공학석사학위논문

단백질 동역학 해석을 위한 유한요소 모델 생성 기술

A Study on Constructing Finite Element Models for Protein Dynamics

2016년 7월

서울대학교 대학원

기계항공공학부

윤 기 석

단백질 동역학 해석을 위한 유한요소 모델 생성 기술

A Study on Constructing Finite Element Models for Protein Dynamics

지도교수 김 도 년

이 논문을 공학석사 학위논문으로 제출함

2016년 7월

서울대학교 대학원

기계항공공학부

윤 기 석

윤기석의 공학석사 학위논문을 인준함

2016년 6월

위 원 장 _____

부위원장 _____

위 원 _____

Abstract

Proteins are essential working unit in life's body whose functional characteristics are majorly determined by those dynamic properties. Conventional all-atom based simulation methods, such as molecular dynamics (MD) simulation or all-atom normal mode analysis (ATM-NMA) provide important insights into dynamic behaviors of proteins and their functions but those method are frequently limited due to computational complexity. As an alternative, finite element (FE) method can be used for coarse-grained simulation. Finite element method provides an efficient way of analyzing proteins' dynamics related to their biological functions. The method has been successful in obtaining quite accurate results with reasonable computational time.

Although there are various approaches to generating a finite element protein model for protein dynamics, it has not been critically investigated whether and how these different approaches affect the predictions. Various approaches in generating FE protein model, including different types of protein-surface definitions and those parameters, are critically investigated in this study to present the robustness of FE-based protein models. Then we provide a guideline on the construction and the use of these FE models.

Keywords : Finite element analysis, normal mode analysis, molecular surface, mean-square fluctuation, ATM-NMA, ENM

Student Number : 2014-22493

Table of Contents

List of Figures

List of Tables

Nomenclature

Symbols

$\rho_{Pearson}$	Pearson correlation coefficient
X_i, Y_i	MSF amplitudes at atom i
μ_X, μ_Y	Mean values of MSF
O_k	Mode overlap in mode k
φ_k, ψ_k	Eigenvectors of mode k
$\rho_{1.5}$	Correlation coefficient when 1.5Å probe is used
$\rho_{7.5}$	Correlation coefficient when 7.5Å probe is used
ρ_{var}	Correlation coefficient when the material properties are varied
ρ_{const}	Correlation coefficient when the material properties are constant

Abbreviations

MD	Molecular dynamics
ATM-NMA	All-atom normal mode analysis
NMA	Normal mode analysis
FE	Finite element
ENM	Elastic network model
ANM	Anisotropic network model
SES	Solvent-excluded surface
EDCS	Electron density contour surface
MSF	Mean-square fluctuation

1. Introduction

Since the conformational dynamics of proteins plays an important role in determining their biological functions in life's body, a lot of efforts, such as molecular dynamics (MD) simulation, all-atom normal mode analysis, elastic network model, and finite element protein model, have been made to predict the conformational dynamics of proteins and to finally understand their functions.[1]

Molecular dynamics (MD) simulation is a classical and a widely-used method to accurately predict the time-dependent trajectories of atoms by solving equations of motion based on full interatomic potential function.[2,3,4,5] Although the method provides the solutions in quite a detailed manner, its demand for heavy calculation confines to analyzing only the small systems with the limited size of molecules and the limited time-scale. Instead, researchers concentrated on searching only the possible dominant motions by performing a normal mode analysis (NMA) perceiving that only the large collective motions of proteins participate in actual functioning. All-atom normal mode analysis (ATM-NMA) is a representative NMA method which provides the several lowest vibrating modes of a protein structure by solving an eigenvalue problem of a Hessian matrix extracted from the energy-minimized structure with full-interatomic potential associated.[6, 7, 8] Although the method effectively reduces the amount of calculation compared to the MD simulation, ATM-NMA is still computationally burdensome since it requires laborious energy-minimization process as well as extracting a Hessian matrix, the second derivative of the full potential field functional. For this reason, numerous coarse-grained models are suggested to predict the dynamics of proteins. Tirion had suggested that a simplified single-parameter potential could represent the

biologically important motions of biomolecules.[9] Instead of considering full interatomic potential function, only a single parameter is used to represent the interactions between atoms to calculate the normal modes of a protein structure. Despite its large simplification, the method was quite successful in predicting the large, slow motions of a protein. Later, even more simplified method is suggested by taking only representative atoms into account in calculation.[10, 11 12] These methods are commonly known as elastic network model (ENM) meaning that the atom pairs are considered to be inter-connected by virtual springs with an adjustable spring constant. ENM is also well known to provide prominent results with predominantly reduced computational cost. For this advantage, many researchers have proposed variations of ENM to improve its accuracy asnd applicability.

