




Determining Optimal Coarse-Grained Representation for Biomolecules Using Internal Cluster Validation Indexes

Zhenliang Wu,^[a] Yuwei Zhang,^[a] John Zenghui Zhang ^[a,b] Kelin Xia ^{*,[c,d]} and Fei Xia ^{*,[a,b]}

The development of ultracoarse-grained models for large biomolecules needs to derive the optimal number of coarse-grained (CG) sites to represent the targets. In this work, we propose to use the statistical internal cluster validation indexes to determine the optimal number of CG sites that are optimized based on the essential dynamics coarse-graining method. The calculated curves of Calinski-Harabasz and Silhouette Coefficient indexes exhibit the extrema corresponding to the similar CG

numbers. The calculated ratios of the optimal CG numbers to the residue numbers of fine-grained models are in the range from 4 to 2. The comparison of the stability of index results indicates that Calinski-Harabasz index is the better choice to determine the optimal CG representation in coarse-graining. © 2019 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.26070

Introduction

All-atom molecular dynamics (MD) simulation of large biomolecules^[1–3] can be prohibitively expensive. To solve this problem, many researchers have devoted to developing the so-called coarse-grained (CG) models^[4–8] for accelerating simulation. The CG models could be constructed from the fine-grained models such as all-atom model by using the “bottom-up” strategies.^[9,10] By resorting to the strategies, we could set up a connection between the statistical properties of fine-grained ensembles and CG ensembles.^[11–21] Construction of CG models based on all-atom models needs a preliminary definition of how to map fine-grained models to CG models. For example, the CG representation of amino acids in protein is usually defined according to their chemical groups. In the high-resolution Martini CG models,^[22–24] each chemical group in the amino acid is usually represented by more than two CG beads. However, for the construction of low-resolution ultracoarse-grained (UCG) model,^[17,18,25–29] how to map a target biomolecule into its UCG representation is more complicated, since the number of UCG sites is usually far less than the number of residues in proteins.

Coarse-graining a protein with a given sequence composed of M residues into the number N ($N < M$) UCG beads involves two major optimization problems in mathematics.^[16,17,30] First, if the sequential M residues are CG into a specific number of N UCG beads, it means that the protein sequence needs to be divided into N clusters that are separated by $N-1$ boundaries along the protein backbone, which refers to the problem of boundary optimization. Second, the coarse-graining number N itself should be considered as a variable of defined property function $f(N)$ of target system, which means that the magnitude of N should make the property function $f(N)$ the most optimal. Thus, determining the optimal number N needs to optimize $f(N)$ with respect to the variable N , namely, the problem of N optimization.

In order to solve the problem of boundary optimization, Zhang et al. proposed a combined algorithm^[7,17,18] composed

of simulated annealing^[31] and steepest descent^[32] (SASD) to obtain the optimal $N-1$ boundaries dividing the protein sequence into N clusters, based on their proposed essential dynamics coarse-graining (ED-CG) method.^[17,18] Xia and coworkers developed two efficient and rapid algorithms,^[16,33] namely, the stepwise optimization imposed with boundary constraint (SOBC)^[33] and stepwise local iterative optimization (SLIO)^[16] for coarse-graining. The further comparison of SASD, SOBC, and SLIO combined with ED-CG demonstrates that SLIO is the most accurate and fast algorithm for coarse-graining.^[30] In addition, Koehl et al. utilized the method of renormalization group^[34,35] to optimize the CG beads for large biomolecules. Chen and Habeck also introduced a Bayesian approach^[36] for coarse-graining biomolecular structures. Zhang and Voth proposed a density-based ED-CG method^[18] to coarse-grain a given protein without the sequence known. Recently, our group

[a] Z. Wu, Y. Zhang, J. Z. Zhang, F. Xia
Shanghai Engineering Research Center of Molecular Therapeutics and New Drug Development, School of Chemistry and Molecular Engineering, East China Normal University, Shanghai 200062, China
E-mail: fxia@chem.ecnu.edu.cn

[b] J. Z. Zhang, F. Xia
NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China

[c] K. Xia
Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, 637371, Singapore
E-mail: xiakelin@ntu.edu.sg

[d] K. Xia
School of Biological Sciences, Nanyang Technological University, 637371, Singapore

Contract Grant sponsor: Nanyang Technological University; Contract Grant number: Startup Grant M4081842; Contract Grant sponsor: National Natural Science Foundation of China; Contract Grant numbers: 21433004, 21673185, 21773065, 21873078; Contract Grant sponsor: Singapore Ministry of Education Academic Research Fund; Contract Grant numbers: Tier 1RG31/18, Tier 2 MOE2018-T2-1-033

© 2019 Wiley Periodicals, Inc.

developed a new convolutional^[37] and K-means^[38] coarse-graining method^[29] to construct UCG models from the low-resolution data of cryoelectron microscopy.^[39]

On the other hand, in order to solve the problem of N optimization, Sinitskiy et al. proposed an empirical formula^[40] to determine the optimal N to represent large biomolecules. Xia and coworkers developed the SLIO algorithm to determine the optimal coarse-graining number N . In particular, it is highlighted that the SLIO algorithm does not depend on any empirical parameters and it could derive the change of the values of the function $f(N)$ with respect to the variable N . The previous results optimized by SLIO show that the curve of $f(N)$ change monotonously with N , as well as the increments of $f(N)$ tend to approach a constant as N becomes large.^[30] However, the monotonous curve of $f(N)$ did not exhibit any extrema so that we needed to set a threshold to judge whether the function has been converged.

