Science 27 June 2014:
Vol. 344no. 6191 pp. 1492-1496
DOI:10.1126/science.1242072

Clustering by fast search and find of density peaks

Alex Rodriguez, AlessandroLaio
SISSA(ScuolaInternazionaleSuperiorediStudiAvanzati),viaBonomea265,I-34136Trieste,Italy.
E-mail:laio@sissa.it(A.L.); alexrod@sissa.it(A.R.)

Abstract

Cluster analysis is aimed at classifying elements into categories on the basis of their similarity. Its applications range from astronomy to bioinformatics, bibliometrics, and pattern recognition. We propose an approach based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. This idea forms the basis of a clustering procedure in which the number of clusters arises intuitively, outliers are automatically spotted and excluded from the analysis, and clusters are recognized regardless of their shape and of the dimensionality of the space in which they are embedded. We demonstrate the power of the algorithm on several test cases.

Clustering algorithms attempt to classify elements into categories, or clusters, on the basis of their similarity. Several different clustering strategies have been proposed (*1*), but no consensus has been reached even on the definition of a cluster. In K-means (*2*) and K-medoids (*3*) methods, clusters are groups of data characterized by a small distance to the cluster center. An objective function, typically the sum of the distance to a set of putative cluster centers, is optimized (*3*–*6*) until the best cluster centers candidates are found. However, because a data point is always assigned to the nearest center, these approaches are not able to detect nonspherical clusters (*7*). In distribution-based algorithms, one attempts to reproduce the observed realization of data points as a mix of predefined probability distribution functions (*8*); the accuracy of such methods depends on the capability of the trial probability to represent the data.

Clusters with an arbitrary shape are easily detected by approaches based on the local density of data points. In density-based spatial clustering of applications with noise (DBSCAN) (*9*), one chooses a density threshold, discards as noise the points in regions with densities lower than this threshold, and assigns to different clusters disconnected regions of high density. However, choosing an appropriate threshold can be nontrivial, a drawback not present in the mean-shift clustering method (*10*, *11*). There a cluster is defined as a set of points that converge to the same local maximum of the density distribution function. This method allows the finding of nonspherical clusters but works only for data defined by a set of coordinates and is computationally costly.

Here, we propose an alternative approach. Similar to the K-medoids method, it has its basis only in the distance between data points. Like DBSCAN and the mean-shift method, it is able to detect nonspherical clusters and to automatically find the correct number of clusters. The cluster centers are defined, as in the mean-shift method, as local maxima in the density of data points. However, unlike the mean-shift method, our procedure does not require embedding the data in a vector space and maximizing explicitly the density field for each data point.

The algorithm has its basis in the assumptions that cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density. For each data point i, we compute two quantities: its local density ρ_i and its distance δ_i from points of higher density. Both these quantities depend only on the distances d_{ij} between data points, which are assumed to satisfy the triangular inequality.

The local density ρ_i of data point i is defined as

$$\rho_i = \sum_j \chi\left(d_{ij} - d_c\right) \qquad (1)$$

where $\chi\left(x\right) = 1$ if $x < 0$ and $\chi\left(x\right) = 0$ otherwise, and d_c is a cutoff distance. Basically, ρ_i is equal to the number of points that are closer than d_c to point i. The algorithm is sensitive only to the relative magnitude of ρ_i in different points, implying that, for large data sets, the results of the analysis are robust with respect to the choice of d_c.

δ_i is measured by computing the minimum distance between the point i and any other point with higher density:

$$\delta_i = \min_{j:\rho_j > \rho_i}\left(d_{ij}\right) \qquad (2)$$

For the point with highest density, we conventionally take $\delta_i = \max_j\left(d_{ij}\right)$. Note that δ_i is much larger than the typical nearest neighbor distance only for points that are local or global maxima in the density. Thus, cluster centers are recognized as points for which the value of δ_i is anomalously large.

