c)^2}$$

Here r_c is the geometric center of the molecule. As an alternative the sum might be mass weighted (in which case r_c would be the center of mass).

2.7 Cell parameters

Most molecular dynamics simulations are performed in a confined cell. When the parameters of this cell are allowed to change (e.g. in any constant pressure or constant stress simulation), they might be used as collective variable. Even though technically the are not function of the atomic coordinates, they influence the way atoms interact. For instance, a solid and a liquid might be distinguished by the different density.

2.8 Periodic boundary conditions

Some CVs are calculated taking into account implicitly the periodic boundary conditions. This is for instance the case of a distance between two atoms. The two atoms might be on opposite sides of the simulation box but still their distance be small through a periodic image. However, some variables do not play well with periodic boundary conditions and are usually computed without. Examples are: gyration radius, centers of molecules, RMSDs, etc. For this variables to be properly defined one should make sure to specify how the periodic boundaries should be solved before their calculation.

2.9 Other possible collective variables

The list of collective variables that might be used to analyze molecular systems is very long, and we will not mention all of them here. Keep in mind that any function of the atomic coordinates or of the cell parameters can in principle be used. It is much safer to use CVs that are invariant for rigid translations of the whole system (thus avoid using positions directly, rather use distances). Any function of the variables mentioned above can be used. Other variables that are commonly used in specific contexts are for instance: path collective variables, Steinhard parameters, secondary structure content, principal components, experimental observables (e.g. NMR, SAXS), etc.

3 Free-energy landscapes

If the system is at thermodynamic equilibrium, the probability of finding a given value of q and v is the Boltzmann distribution

$$P(q,v) \propto e^{-\frac{K(v)+U(q)}{k_B T}} \tag{1}$$

Here $U(q)$ is the potential energy and $K(v)$ is the kinetic energy, k_B is the Boltzmann constant and T is the temperature. The equivalent of the Boltzmann distribution for a CV is obtained by marginalization:

$$P(s) \propto \int dqdvP(q,v)\delta(s(q,v)-s)$$

This distribution can be expressed as

$$P(s) \propto e^{-\frac{F(s)}{k_B T}}$$

where $F(s)$ is the free-energy associated to the value s of the collective variable. F is also called potential of mean force since its gradient corresponds to the average force acting on a collective variable. By straightforward algebra one can express F as

$$F(s) = -k_B T \log \int dqdvP(q,v)\delta(s(q,v)-s) + C$$

where C is an arbitrary constant. Notice that $F(s)$ is nothing more that "the marginal probability expressed in energy units". States (i.e. value of s) with a large free energy are states with a low population.

Provided you have a trajectory that is long enough to be ergodic and visit all the relevant states of a system, you can obtain $F(s)$ in the following way:

1. Accumulate the histogram of the visited values of s.
2. Take its logarithm and multiply by $-k_B T$ to have energy units.

3.1 Changing variables

Always keep in mind that F is just a probability in energy units, and thus is subject to the same rules of a probability density when changing variables. For instance, if you know the probability as a function of the distance between two particles r and want to compute the probability as a function of their coordination $c = (1 + r/r_0)^{-6}$ you should use the following rule

$$P(c)dc = P(r)dr = P(r)\left|\frac{dr}{dc}\right| dc$$

Since $r = r_0\left(c^{-1/6} - 1\right)$, $\left|\frac{dr}{dc}\right| = \frac{1}{6} r_0 c^{-7/6}$, thus $P(c) \propto P(r)c^{-7/6}$. Thus

$$P(c)dc \propto P(r)c^{-\frac{7}{6}}dc$$

As a consequence, if we know $F(r)$, we can compute $F(c)$ as

$$F(c) = F(r_0\left(c^{-1/6} - 1\right)) + \frac{7}{6}k_BT\log c + C$$

Notice that the extra term would add an infinitely deep free-energy well at low c (large distance).

3.2 Entropic effects

Another interesting consequence of the fact that F represents a probability is that it implicitly contains entropic effects. Consider two non-interacting particles and measure the probability of finding them at a given distance. Since all the possible distance vectors are equiprobable, $P(r) \propto r^2$. Thus, the free-energy profile associated to their distance is $F(r) = -2k_BT\log r$. This is an entropic contribution.

3.3 Integrating free-energy landscapes

Sometime instead of being interested in the free-energy associated to a single value of s you might be interested in the stability of a whole free-energy basin. Remember the connection with probabilities, and notice that the probability for a basin A can be obtained by integrating the probability on that basin. Thus the free energy of a basin A will be

$$F_A = -k_BT\log\int_A ds e^{-\frac{F(s)}{k_BT}} + C$$

If you have two metastable basins (e.g. *syn* and *anti*), their free-energy difference is

$$F_{syn} - F_{anti} = -k_BT\log\frac{\int_{syn} ds e^{-\frac{F(s)}{k_BT}}}{\int_{anti} ds e^{-\frac{F(s)}{k_BT}}}$$

So, if the system has 80% probability to be in *syn* and 20% probability to be in *anti*, you can say that $\Delta F = 0.6\text{kcal/mol} \times \log\frac{1}{4} \approx -0.83\text{kcal/mol}$, where $k_BT \approx 0.6\text{kcal/mol}$. You will then know that if you disfavor the *anti* state by 0.83 kcal/mol the two states will have the same probability.