Actually, the “ N optimization” aforementioned is a problem of data classification in nature. The purpose of data classification is to find out the regularity or connection within the data clusters. The problem of coarse-graining a protein in biology could be categorized as a specific kind of data classification without the external information known. The number of CG sites in coarse-graining, namely, the number of clusters, is an important parameter in the data classification. In the classical K-means algorithm, the parameter K is always manually determined just based on the data distribution. In this work, we propose to use the statistical internal cluster validation indexes including the Calinski-Harabasz index (CH index)^[41–43] and Silhouette-Coefficient index (SC index)^[41–43] to determine the optimal number of clusters, namely, the optimal number of CG sites in coarse-graining. The ED-CG method is used in this work to coarse-grain the biomolecules into CG sites. The CG sites derived from ED-CG method could preserve the essential dynamics of intradomains of proteins, as proposed by Zhang et al.^[17] in the previous study.

Theory and Methods

The ED-CG method and SLIO algorithm

In this work, we coarse-grain protein sequences by using the SLIO algorithm^[16] based on the ED-CG method.^[17,30] In the ED-CG method, a property function $f_{\text{ED-CG}}(N)$ is defined as the sum of the squared displacement difference of pairwise $\text{C}\alpha$ atoms, as shown in eq.(1):

$$\chi^2_{\text{ED-CG}} = \frac{1}{3N} \sum_{l=1}^N \frac{1}{n_l} \sum_{t=1}^{n_l} \left(\sum_{i \in l} \sum_{j \geq i \in l} |\Delta \mathbf{r}_i(t) - \Delta \mathbf{r}_j(t)|^2 \right) \quad (1)$$

where N is the number of CG sites and n_l denotes the number of snapshots of $\text{C}\alpha$ atoms extracted from MD trajectories. The $\Delta \mathbf{r}_i$ and $\Delta \mathbf{r}_j$ represent the displacements of the i th and j th $\text{C}\alpha$ atoms, respectively. If the i th $\text{C}\alpha$ atom moves in a direction similar to the j th $\text{C}\alpha$ atom, the squared displacement difference $|\Delta \mathbf{r}_i(t) - \Delta \mathbf{r}_j(t)|^2$ should be small. In the ED-CG method, the optimal number N should make the residual $\chi^2_{\text{ED-CG}}$ achieve minimal.

In order to search for the optimal number N , we developed the SLIO algorithm, which could be combined with ED-CG for coarse-graining biomolecules. SLIO is an accurate and efficient optimization algorithm, which has been demonstrated by tests performed by us.^[16,30] The detailed principle of the SLIO algorithm is described in the previous study.^[16]

Definition of distance metric

1. Through all-atom MD simulation, we can get the coordinates of $\text{C}\alpha$ atoms of frames, denoted as $C_i^w : (x_i^w, y_i^w, z_i^w)$, where C_i^w refers to the coordinates of the i th $\text{C}\alpha$ atoms in the w th frame. W denotes the total number of frames and we chose arbitrary 2000, 3500, and 5000 frames in data analysis.
2. Calculate the equilibrium position coordinates $\bar{C}_i : (\bar{x}_i, \bar{y}_i, \bar{z}_i)$ over all chosen frames, where $\bar{x}_i = \frac{1}{W} \sum_{w=1}^W x_i^w$, $\bar{y}_i = \frac{1}{W} \sum_{w=1}^W y_i^w$, and $\bar{z}_i = \frac{1}{W} \sum_{w=1}^W z_i^w$. Calculate the displacement vector $\Delta r_i^w : (\Delta x_i^w, \Delta y_i^w, \Delta z_i^w)$ of each atom in each frame, where $\Delta x_i^w = x_i^w - \bar{x}_i$, $\Delta y_i^w = y_i^w - \bar{y}_i$, and $\Delta z_i^w = z_i^w - \bar{z}_i$. The calculated displacement vector Δr_i^w in Table 1 could be considered as a $3 \times M$ dimensional vector for index calculation, where M denotes the number of $\text{C}\alpha$ atoms in a frame.
3. Define the distance metric of the data points based on the data set. In the traditional clustering algorithm, the Euclidean distance is always used to measure the distance or the dissimilarity between data points. The distance metric is defined in eq. (2):

$$d^2(i, j) = \|\bar{C}_i - \bar{C}_j\|_2^2 * W + \sum_{w=1}^W \|\Delta r_i^w - \Delta r_j^w\|_2^2 \quad (2)$$

where the first term represents the similarity of equilibrium positions scaled by the number of frames W and the second term measures the difference of displacement vectors over the total W frames.

Table 1. The initial data set in this article. The equilibrium position coordinates of each $\text{C}\alpha$ atom and displacement vectors of each frame are used as the features for internal cluster validation calculation.