This observation, which is the core of the algorithm, is illustrated by the simple example in Fig. 1. Figure 1A shows 28 points embedded in a two-dimensional space. We find that the density maxima are at points 1 and 10, which we identify as cluster centers. Figure 1B shows the plot of δ_i as a function of ρ_i for each point; we will call this representation the decision graph. The value of δ for points 9 and 10, with similar values of ρ, is very different: Point 9 belongs to the cluster of point 1, and several other points with a higher ρ are very close to it, whereas the nearest neighbor of higher density of point 10 belongs to another cluster. Hence, as anticipated, the only points of high δ and relatively high ρ are the cluster centers. Points 26, 27, and 28 have a relatively high δ and a low ρbecause they are isolated; they can be considered as clusters composed of a single point, namely, outliers.

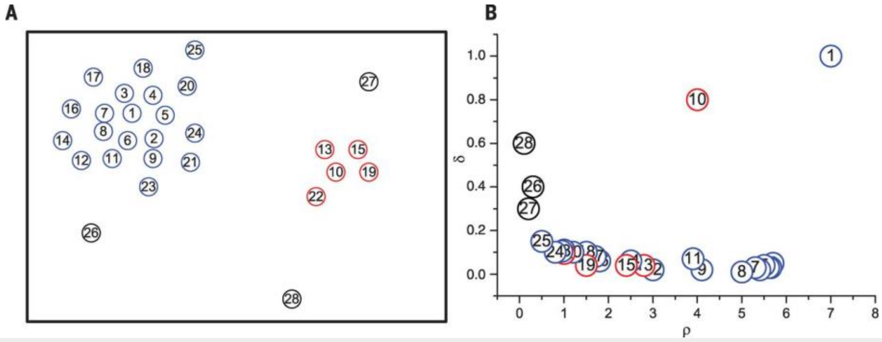

Fig. 1The algorithm in two dimensions.

(**A**) Point distribution. Data points are ranked in order of decreasing density. (**B**) Decision graph for the data in (A). Different colors correspond to different clusters.

After the cluster centers have been found, each remaining point is assigned to the same cluster as its nearest neighbor of higher density. The cluster assignment is performed in a single step, in contrast with other clustering algorithms where an objective function is optimized iteratively (2, 8).

In cluster analysis, it is often useful to measure quantitatively the reliability of an assignment. In approaches based on the optimization of a function (2, 8), its value at convergence is also a natural quality measure. In methods like DBSCAN (9), one considers reliable points with density values above a threshold, which can lead to low-density clusters, such as those in **Fig. 2E**, being classified as noise. In our algorithm, we do not introduce a noise-signal cutoff. Instead, we first find for each cluster a border region, defined as the set of points assigned to that cluster but being within a distance d_c from data points belonging to other clusters. We then find, for each cluster, the point of highest density within its border region. We denote its density by ρ_b. The points of the cluster whose density is higher than ρ_b are considered part of the cluster core (robust assignation). The others are considered part of the cluster halo (suitable to be considered as noise).

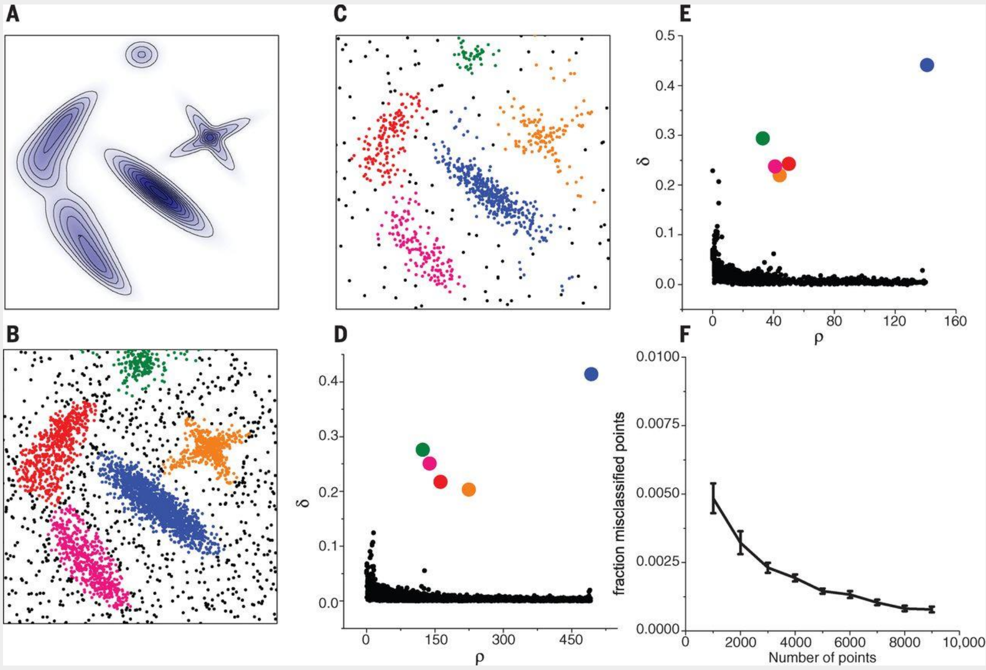

Fig.

