Comparing with other community algorithms

As it was introduced in the paper, one may be tempted to apply standard algorithms for community detection in complex networks to the problem of unveiling the topology of free energy landscapes. Several authors have made use of current community detection tools [1–6] with diverse degree of success. However, there is no clear conclusion about what method is the most suitable for the problem [7]. Here we sketch the results obtained using two different approaches [5, 8] to the community detection that have already been used in previous related works. These results show that our method outperforms any of the proposed algorithms both in the computational complexity and the success in relating the CMN structure to the FEL.

Maximization of Modularity

A number of community detection algorithms have been introduced in the last years making use of the idea of modularity [4, 6]. The idea is to find a network partition into disconnected clusters so that a given function, the modularity of the network partition, is maximal. The modularity of a given partition accounts for the probability of having edges connecting nodes belonging to the same cluster in the network minus the expected probability in a randomized version of the network (having the same number of nodes and edges and preserving the strength of the nodes). Therefore, modularity provide a way to quantify the quality of a given network partition, *i.e.* the larger the modularity the better the partitioning is.

The modularity, Q, for weighted and directed networks has been introduced in ref. [8, 9] and, applied to our CMN, it takes the following form:

$$Q = \sum_{ij} [P_i P_{ji} - P_i P_j] \delta(c_i, c_j) \tag{1}$$

where the function $\delta(x, y)$ is the Kronecker delta function that takes value 1 if $x = y$ and 0 otherwise. The label c_i accounts for the community where node i is assigned in the network partition.

A number of optimization algorithms have been used to look for maximal modularity partitions. We have chosen the spectral decomposition introduced in ref. [8] and a deterministic greedy search algorithm [10, 11] to compare with the results obtained for the funnel-like CMN using the *SSD* algorithm. With these two methods we have obtained two optimal partitions of the CMN having $Q_{max} = 0.54$ and $Q_{max} = 0.53$ respectively. On the other hand, the modularity value for the basins detected by *SSD* only reaches a value $Q = 0.116$. This would seem a bad test for the *SSD*, however, as it is also shown in ref. [7], the community structure obtained by modularity-based methods is far from the basins shown in Figure 1C *of main text*. The most approximated decomposition is obtained by the greedy search algorithm where 11 communities were detected. In this case the most populated basin detected for the funnel-like potential, the basin called a in Figure 1C *of main text*, was split in 5 different communities.

This check was not so unwise if one observes that Q is related with the relaxation times of a stochastic Markov chain represented by the matrix $\mathbf{S}$ in the article:

$$Q = \sum_{\alpha} \sum_{k \neq 1} (\tau_k + 1) \Big(\sum_{i \in \alpha} \xi_i^k \Big)^2, \tag{2}$$

where α is the community index, i is the node index, and τ_k and $\vec{\xi}^k$ are the eigenvalues -relaxation times- and eigenvectors -relaxation modes- of the stochastic matrix $\mathbf{M} = \mathbf{S} - \mathbf{1}$.

Despite of the possible interpretation of communities based on modularity in CMNs, the Q value obtained for the *SSD* basins is quite farway from being the maximum.

Markov Clustering algorithm

Alternatively to the use of modularity optimization approaches, a second type of algorithms based on random walks has been introduced to infer the structure of complex networks. The most interesting algorithm of this type is the Markov Clustering (MCL) approach [5] since it has been proved to be the best suited method for detecting basins in CMNs [7]. MCL starts with the stochastic matrix $\mathbf{S}$ and iteratively alternates sequences of two matrix operations (namely expansion and inflation) until the convergence to an invariant matrix (from which network partition is finally obtained) is achieved. Expansion and inflation transformations correspond to matrix squaring and matrix Hadamard p-power respectively. These two operations are followed to a normalization step to assure that the resulting matrix is stochastic at the end of each iteration. The detected network partition depends crucially on the choice of the parameter $p \geq 1$, often called granularity. For its minimum value $p = 1$ only one community is detected, the whole network. Therefore, the success on finding the right network community structure is achieved by a fine tuning of p.