No.C	Coordinates			Vector Δr^1			Vector Δr^2			Vector Δr^W		
1	\bar{x}_1	\bar{y}_1	\bar{z}_1	Δx_1^1	Δy_1^1	Δz_1^1	Δx_1^2	Δy_1^2	Δz_1^2	Δx_1^W	Δy_1^W	Δz_1^W
2	\bar{x}_2	\bar{y}_2	\bar{z}_2	Δx_2^1	Δy_2^1	Δz_2^1	Δx_2^2	Δy_2^2	Δz_2^2	Δx_2^W	Δy_2^W	Δz_2^W
.....	
M	\bar{x}_M	\bar{y}_M	\bar{z}_M	Δx_M^1	Δy_M^1	Δz_M^1	Δx_M^2	Δy_M^2	Δz_M^2	Δx_M^W	Δy_M^W	Δz_M^W

Internal cluster validation indexes

The cluster validation indexes^[41–43] can be divided into two categories: external and internal cluster validation indexes. The external cluster validation indexes are based on external supervised learning information of data, while the internal indexes are based only on the intrinsic properties of data distribution itself. Since there is no supervised learning information provided for the problem of coarse-graining, external cluster validation indexes cannot be used in this case. We choose four commonly used internal indexes including the CH index,^[41–43] SC index,^[41–43] Davies–Bouldin index,^[43] and Dunn index^[43] for coarse-graining, and the curves of the Davies–Bouldin index and the Dunn index are unstable and discontinuous. Thus, we mainly discuss the coarse-graining results by using the CH and SC indexes.

The CH index^[41–43] is defined in eq. (3):

$$\text{CH}(N) = \frac{\text{SSB}/(N-1)}{\text{SSW}/(M-N)} \quad (3)$$

where $\text{SSB} = \sum_{k=1}^N [n_k * d^2(\bar{X}_k, \bar{X})]$, $\text{SSW} = \sum_{k=1}^N \sum_{i \in I_k} d^2(X_i, \bar{X}_k)$, $X_i = (\bar{C}_i, \Delta r_i^1, \Delta r_i^2, \dots, \Delta r_i^w)$, $\bar{X}_k = \frac{1}{n_k} \sum_{i \in I_k} X_i$ ($I_k = \{i_1, i_2, \dots, i_{n_k}\}$ is the index set for all the X_i in the k th cluster), and $\bar{X} = \frac{1}{M} \sum_{i=1}^M X_i$. N denotes the variable of cluster number and M denotes the number of total C α atoms. The notations SSB and SSW denote the intercluster and intracluster similarity of clusters, respectively. The CH index is calculated according to eq. (3) and its maximum corresponds to the optimal number of CG sites N .

The SC index^[41–43] is defined as in eq. (4):

$$\text{SC}(N) = \frac{1}{M} \sum_{i=1}^M s(X_i); s(X_i) = \frac{b(X_i) - a(X_i)}{\max[a(X_i), b(X_i)]} \quad (4)$$

where $a(X_i) = \frac{1}{n_k - 1} \sum_{i \in I_k, j \in I_k, i \neq j} d^2(X_i, X_j)$ and $b(X_i) = \min \left\{ \frac{1}{n_k} \sum_{i \in I_k, j \in I_k} d^2(X_i, X_j) \right\}$. What is different from CH index

is that the calculation of $a(X_i)$ and $b(X_i)$ in the SC index does not involve the center point of each cluster, but is described by the distance metric between every two data points within the cluster or between the clusters. The index is calculated based on different number of CG sites, and the value of N corresponding to the maximum is the optimal number of CG sites.

Results and Discussion

SSB, SSW, and CH index for Ras

In order to obtain the relationship between the values of ED-CG function^[17] and the variable N , we employed the SLIO algorithm^[16] to optimize the initial clusters. The SLIO is one of the efficient coarse-graining algorithms developed by our group. One of its advantages is that it could give the results of optimized function values of ED-CG with respect to the variable N , whereas the SASD^[17] and SOBC³³ algorithms could not achieve it. Our previous coarse-graining results^[30] revealed that the