2Results for synthetic point distributions.

(**A**) The probability distribution from which point distributions are drawn. The regions with lowest intensity correspond to a background uniform probability of 20%. (**B** and **C**) Point distributions for samples of 4000 and 1000 points, respectively. Points are colored according to the cluster to which they are assigned. Black points belong to the cluster halos. (**D** and **E**) The corresponding decision graphs, with the centers colored by cluster. (**F**) The fraction of points assigned to the incorrect cluster as a function of the sample dimension. Error bars indicate the standard error of the mean.

In order to benchmark our procedure, let us first consider the test case in **Fig. 2**. The data points are drawn from a probability distribution with nonspherical and strongly overlapping peaks (**Fig. 2A**); the probability values corresponding to the maxima differ by almost an order of magnitude. In **Fig. 2, B and C**, 4000 and 1000 points, respectively, are drawn from the distribution in **Fig. 2A**. In the corresponding decision graphs (**Fig. 2, D and E**), we observe only five points with a large value of δ and a sizeable density. These points are represented in the graphs as large solid circles and correspond to cluster centers. After the centers have been selected, each point is assigned either to a cluster or to the halo. The algorithm captures the position and shape of the probability peaks, even those corresponding to very different densities (blue and light green points in Fig. 2C) and nonspherical peaks. Moreover, points assigned to the halo correspond to regions that by visual inspection of the probability distribution in **Fig. 2A** would not be assigned to any peak.

To demonstrate the robustness of the procedure more quantitatively, we performed the analysis by drawing 10,000 points from the distribution in **Fig. 2A**, considering as a reference the cluster assignment obtained on that sample. We then obtained reduced samples by retaining only a fraction of points and performed cluster assignment for each reduced sample independently. **Figure 2F** shows, as a function of the size of the reduced sample, the fraction of points assigned to a cluster different than the one they were assigned to in the reference case. The fraction of misclassified points remains well below 1% even for small samples containing 1000 points.

Varying d_c for the data in **Fig. 2B** produced mutually consistent results (fig. S1). As a rule of thumb, one can choose d_c so that the average number of neighbors is around 1 to 2% of the total number of points in the data set. For data sets composed by a small number of points, ρ_i might be affected by large statistical errors. In these cases, it might be useful to estimate the density by more accurate measures (*10*, *11*).

Next, we benchmarked the algorithm on the test cases presented in **Fig. 3**. For computing the density for cases with few points, we adopted the exponential kernel described in (*11*). In **Fig. 3A**, we consider a data set from (*12*), obtaining results comparable to those of the original article, where it was shown that other commonly used methods fail. In **Fig. 3B**, we consider an example with 15 clusters with high overlap in data distribution taken from (*13*); our algorithm successfully determines the cluster structure of the data set. In **Fig. 3C**, we consider the test case for the FLAME (fuzzy clustering by local approximation of membership) approach (*14*), with results comparable to the original method. In the data set originally introduced to illustrate the performance of path-based spectral clustering (*15*) shown in **Fig. 4D**, our algorithm correctly finds the three clusters without the need of generating a connectivity graph. As comparison, in figs. S3 and S4 we show the cluster assignations obtained by K-means (*2*) for these four test cases and for the example in **Fig. 2**. Even if the K-means optimization is performed with use of the correct value of K, the assignations are, in most of the cases, not compliant with visual intuition.