On the other hand, finite element (FE) approach, which is of major interest in this study, had also been suggested as a prominent coarse-grained model. The method treats a protein as a homogenous isotropic elastic material encapsulated by a molecular surface representing the protein, and computes the normal modes of the protein structure based on the theory of finite element method (FEM).[13, 14, 15, 16, 17] The FE method also provides prominent results with great efficiency. Since the FE model focuses only on the boundary shape of surface, the total degree of freedom can be determined with reasonable range regardless of the number of atoms in the protein structure. In fact, there have been several attempts to adopt the FE method into biomolecule physics, such as identifying electrostatic characteristics or characterizing mechanical properties with simulating the nanoindentation experiments of biomolecules.[18, 19, 20]

The desirable advantage of FEM lies in its versatility and applicability. Since FE method is well established in classical engineering field, there are many chances to apply its computational techniques to a protein dynamics simulation for improvement in both accuracy and efficiency. Furthermore, since the FE method models the distinct surface of a protein, it is possible to consider the external effects such as viscous friction, electrostatics or other physics to understand its dynamics or functions more closely.

Despite these potential advantages, the finite element approach of protein dynamics has not been critically evaluated unlike other popular coarse grained method such as ENM. There are several ways and parameters to construct a finite element protein models, and these parameters can affect the final shape or properties of finite element models. However, it has not been critically investigated yet how different approaches and different parameters in constructing a FE model affect the prediction of dynamics of proteins. For this reason, we investigate several different datasets of FE protein models generated with different parameters to identify how the different parameters affect the results. We evaluate and compare result of the FE models with more precise method, the ATM-NMA, and provide a guide to construct a proper FE protein model.

2. Method

2.1 FE-based NMA

FE-based NMA consists of several steps including the construction of a molecular surface, the generation of a three-dimensional volumetric mesh, and the calculation of normal modes and natural frequencies (Figure 1).

2.1.1 Defining the molecular surface

It begins with defining a molecular surface of the target protein, typically either solvent-excluded surface (SES) or electron density contour surface (EDCS), that is used to build a three-dimensional FE model (Figure 2). SES is calculated by rolling a spherical probe, representing a solvent molecule, over the Van der Waals surface of the target protein structure.[21] We obtain the initial SES in this study using a freely available program, MSMS version 2.6.1[22] (http://mgltools.scripps.edu/packages/MSMS/), that calculates the analytic surface first and then generates a triangulated surface mesh from the input atomic coordinates (Figure 2B). The triangulated surface data consists of the vertex file containing the coordinates of the vertices and the face file containing the vertex indices related to the face connectivity. These data can be converted into STL file format which can be handled in various mesh-handling programs. On the other hand, EDCS is defined as a contour surface of the three-dimensional electron density map that can be determined experimentally or simulated from the atomic coordinates. Here, we build the simulated EDCS from the same atomic coordinates used for constructing the SES using an open source program, UCSF Chimera