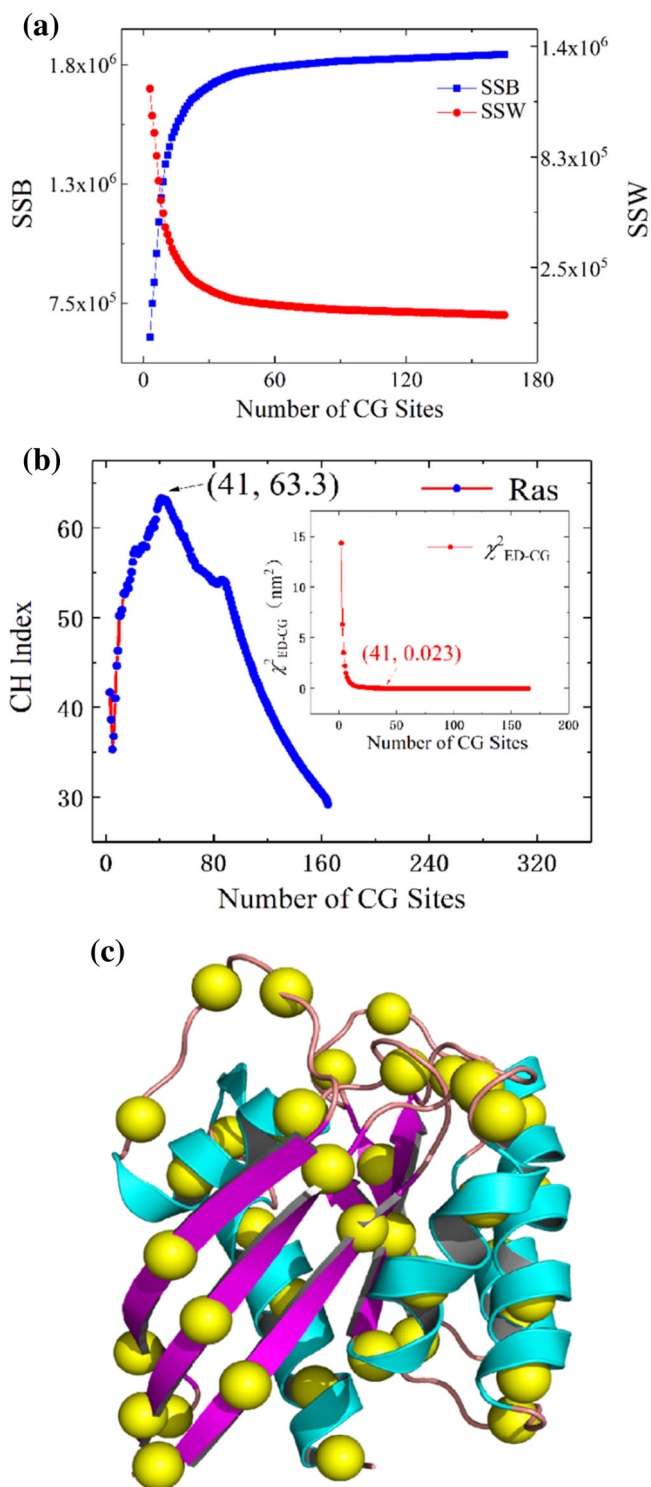


Figure 1. (a) Calculated values of SSW and SSB of Ras system with respect to the number of CG sites, denoted by the red and blue curves respectively. (b) The calculated CH index curve for Ras, with the maximal value 63.3 corresponding to the number of CG sites 41. The inset shows the optimized functional values of ED-CG by SILO algorithm. (c) The cartoon representation of Ras with its optimal 41 CG sites. [Color figure can be viewed at wileyonlinelibrary.com]

values of ED-CG function $f_{\text{ED-CG}}(N)$ optimized by SLIO decreased with the CG number N smoothly and monotonically. Also, the functional increments approached to zero as N adopted a large

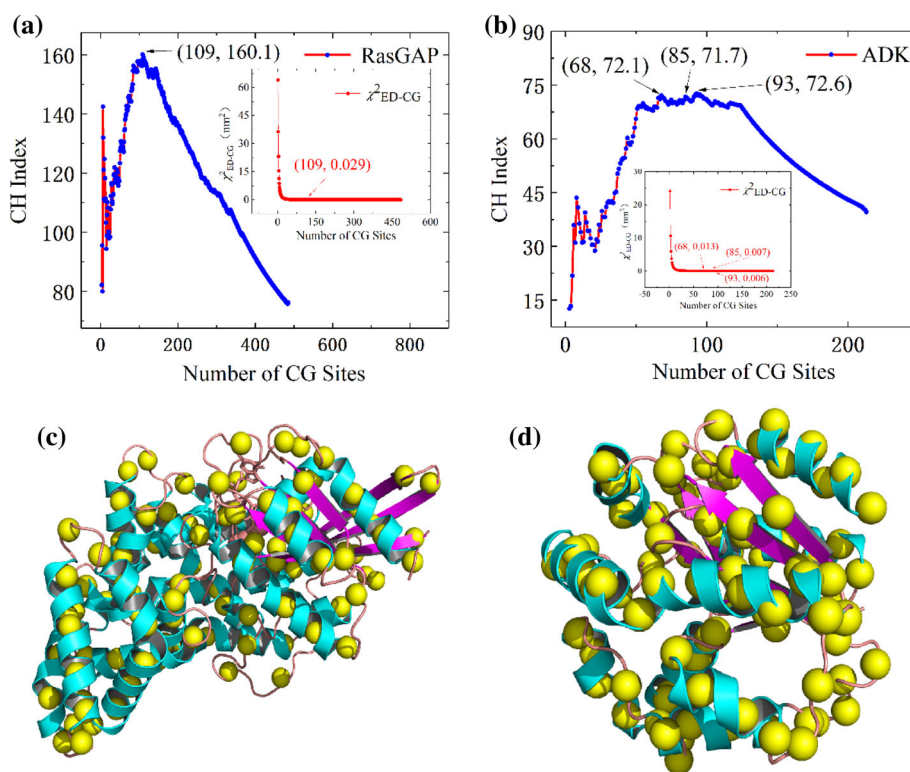


Figure 2. Calculated blue curves of CH index for (a) RasGAP and (b) ADK systems, with the insets showing the optimized functional values of ED-CG by SILO algorithm. The maximal value 160.1 of CH index in (a) corresponds to the number of CG sites 109 and the three maximal values 72.6, 72.1, and 71.7 of CH index in (b) correspond to 93, 68, and 85, respectively. The cartoon representations of RasGAP and ADK with their optimal CG sites are shown in (c) and (d), respectively. [Color figure can be viewed at wileyonlinelibrary.com]

number. In the previous work, we set an empirical threshold to judge whether the optimized functional values have achieved convergence. If the increment of $f_{\text{ED-CG}}(N)$ was smaller than the threshold in certain round, the corresponding N would be considered as an optimal number to represent the target biomolecule. However, it had to be pointed that the resulted N aforementioned did not correspond to a minimum of the curve of $f_{\text{ED-CG}}(N)$ rigorously. The reason lies in the fact that the ED-CG method itself actually only considers the intracluster similarity according to its definition. Thus, it is expected that we can resort to the CH index whose definition considers the intracluster and intercluster similarities to determine the optimal N .