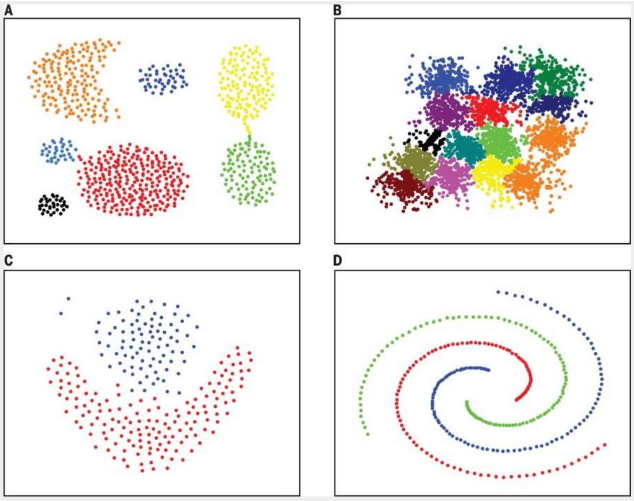

Fig. 3Results for test cases in the literature.

Synthetic point distributions from (*12*) (**A**), (*13*) (**B**), (*14*) (**C**), and (*15*) (**D**).

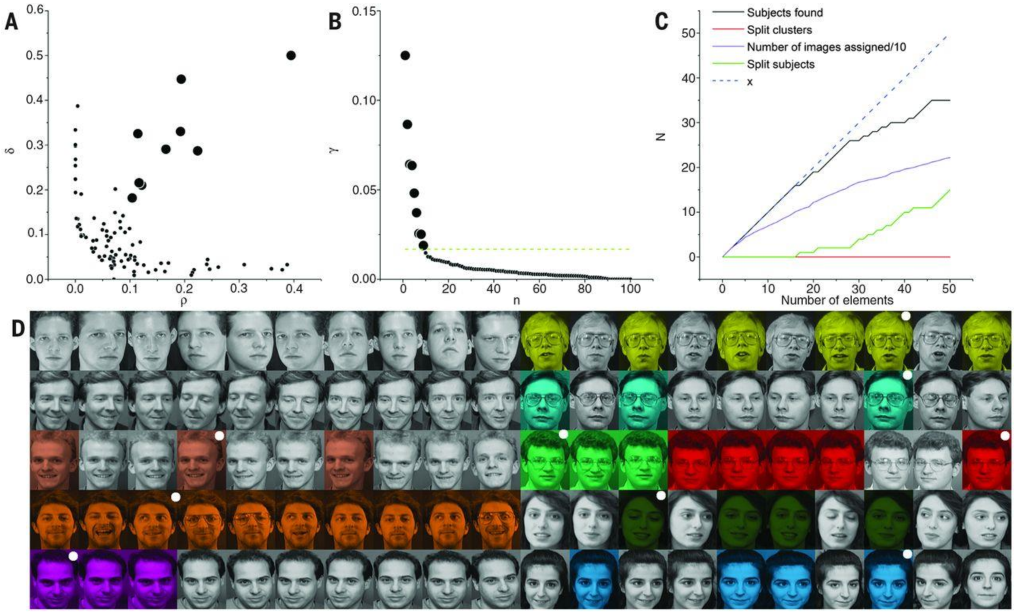

Fig. 4Cluster analysis of the Olivetti Face Database.

(**A**) The decision graph for the first hundred images in the database (*18*). (**B**) The value of $\gamma_i = \rho_i \delta_i$ in decreasing order for the data in (A). (**C**) The performance of the algorithm in recognizing subjects in the full database as a function of the number of clusters: number of subjects recognized as individuals (black line), number of clusters that include more than one subject (red line), number of subjects split in more than one cluster (green), and number of images assigned to a cluster divided by 10 (purple). (**D**) Pictorial representation of the cluster assignations for the first 100 images. Faces with the same color belong to the same cluster, whereas gray images are not assigned to any cluster. Cluster centers are labeled with white circles.

The method is robust with respect to changes in the metric that do not significantly affect the distances below d_c, that is, that keep the density estimator in Eq. 1 unchanged. Clearly, the distance in Eq. 2 will be affected by such a change of metric, but it is easy to realize that the structure of the decision graph (in particular, the number of data points with a large value of δ) is a consequence of the ranking of the density values, not of the actual distance between far away points. Examples demonstrating this statement are shown in fig. S5.