3.4 Overlapping metastable states

The free-energy landscape $F(s)$ associated to a collective variable can always be defined and computed as discussed above (if a long enough simulation can be performed). However, depending on the choice of the collective variable, it might be totally uninformative. Imagine for instance a case where the typical values assumed by s in reactant and product states are overlapping. Knowing the value of s will not be enough to decide in which metastable state the system is located.

Remember that s is just a function of q and v, so given q and v I can obtain a unique s. Conversely, there are many possible values of q and v that correspond to the same value of s. What I should check is that all these values (i.e. all these conformations) belong to "the same state". The definition of "belonging to the same state" depends on which time scale I am investigating. Clearly, if the typical time to isomerize a bond is on the order on 10^{-10} seconds, on the timescale of 10^{-6} seconds two isomers would correspond to the same state. This won't be true if I want to analyze the system on the timescale of picoseconds.

4 Enhancing transitions

Let's now imagine that the value of the collective variable is sufficient to completely characterize the state of the system. As discussed, knowing s does not tell us the precise position of all the atoms of the system. However, if knowing s is sufficient to characterize the system on timescale τ, the probability of observing a given value of the CV s' after a time τ will only depend on the previous value of s:

$$P(s', t+\tau|s, t) = M(s' \leftarrow s, \tau)$$

The dynamics of s will be Markovian on timescale τ.

Let's also consider s to be a discrete rather than continuous variable. This is not a strong assumption, since it can be relaxed by assuming the discrete values to be very dense. We consider a simple system where s can only take three possible values (A, B and C), such that to go from A to C it is necessary to cross$B$:

$$A \rightleftharpoons B \rightleftharpoons C$$

If $P(B) \ll P(A) \approx P(C)$ then the transition between A and C will require a lot of time. This can be interpreted as a "free energy barrier", since $F(B)$ will be much larger than $F(A)$ and $F(C)$. As you will learn later, by simply adding a bias potential that encourages the system to visit B you will be able to greatly enhance the probability to see a transition between A and C. This is basically the same way enzymes work: by relatively stabilizing the transition state of a reaction.

4.1 Again on changing variables

The fact that the free-energy function is affected by a change of variable should make you think about the meaning of a "free energy barrier". Imagine that you have a system which exhibits a free-energy barrier on the distance between two atoms $F(r)$. It is always possible to change variable and describe the system using as a CV a non linear function of the distance between the two atoms such that $F(s)$ does not have any free-energy barrier. When looking at the dynamics of s, we will see that although there is no barrier (i.e. all values of s are equiprobable) still the dynamics will be very slow. A high free-energy barrier

in r is translated into a small diffusion coefficient in s. Thus the presence of a barrier does not guarantee that the transition is a slow process, and viceversa.

In the three state example above, one might think that the state B has a very low probability not because transitions $A \to B$ and $C \to B$ are very infrequent, but because transitions $B \to A$ and $B \to C$ are very frequent.

Which is the solution of this paradox? A free-energy barrier is going to lead to rare transitions only if it is observed of a CV for which the diffusion constant is approximately constant. Notice that a change of a few kcal/mol in a barrier would be compensated by a change of orders of magnitute of the diffusion constant, so that typically barriers of several kcal/mol are "true" barriers. Still, when using highly nonlinear CVs one should pay attention to this issue.

4.2 Troubles with transition states

To be useful in enhanced sampling techniques, CVs should be able to identify transition states. Let's make an example where this is not verified. Let's start with a Markovian system with three states (A, B, and C, as above), where $P(B) \ll P(A) \approx P(C)$. Now let's make a further coarse graining and define a CV whose value is

- $s = 0$ when the system is in A or B
- $s = 1$ when the system is in C

Our variable s is capable to distinguish A and C (the metastable minima) but not capable to distinguish A from B (the transition state from one of the minima). Clearly, in this case stabilizing B (and thus also A) would not increase the probability to see a transition between A and C.

5 Conclusions

Designing CVs with the properties above is far from trivial and highly system dependent. Often, they can only be found by trial and error. In the case of analysis, trial and error means "analyze multiple times the same simulation". In the case of biased MD, trial and error means "running MD again", which can be painful. A number of methods that allow to automatize, at least partially, this search is available. Learning how to combine existing CVs in PLUMED can significantly speedup your workflow. Finally, consider that once you will have found a CV that works well for biased sampling, the choice of the CV should be considered a *result* of your work and will tell you which are the important physical or chemical processes underlying the phenomenon you are studying.

6 Summary

- Collective variables are generic functions of the microscopic coordinates of a system (hypothetically including velocities, though this is very rarely done).

- To be useful in analysis, a collective variable should at least be able to distinguish different metastable conformations.
- To be useful in enhanced sampling methods based on collective variables (e.g umbrella sampling, metadynamics, etc; you will learn more later about these methods), a collective variable should in addition be able to correctly distinguish the transition states from the metastable conformations.
- When used for biasing trajectories, variables should be made continuous.