To estimate the CH index in eq. (3), we need to calculate SSB and SSW at first. In the first case of coarse-graining system, we chose the important oncogenic Ras protein,^[44–48] since our group aimed to develop a UCG model of Ras for its functional study. Ras protein has a total number of 166 residues and its maximal residue-based CG number N in theory is 166. We conducted a routine MD simulation and extracted the snapshots from its trajectory for the CH index calculation. All the simulations for target systems in this study were carried out for 100 ns under the NPT ensemble using AMBER 18 package.^[49,50] The calculated results of SSB and SSW are shown in Figure 1a. It can be seen that both SSB and SSW change monotonically with the increased of number of CG sites and both of them tend to approach constant. It is easy to understand the increased and decreased tendency of SSB and SSW, since the increased N definitely leads to the increased intercluster difference and decreased intracluster difference according to their definitions in the computational section.

On the basis of the results of SSB and SSW, CH index is calculated according to eq. (3) and the results are shown in Figure 1b. The inset of Figure 1b shows that the optimized values of ED-CG function approach to zero as the number N of CG sites becomes larger. This curve does not exhibit an extremum and the optimal number N cannot be determined rigorously in mathematics. However, the calculated curve of the CH index with respect to N in Figure 1b shows that the distribution of CH index appears to be unimodal and its peak is very obvious. The maximal value of CH index is 63.3 and corresponds to the optimal CG number $N = 41$. The total number of residues in Ras is 166 and the ratio of residues to the optimal number of CG sites is 4.05:1, approximately being 4:1. This ratio is in line with our expectation on the resolution of CG representation of target biomolecule. It means that if the ratio is too small such as 1:1, the CG model has the same resolution as the fine-grained model and cannot save any computational time in simulation. If the ratio is too large, this means that the CG sites are too coarse to represent the crucial secondary structures in protein, which has been pointed out by us in the previous coarse-graining of biomolecules.^[28] Figure 1c clearly shows the optimal 41 CG sites derived from the CH index and these CG sites represent the crucial secondary structures such as the α -helices, β -sheets, and flexible loops of Ras.

CH indexes of RasGAP and ADK

To validate the CH index for determining the optimal CG number further, we also calculated the CH indexes of the RasGAP complex^[51] and the adenylate kinase (ADK),^[52] as shown in Figure 2. Figure 2a shows that the CH index distribution of

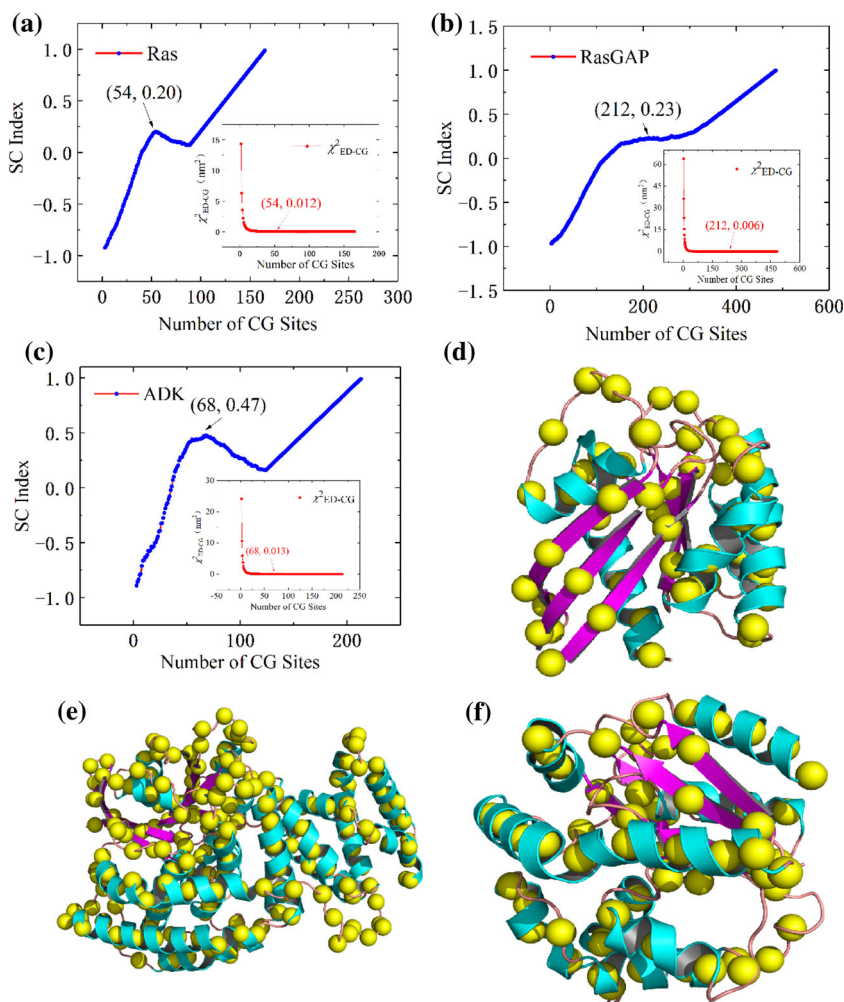


Figure 3. (a)–(c) Calculated blue curves of SC index for Ras, RasGAP, and ADK, respectively. The optimal CG sites indicated by extrema are 54, 212 and 68 in (a)–(c), respectively. The cartoon representations of Ras, RasGAP, and ADK with their optimal CG sites are visualized in (d)–(f), respectively. [Color figure can be viewed at wileyonlinelibrary.com]

RasGAP is similar to that of Ras and the maximum of the CH index curve with the value 160.1 corresponds to the number $N = 109$. The number of residues of RasGAP is 486 and the ratio of optimal CG sites to the whole number is about 4.45:1. The ratio 4.45 is similar to that ratio 4.05 of the Ras system, which seems to be a reasonable value to determine the resolution of CG model relative to the fine-grained model.