Our approach only requires measuring (or computing) the distance between all the pairs of data points and does not require parameterizing a probability distribution (8) or a multidimensional density function (10). Therefore, its performance is not affected by the intrinsic dimensionality of the space in which the data points are embedded. We verified that, in a test case with 16 clusters in 256 dimensions (16), the algorithm finds the number of clusters and assigns the points correctly (fig. S6). For a data set with 210 measurements of seven x-ray features for three types of wheat seeds from (17), the algorithm correctly predicts the existence of three clusters and classifies correctly 97% of the points assigned to the cluster cores (figs. S7 and S8).

We also applied the approach to the Olivetti Face Database (18), a widespread benchmark for machine learning algorithms, with the aim of identifying, without any previous training, the number of subjects in the database. This data set poses a serious challenge to our approach because the “ideal” number of clusters (namely of distinct subjects) is comparable with the number of elements in the data set (namely of different images, 10 for each subject). This makes a reliable estimate of the densities difficult. The similarity between two images was computed by following (19). The density is estimated by a Gaussian kernel (11) with variance $d_c = 0.07$. For such a small set, the density estimator is unavoidably affected by large statistical errors; thus, we assign images to a cluster following a slightly more restrictive criterion than in the preceding examples. An image is assigned to the same cluster of its nearest image with higher density only if their distance is smaller than d_c. As a consequence, the images further than d_c from any other image of higher density remain unassigned. In **Fig. 4**, we show the results of an analysis performed for the first 100 images in the data set. The decision graph (**Fig. 4A**) shows the presence of several distinct density maxima. Unlike in other examples, their exact number is not clear, a consequence of the sparsity of the data points. A hint for choosing the number of centers is provided by the plot of $\gamma_i = \rho_i \delta_i$ sorted in decreasing order (**Fig. 4B**). This graph shows that this quantity, that is by definition large for cluster centers, starts growing anomalously below a rank order ~ 9. Therefore, we performed the analysis by using nine centers. In **Fig. 4D**, we show with different colors the clusters corresponding to these centers. Seven clusters correspond to different subjects, showing that the algorithm is able to “recognize” 7 subjects out of 10. An eighth subject appears split in two different clusters. When the analysis is performed on all 400 images of the database, the decision graph again does not allow recognizing clearly the number of clusters (fig. S9). However, in **Fig. 4C** we show that by adding more and more putative centers, about 30 subjects can be recognized unambiguously (fig. S9). When more centers are included, the images of some of the subjects are split in two clusters, but still all the clusters remain pure, namely include only images of the same subject. Following (20) we also computed the fraction of pairs of images of the same subject correctly associated with the same cluster (r_{true}) and the fraction of pairs of images of different subjects erroneously assigned to the same cluster (r_{false}). If one does not apply the cutoff at d_c in the assignation (namely if one applies our algorithm in its general formulation), one
obtains $r_{\mathrm{true}} \sim 68\%$ and $r_{\mathrm{false}} \sim 1.2\%$ with ~ 42 to ~ 50centers, a performance comparable to a state-of-the-art approach for unsupervised image categorization (20).

Last, we benchmarked the clustering algorithm on the analysis of a molecular dynamics trajectory of trialanine in water at 300 K (21). In this case, clusters will approximately correspond to kinetic basins, namely independent conformations of the system that are stable for a substantial time and separated by free energy barriers, that are crossed only rarely on a microscopic time scale. We first analyzed the trajectory by a standard approach (22) based on a spectral analysis of the kinetic matrix, whose eigenvalues are associated with the relaxation times of the system. A gap is present after the seventh eigenvalue (fig. S10), indicating that the system has eight basins; in agreement with that, our cluster analysis (fig. S10) gives rise to eight clusters, including conformations in a one-to-one correspondence with those defining the kinetic basins (22).