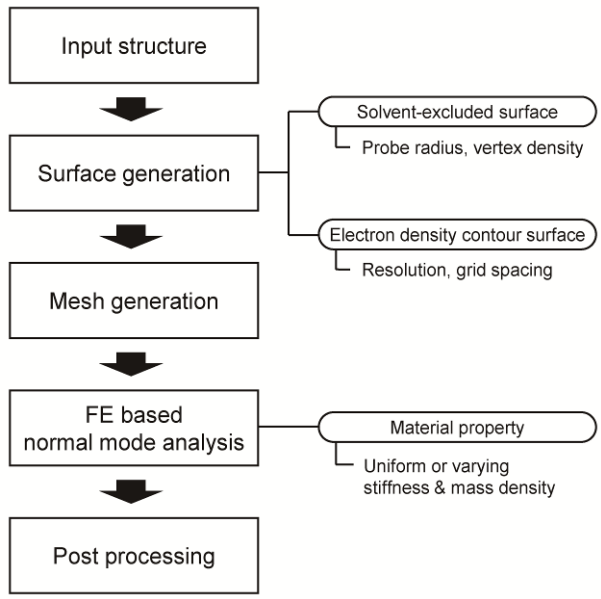

Figure 1. The process for generating finite element protein models. Each step contains several parameters that could affect the property of FE model. Two accustomed surface definitions, SES and EDCS, can be used in determining the surface of the protein structure, and each surface can be determined by several variables determining the final form of the surface. Mechanical properties such as Young's modulus and mass density can be assigned distinctively to each element in FE model. Number of modes to be extracted in NMA process also can be controlled.

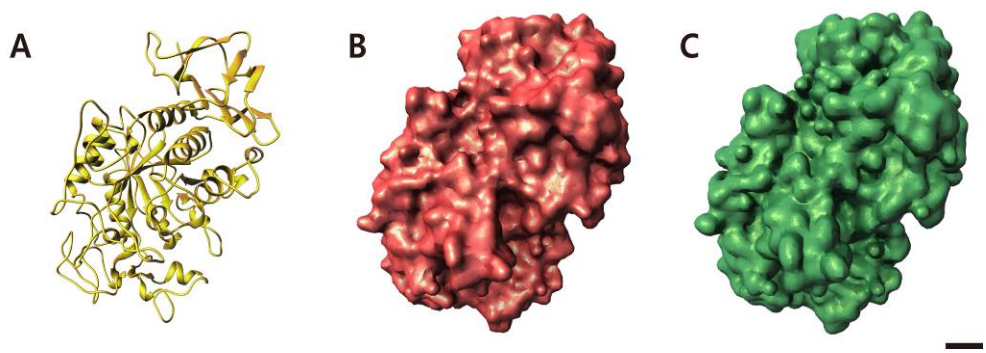

Figure 2. (A) Crystal structure of amylase (Protein Data Bank ID: 1SMD). (B) Solvent-excluded surface of amylase generated with the probe radius 1.5 Å and the vertex density 1.0. (C) Electron density contour surface of amylase generated with the resolution 5.0 Å. Scale bar represents 10 Å.

version 1.10.1[23] (http://www.cgl.ucsf.edu/chimera/). The command molmap is used to generate a density map from the PDB file. It computes the triangulated contour surface for the user-supplied resolution and contour level assuming Gaussian distribution of electron densities for each atom in three dimensions whose amplitude and width are proportional to the atomic number and the resolution, respectively (Figure 2C).

2.1.2 Mesh filtering steps

The initial molecular surfaces frequently contain several defects including self-intersection, non-manifolds, and isolated components that hinder the creation of clean and closed surfaces prerequisite to generate the volumetric mesh. Also, the initial molecular surface of supramolecular proteins often consists of excessively many faces deteriorating the computational efficiency. Hence, successive mesh cleaning and reduction are usually performed in practice using a carefully designed sequence of mesh filtering algorithms[15, 16] available in, for example, an open source program, MeshLab[24] (http://meshlab.sourceforge.net). Automated mesh cleaning procedure used in this study is illustrated in detail in the following paragraph.

The filtering processes start with “Merge close vertices” filter to merge all the vertices nearer than the preset threshold. The intersecting faces are removed by “Select self-intersecting faces” and “Delete” options. Likewise, faces attached to the non-manifold edges and non-manifold vertices are removed with “Select non-manifold edges”, “Select non-manifold vertices”, and “Delete” options.