Figure 2b shows the calculated results of the CH index for the ADK system. The contour of the CH index curve of ADK is remarkably different from the ones of Ras and RasGAP systems. The CH index distribution of ADK has a flat plateau rather than a sharp peak of RasGAP. It can be seen that the values of the

CH index do not change drastically at this rugged plateau, and it has quite a few maxima with similar values close to each other in the range from $N = 50$ to 120. For instance, the three maxima with the calculated CH index values 72.1, 71.7, and 72.6 are close to each other and correspond to the CG number $N = 68, 85,$ and 93. The total number of residues of ADK is 214 and the corresponding CG ratios are 3.1, 2.5, and 2.3 for the three numbers, respectively.

We note that the global maximum of the CH index with the magnitude 72.6 corresponds to a coarse-graining ratio 2.3, which indicates that one CG bead represents nearly two particles of a fine-grained model. However, such a ratio is considered to be small for mapping a fine-grained model to a CG model, while other larger ones such as 3.1 seems to be a more reasonable alternative for mapping. Therefore, it appears that determining the optimal number of CG sites depends on the specific case. In the case of ADK whose CH index has a flat plateau, we propose that we can choose a certain maximum of CH index that results in a more reasonable coarse-graining ratio, not necessarily a global maximum. If we pursue a higher accuracy to represent the original fine-grained model by CG model, we can use a small ratio to map the fine-grained model to the CG model. On the contrary, if much less computational cost is required for the simulation with CG model, a larger ratio needs

Table 2. The optimal CG sites determined by CH and SC indexes and calculated ratios of the number of CG sites to the total number of $C\alpha$ atoms for Ras, RasGAP, and ADK systems, respectively.

Systems		CH index	SC index
Ras	Optimal number	41	54
	Ratios	4.05	3.02
RasGAP	Optimal number	109	212
	Ratios	4.46	2.29
ADK	Optimal number	68, 85, 93	68
	Ratios	3.15, 2.52, 2.30	3.15

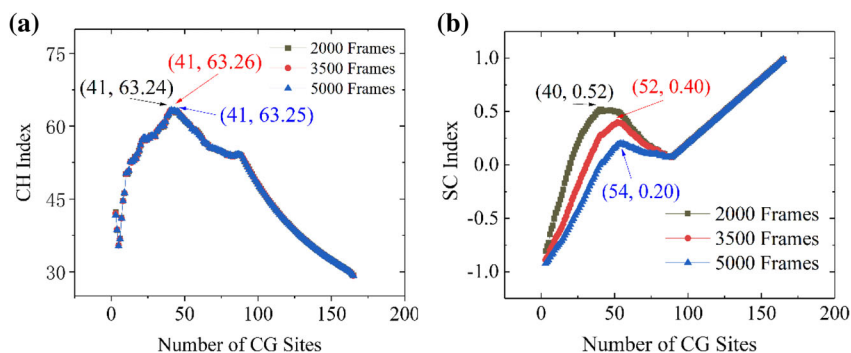


Figure 4. (a) Calculated CH index curves with 2000, 3500, and 5000 frames extracted from MD trajectory of Ras. All the maxima of the three curves correspond to the number of CG sites 41. (b) Calculated SC index curves with 2000, 3500, and 5000 frames extracted from MD trajectory of Ras. The maxima of the black, red, and blue curves correspond to 40, 52, and 54 CG sites. [Color figure can be viewed at wileyonlinelibrary.com]

to be selected for the construction of the CG model. The optimal CG sites of RasGAP and ADK with 109 and 93 determined by CH index are visualized in Figures 2c and 2d, respectively.

Estimation of SC index

As what is defined in the computational section, the SC index is different from the CH index and it depends only on the distance metrics between data points, rather than the centers of clusters. Figures 3a–3c shows the calculated results of SC indexes for Ras, RasGAP, and ADK systems, respectively. Figure 3a clearly shows that the SC index curve is smoother than that of the CH index, as well as the maximum is quite obvious and unique. It has a maximum as the CG number adopts 54. The ratio of the number of residues to the optimal number 54 is about 3.1:1. This ratio 3.1 is slightly lower than that obtained by CH index. For the RasGAP and ADK systems, the optimal CG sites indicated by the SC index curves are 212 and 68, with the corresponding ratios being 2.3:1 and 3.2:1, respectively. The SC index curves of Ras and ADK systems have obvious peaks while this not visible.