Identifying clusters with density maxima, as is done here and in other density-based clustering algorithms (9,10), is a simple and intuitive choice but has an important drawback. If one generates data points at random, the density estimated for a finite sample size is far from uniform and is instead characterized by several maxima. However, the decision graph allows us to distinguish genuine clusters from the density ripples generated by noise. Qualitatively, only in the former case are the points corresponding to cluster centers separated by a sizeable gap in ρ and δ from the other points. For a random distribution, one instead observes a continuous distribution in the values of ρ and δ. Indeed, we performed the analysis for sets of points generated at random from a uniform distribution in a hypercube. The distances between data points entering in Eqs. 1 and 2 are computed with periodic boundary conditions on the hypercube. This analysis shows that, for randomly distributed data points, the quantity $\gamma_i = \rho_i \delta_i$ is distributed according to a power law, with an exponent that depends on the dimensionality of the space in which the points are embedded. The distributions of γ for data sets with genuine clusters, like those in **Figs. 2** to **4**, are strikingly different from power laws, especially in the region of high γ (fig. S11). This observation may provide the basis for a criterion for the automatic choice of the cluster centers as well as for statistically validating the reliability of an analysis performed with our approach.

Supplementary Materials

www.sciencemag.org/content/344/6191/1492/suppl/DC1

Figs. S1 to S11

Data S1

Full paper: http://www.sciencemag.org/content/344/6191/1492.full

ReferencesandNotes

1. R. Xu, D. Wunsch 2nd, Survey of clustering algorithms. IEEE Trans. Neural Netw. 16, 645–678 (2005). CrossRefMedlineWeb of Science Search Google Scholar
2. J. MacQueen, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, L. M. Le Cam, J. Neyman, Eds. (Univ. California Press, Berkeley, CA, 1967), vol. 1, pp. 281–297.
3. L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344 (Wiley-Interscience, New York, 2009).

4. B. J. Frey, D. Dueck, Clustering by passing messages between data points. Science 315, 972–976 (2007). Abstract/FREE Full Text
5. J. H. Ward Jr., Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244 (1963). CrossRef Search Google Scholar
6. F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition (Wiley, New York, 1999).
7. A. K. Jain, Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 31, 651–666 (2010). CrossRefWeb of Science Search Google Scholar
8. G. J. McLachlan, T. Krishnan, The EM Algorithm and Extensions (Wiley Series in Probability and Statistics vol. 382, Wiley-Interscience, New York, 2007).
9. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, E. Simoudis, J. Han, U. Fayyad, Eds. (AAAI Press, Menlo Park, CA, 1996), pp. 226–231.
10. K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans. Inf. Theory 21, 32–40 (1975). CrossRefWeb of Science Search Google Scholar
11. Y. Cheng, Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. 17, 790 (1995). CrossRefWeb of Science Search Google Scholar
12. A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation. ACM Trans. Knowl. Discovery Data 1, 4 , es (2007). CrossRef Search Google Scholar
13. P. Fränti, O. Virmajoki, Iterative shrinking method for clustering problems. Pattern Recognit. 39, 761–775 (2006). CrossRefWeb of Science Search Google Scholar
14. L. Fu, E. Medico, FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics 8, 3 (2007). CrossRefMedline Search Google Scholar
15. H. Chang, D.-Y. Yeung, Robust path-based spectral clustering. Pattern Recognit. 41, 191–203 (2008). CrossRefWeb of Science Search Google Scholar
16. P. Fränti, O. Virmajoki, V. Hautamäki, Fast agglomerative clustering using a k-nearest neighbor graph. IEEE Trans. Pattern Anal. Mach. Intell. 28, 1875–1881 (2006). CrossRefMedlineWeb of Science Search Google Scholar
17. M. Charytanowicz et al., Information Technologies in Biomedicine (Springer, Berlin, 2010), pp. 15–24.
18. F. S. Samaria, A. C. Harter, in Proceedings of 1994 IEEE Workshop on Applications of Computer Vision (IEEE, New York, 1994), pp. 138–142.
19. M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, M. K. Markey, Complex wavelet structural similarity: A new image similarity index. IEEE Trans. Image Process. 18, 2385–2401 (2009). CrossRefMedlineWeb of Science Search Google Scholar
20. D. Dueck, B. Frey, ICCV 2007. IEEE 11th International Conference on Computer Vision (IEEE, New York, 2007), pp. 1–8.
21. F. Marinelli, F. Pietrucci, A. Laio, S. Piana, A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. PLOS Comput. Biol. 5, e1000452 (2009). CrossRefMedline Search Google Scholar
22. I. Horenko, E. Dittmer, A. Fischer, C. Schütte, Automated model reduction for complex systems exhibiting metastability. Multiscale Model. Simulation 5, 802–827 (2006). CrossRefWeb of Science Search Google Scholar