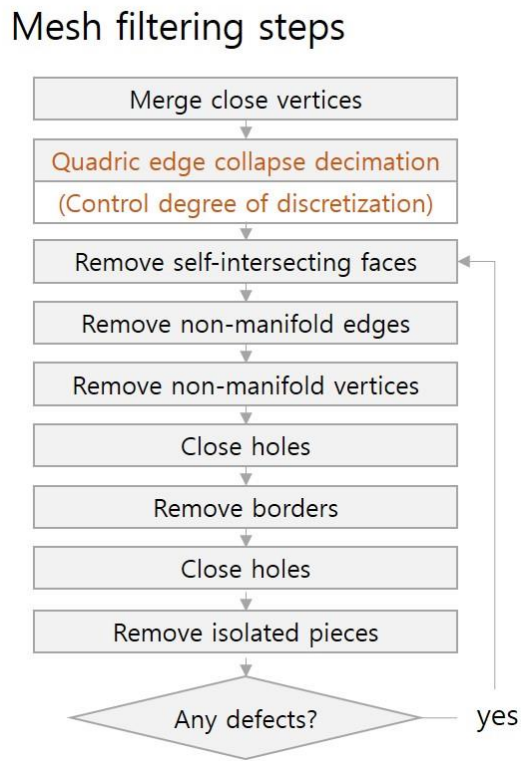

Figure 3. An automated procedure for repairing a molecular surface using several mesh filtering algorithms.

Consequently, the surface with holes where the defected faces removed is obtained. The surface with holes is to be filled with proper faces with “Close holes” option. Faces on the boundaries are examined, deleted, and again filled with “Select borders” and “Delete” options in case that the faces are not perfectly filled with “Close holes” option. Lastly, “Remove isolated pieces” filter is applied to eliminate isolated floating components. This filter searches for and eliminate the components whose number of connected faces are less than the user-set value. The value of 100 is used in this study, since the main component contains at least more than 1000 faces. These series of filtering process is repeated until the input surface meets the condition for proper FE model. The surface conditions for proper FE model are following: non-intersecting, manifold, single-component, and fully closed. In addition, “Quadric edge collapse decimation” filter can be added ahead if the size of the mesh is desired to be reduced. This filter reduces the number of faces as a percentage of the initial size. Since the sudden drop of the mesh size causes many defects, it is recommended to use the percentage value of 0.8 or higher and executed repeatedly until reaching the target number of faces.

2.1.3 FE model and normal mode analysis

The three-dimensional finite element model consisting of tetrahedrons is then constructed from the clean molecular surface using tessellation algorithms such as advancing front method[25, 26] and Delaunay triangulation[27, 28]. The mass and stiffness matrices are determined usually by assuming proteins as isotropic elastic materials described by the mass density, Young’s modulus and Poisson’s ratio. In general, the

mass density of 1.45 g/cm^3 corresponding to the mean density of proteins[29] and Poisson's ratio of 0.3 are used. Young's modulus serves as an adjustable parameter for fitting the mean-square fluctuation (MSF) amplitudes calculated using FE models to those obtained using ATM-NMA solutions or experimental Debye-Waller factors, which normally ranges from 2 to 5 GPa.[13, 15, 17] Normal mode analysis is finally performed using these FE mass and stiffness matrices to obtain a set of lowest normal modes and corresponding natural frequencies. Here, we use the commercial FE analysis software, ADINA version 9.0.1 (ADINA R&D, Inc., Watertown, MA, USA), to generate the three-dimensional FE models and calculate 200 lowest normal modes excluding six rigid-body modes.

Eigenvectors of alpha carbons are computed by interpolating the eigenvectors of nodes nearby. The process starts with setting up a linear interpolation function appropriate for a tetrahedral element, and searching for the tetrahedral element which contains the alpha carbon by using positional data of each atom. For each element containing the atom, 4 eigenvectors of belonging nodes are linearly interpolated depending on the distances to the atom and summed up finally to decide the eigenvector of associated atom.

2.2 Parametric study of FE models

We investigate the effect of FE model parameters on the solution accuracy by varying dominant parameter values systematically and comparing the results with ATM-NMA solutions. Since the level of approximation in constructing the molecular surface mainly determines the quality of FE models, we particularly focus on the parameters in the surface generation step.

When SES is used as the molecular surface, the probe radius is a primary parameter to be considered as it determines the analytical surface geometry. While it is typical to use 1.4 or 1.5 Å for it representing the size of a single water molecule, it is sometimes desirable to use a larger probe when we want to generate a smoother surface especially of large proteins[13]. We generate the SES-based FE models using several probe radii ranging from 1.0 to 7.5 Å and study their effect on the solution accuracy. Another important feature of the FE model is the level of discretization approximating the analytical molecular surface. The fineness of discretized molecular surfaces can be controlled using two parameters: the vertex density determining the number of vertices per unit area used when triangulating the analytical SES initially and the face reduction ratio used when coarsening the initial surface mesh further. We explore their effect by using three vertex densities (1.0, 2.0, and 3.0 vertex/Å^2) for models without any face reduction and four face reduction ratios (1/2, 1/4, 1/8, and 1/16) when coarsening the initial surface mesh constructed using the vertex density of 1.0 vertex/Å^2 via quadric edge collapse decimation[30] that reduces the number of triangles on the surface preserving its original shape as much as possible.