In addition, it can be seen that when the number of CG sites is larger than half of the residue numbers ($N \geq \frac{M}{2}$), the SC index begins to increase with the cluster number in a linear-like way. This is due to the reason that when the number of CG sites is large ($N \geq \frac{M}{2}$), more and more individual atom forms a cluster by itself and $a(X_i)$ will be equal to zero in eq. (4). With the further increase in the number of CG sites (N), more and more $a(X_i)$ will be equal to zero. At the same time, $b(X_i)$ will remain relatively the same, and thus, $s(X_i)$ will systematically increase. This is why there is a linearly-like increase of the SC index. Traditionally, both the CH index and the SC index are considered for the measuring of clustering with a relatively large number of elements inside each cluster. However, in this paper, the size of each cluster can be much smaller. And SC index may not be a good measurement when the size of the cluster reduces to smaller than two. The optimal CG sites of Ras, RasGAP, and ADK derived from the curves of SC index are shown in Figures 3c–3e, respectively.

In summary, Table 2 shows all the results of optimal number of CG sites determined by CH index and SC index for Ras, RasGAP, and ADK, respectively. The comparison of results indicate that the ratios determined by SC index are a little smaller than those determined by CH index for Ras and RasGAP systems, while they are similar for ADK. In the practical

application, a large ratio such as 4.0 or 3.0 might be the better choice for mapping a fine-grained model to a CG model.

Stability of CH and SC indexes

Since both CH and SC indexes were calculated based on the frames extracted from MD trajectories, we tested the stability of the CH and SC indexes for the Ras system by using different number of frames. Figures 4a and 4b show the calculated curves of CH and SC indexes with 2000, 3500, and 5000 frames extracted from Ras trajectories, respectively. Figure 4a shows that three curves of CH indexes calculated from 2000, 3500, and 5000 frames almost completely overlap with each other, which indicates that the CH index results are not dependent on the number of frames. It also means that we can use a less number of frames to calculate the CH index due to its stability.

In contrast, Figure 4b shows that the calculated curves of SC indexes with different frames are severely dependent on the number of frames. The three SC index curves differ in the peak locations with each other and the maxima of curves calculated from 2000, 3500, and 5000 frames correspond to the CG numbers 40, 52, and 54, respectively. Among them, the black curve calculated from 2000 frames has the largest local maximal value 0.52, and the corresponding $N = 40$ is the smallest. As the number of frames increases, the values of the calculated maxima gradually decrease and the corresponding N increases. The compared results reveal that the SC index is dependent on the number of MD frames and less stable than the CH index.

Conclusions

In this work, we propose to use internal cluster validation indexes such as the CH index and the SC index to determine the optimal CG sites in coarse-graining. A distance metric was designed and combined with the ED-CG method for index calculation. The CH and SC indexes were calculated for three biomolecular systems including Ras, RasGAP, and ADK, respectively. All the ratios determined by the CH and SC indexes are in the range of 4–2, which means that the number of CG sites could be the 1/4–1/2 of the total residue number of the fine-grained models. The calculated ratios are in line with the expectation for the resolution of CG models. The test results also reveal that the CH index is more stable in using different number of frames than the SC index. Thus, we suggest using the CH index to determine the optimal number of CG sites of biomolecules in future coarse-graining.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. 21773065, 21433004, 21673185 and 21873078), Nanyang Technological University Startup Grant M4081842, Singapore Ministry of Education Academic Research fund Tier 1 RG31/18, Tier 2 MOE2018-T2-1-033. We acknowledge the support of the NYU-ECNU Center for Computational Chemistry at NadenylateYU Shanghai. We also thank the ECNU Public Platform for Innovation(001) for providing computer time.

Keywords: coarse-graining · optimal CG sites · internal cluster validation index · CH index · SC index

How to cite this article: Z. Wu, Y. Zhang, J. Z. Zhang, K. Xia, F. Xia. *J. Comput. Chem* **2020**, *41*, 14–20. DOI: 10.1002/jcc.26070