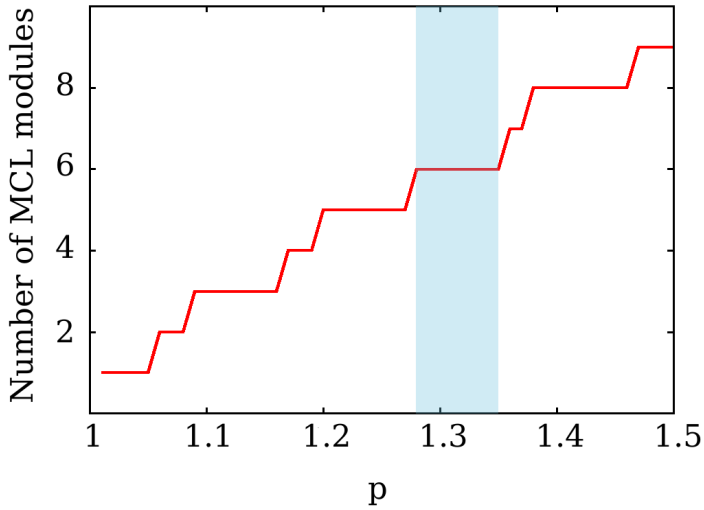

Figure S3. Markov Clustering algorithm applied to the funnel-like potential. The figure shows the number of network clusters found by MCL as a function of the granularity parameter p. As it is shown there is a range of granularity values for which MCL is able to detect the same number of clusters as the SSD algorithm.

We have performed the MCL algorithm over the CMN of the funnel-like network, see Figure S3. Unlike modularity-based algorithms, the MCL analysis points out that the network is divided into six communities (basins) for a small range of p, $1.28 \leq p \geq 1.35$ ($1.5 \leq p < 2.0$ is recommended in the literature), in agreement with the partition found using the *SSD* algorithm (see Table S1). However, the non-deterministic character of the method would tempt, if we do not have prior knowledge about the structure of the FEL, to evaluate the quality of the partitions by computing their modularity. This would lead us to a CMN partition in disagreement with the number of basins present in the FEL. Therefore, the free parameter p makes this algorithm unsuitable for analyzing general CMNs since one should have enough prior knowledge about the system to distinguish the best p value.

Finally, let us remark that both modularity optimization and MCL are computationally far more complex than the SSD algorithm. In particular, the modularity optimization scales as $\mathcal{O}(N^3)$ for the spectral decomposition and as $\mathcal{O}(mN)$ when greedy algorithm is implemented. On the other hand,

Table S1. MCL modules compared with SSD basins. The number of nodes of SSD basins are compared with those of the MCL modules ($p = 1.3$) for the funnel-like potential.

SSD Basins \ MCL Modules	1	2	3	4	5	6
a	56	3				
b		29				
c			54			
d				24		
e				1	49	
f						25

MCL has a time complexity $\mathcal{O}(N^3)$. To our knowledge the unique community detection algorithm that scales linearly in time (specifically as $\mathcal{O}(L + N)$) (with L being the number of links in the network) was proposed in ref. [12]. However, this latter method assumes a prior knowledge of the number of network clusters and, moreover, these cluster have to be of equal size. As a conclusion, the comparison between SSD algorithm and standard community detection [13] methods yields a positive assessment on the convenience of using SSD based both on its better performance and linear time complexity. This makes the use of the "Stochastic steepest descent" technique the most appropriate method for analyzing molecular dynamics data from systems of many degrees of freedom such as proteins.

References

1. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci U S A 101: 2658–2663.

2. Donetti L, Muñoz MA (2004) Detecting network communities: a new systematic and efficient algorithm. J Stat Mech 2004: P10012-1–P10012-15.

3. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci U S A 99: 7821–7826.

4. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69: 026113-1–026113-15.

5. van Dongen SM (2000) Graph Clustering by Flow Simulation. Ph.D. thesis, University of Utrecht, The Netherlands.

6. Newman MEJ (2006) From the cover: Modularity and community structure in networks. Proc Natl Acad Sci U S A 103: 8577–8582.

7. Gfeller D, De Los Rios P, Caflisch A, Rao F (2007) From the cover: Complex network analysis of free-energy landscapes. Proc Natl Acad Sci U S A 104: 1817–1822.

8. Leicht EA, Newman MEJ (2008) Community structure in directed networks. Phys Rev Lett 100: 118703-1–118703-4.

9. Newman MEJ (2004) Analysis of weighted networks. Phys Rev E 70: 056131-1–056131-9.

10. Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69: 066133-1–066133-5.

11. Wakita K, Tsurumi T (2007) Finding community structure in mega-scale social networks. arXiv:cs/0702048v1 Available: http://arxiv.org/abs/cs/0702048v1. Accessed 8 February 2007.

12. Wu F, Huberman B (2004) Finding communities in linear time: a physics approach. Eur Phys J B 38: 331–338.

13. Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech 9: P09008-1–P09008-10.