- [1] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, D. E. Shaw, *Curr. Opin. Struct. Biol.* **2009**, *19*, 120.
- [2] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, D. E. Shaw, *Annu. Rev. Biophys.* **2012**, *41*, 429.
- [3] A. C. Pan, H. Xu, T. Palpant, D. E. Shaw, *J. Chem. Theory Comput.* **2017**, *13*, 3372.
- [4] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, A. Kolinski, *Chem. Rev.* **2016**, *116*, 7898.
- [5] A. J. Pak, G. A. Voth, *Curr. Opin. Struct. Biol.* **2018**, *52*, 119.
- [6] H. Gohlke, M. F. Thorpe, *Biophys. J.* **2006**, *91*, 2115.
- [7] Z. Zhang, K. Y. Sanbonmatsu, G. A. Voth, *J. Am. Chem. Soc.* **2011**, *133*, 16828.
- [8] M. G. Saunders, G. A. Voth, *Annu. Rev. Biophys.* **2013**, *42*, 73.
- [9] W. G. Noid, *J. Chem. Phys.* **2013**, *139*, 090901.
- [10] N. J. Dunn, W. G. Noid, *J. Chem. Phys.* **2015**, *143*, 243148.
- [11] Y. Miao, P. Ortoleva, *J. Chem. Phys.* **2006**, *125*, 44901.
- [12] S. Pankavich, Y. Miao, J. Ortoleva, Z. Shreif, P. Ortoleva, *J. Chem. Phys.* **2008**, *128*, 234908.
- [13] J. W. Chu, S. Izveko, G. A. Voth, *Mol. Simul.* **2006**, *32*, 211.
- [14] J. W. Chu, G. A. Voth, *Biophys. J.* **2006**, *90*, 1572.
- [15] S. Izvekov, G. A. Voth, *J. Chem. Phys.* **2005**, *123*, 134105.
- [16] M. Li, J. Z. Zhang, F. Xia, *J. Chem. Theory Comput.* **2016**, *12*, 2091.
- [17] Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfaendtner, G. A. Voth, *Biophys. J.* **2008**, *95*, 5073.
- [18] Z. Zhang, G. A. Voth, *J. Chem. Theory Comput.* **2010**, *6*, 2990.
- [19] M. S. Shell, *J. Chem. Phys.* **2008**, *129*, 144108.
- [20] A. Lyubartsev, A. Mirzoev, L. Chen, A. Laaksonen, *Faraday Discuss.* **2010**, *144*, 43.
- [21] A. Savelyev, G. A. Papoian, *J. Phys. Chem. B* **2009**, *113*, 7785.
- [22] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. de Vries, *J. Phys. Chem. B* **2007**, *111*, 7812.
- [23] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, S. J. Marrink, *J. Chem. Theory Comput.* **2008**, *4*, 819.
- [24] J. J. Uusitalo, H. I. Ingolfsson, P. Akhshi, D. P. Tieleman, S. J. Marrink, *J. Chem. Theory Comput.* **2015**, *11*, 3932.
- [25] J. F. Dama, A. V. Sinititskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, G. A. Voth, *J. Chem. Theory Comput.* **2013**, *9*, 2466.
- [26] A. Davtyan, J. F. Dama, A. V. Sinititskiy, G. A. Voth, *J. Chem. Theory Comput.* **2014**, *10*, 5265.
- [27] J. F. Dama, J. Jin, G. A. Voth, *J. Chem. Theory Comput.* **2017**, *13*, 1010.
- [28] Y. Zhang, Z. Cao, F. Xia, *Chem. Phys. Lett.* **2017**, *681*, 1.
- [29] Y. Zhang, K. Xia, Z. Cao, F. Grater, F. Xia, *Phys. Chem. Chem. Phys.* **2019**, *21*, 9720.
- [30] Y. Zhang, Z. Cao, J. Z. Zhang, F. Xia, *J. Chem. Inf. Model.* **2017**, *57*, 214.
- [31] S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi, *Science* **1983**, *220*, 671.
- [32] S. S. Petrova, A. D. Solov'ev, *Hist. Math.* **1997**, *24*, 361.
- [33] M. Li, J. Z. Zhang, F. Xia, *J. Comput. Chem.* **2016**, *37*, 795.
- [34] P. Koehl, F. Poitevin, R. Navaza, M. Delarue, *J. Chem. Theory Comput.* **2017**, *13*, 1424.
- [35] P. Koehl, *J. Chem. Theory Comput.* **2018**, *14*, 3903.
- [36] Y. L. Chen, M. Habeck, *PLoS One* **2017**, *12*, e0183057.
- [37] Y. Lecun, *Proc. IEEE* **1998**, *86*, 2278.
- [38] J. A. Hartigan, M. A. W. J. Roy, *Stat. Soc. Ser. C. (Appl. Stat.)* **1979**, *28*, 100.
- [39] W. Kuhlbrandt, *eLife* **2014**, *3*, e03678.
- [40] A. V. Sinititskiy, M. G. Saunders, G. A. Voth, *J. Phys. Chem. B* **2012**, *116*, 8363.
- [41] T. Caliński, J. Harabasz, *Commun. Stat.* **1974**, *3*, 1.
- [42] P. J. Rousseeuw, *J. Comput. Appl. Math.* **1987**, *20*, 53.
- [43] E. Rendón; I. Abundez, A. Arizmendi, E. M. Quiroz, *Int. J. Comput. Commun.* **2011**, *5*, 27.
- [44] M. Malumbres, M. Barbacid, *Nat. Rev. Cancer* **2003**, *3*, 459.
- [45] H. R. Kalbitzer, M. Spoerner, P. Ganser, C. Hozsa, W. Kremer, *J. Am. Chem. Soc.* **2009**, *131*, 16714.
- [46] Y. Li, Y. Zhang, F. Grosseruschkamp, S. Stephan, Q. Cui, C. Kötting, F. Xia, K. Gerwert, *J. Phys. Chem. Lett.* **2018**, *9*, 1312.
- [47] F. Xia, T. Rudack, Q. Cui, C. Kötting, K. Gerwert, *J. Am. Chem. Soc.* **2012**, *134*, 20041.
- [48] I. C. Rosnizeck, T. Graf, M. Spoerner, J. Tränkle, D. Filchtinski, C. Herrmann, L. Gremer, I. R. Vetter, A. Wittinghofer, B. König, H. R. Kalbitzer, *Angew. Chem. Int. Ed.* **2010**, *49*, 3830.
- [49] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696.
- [50] D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, H. Gohlke, A. W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. J. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York and P. A. Kollman, 2018, AMBER 2018, University of California, San Francisco.
- [51] K. Scheffzek, M. R. Ahmadian, W. Kabsch, L. Wiesmüller, A. Lautwein, F. Schmitz, A. Wittinghofer, *Science* **1997**, *277*, 333.
- [52] M. B. Berry, B. Meador, T. Bilderback, P. Liang, M. Glaser, G. N. Phillips, Jr., *Proteins: Struct. Funct. Bioinf.* **1994**, *19*, 183.

Received: 9 July 2019

Revised: 15 August 2019

Accepted: 27 August 2019

Published online on 30 September 2019