When simulated EDCS is used as the molecular surface, the resolution and the grid spacing are the primary parameters to be considered. Electron density of an atom is assumed to have three-dimensional Gaussian distribution centered at the atom position with its peak amplitude proportional to the atomic number and its standard deviation proportional to the resolution we choose. Here we generate the EDCS-based FE models using five different resolutions ranging from 4.0 Å to 8.0 Å. Calculated electron densities are generally stored discretely as voxels in a three-dimensional regular grid whose fineness is controlled by the grid spacing, another key parameter in building an EDCS. Iso-electron-density surfaces are often extracted using the marching cubes algorithm[31] that computes the triangulated contour surface for a given threshold density value. The grid spacing should be determined in conjunction with the resolution since a denser grid is required to achieve a fine resolution. We explore its effect by using four grid spacing options (resolution/5, resolution/3, resolution/2, and resolution/1) at each resolution.

It is conventional to use constant Young's modulus and mass density in FE-based NMA. However, these material properties may vary in space and hence affect the solution. To address this issue, we investigate the effect of spatial variation of the material properties on the solution using the EDCS-based FE models. In the pursuit of this purpose, we postulate three hypotheses on spatial variation of the material properties: (1) both Young's modulus and the mass density are constant, (2) Young's modulus is proportional to the electron density while the mass density is constant, and (3) the mass density is proportional to the electron density while Young's modulus is constant.

2.3 Comparison with ATM-NMA results

ATM-NMA results are used as reference solutions to evaluate the solution accuracy of FE models constructed using systematically varied parameter values. We adopt the protein structures and their ATM-NMA results available on an online database (http://www.cs.iastate.edu/~gsong/CSB/NMAdb) established by Na and Guang[32]. The database provides the eigenvectors and eigenvalues corresponding to 100 lowest normal modes of protein structures whose molecular weight ranges from 7.8 kDa to 19.8 kDa as well as their MSF amplitudes and energy-minimized structures.

The Pearson correlation coefficient and the mode overlap are used to quantify the similarity between the FE-based NMA and ATM-NMA solutions in MSF amplitudes and individual normal modes, respectively. Pearson correlation coefficient is defined as

$$\rho_{Pearson} = \frac{\sum_i \left(X_i - \mu_X\right) \cdot \left(Y_i - \mu_Y\right)}{\sqrt{\sum_i \left(X_i - \mu_X\right)^2 \sum_i \left(Y_i - \mu_Y\right)^2}} \tag{1}$$

where X_i and Y_i indicate the MSF amplitudes at atom i with their mean values, μ_X and μ_Y, obtained using the FE-based NMA and ATM-NMA solutions, respectively.

Mode overlap between two different models corresponding to mode k is defined as

$$O_k = 0.5\left(\max_{k-3 \le l \le k+3} \varphi_k \cdot \psi_l + \max_{l-3 \le k \le l+3} \varphi_l \cdot \psi_k \right) \quad (2)$$

where φ_k and φ_l denote the eigenvectors of mode k and mode l, respectively, computed using FE-based NMA while ψ_k and ψ_l represent the eigenvectors of mode k and mode l, respectively, computed using ATM-NMA.[33] Three nearby modes are included to consider a potential mode switching when the modes are closely spaced in their frequencies.

3. Results and Discussion

Here, we present the results for parametric study of the FE-based models constructed for 135 protein structures investigated by Na and Guang[34] and available in the ATM-NMA database. First, the sensitivity of the solution accuracy on the model parameters is discussed for SES-based models. Our investigation through various parameters, including probe radius and degree of discretization, shows the robustness of the solution accuracy of SES-based FE model, yet for some care must be taken for certain proteins with characteristic shapes. Next, we examine the sensitivity of the solution accuracy of EDCS-based FE model depending on model parameters such as the resolution and the grid spacing. Especially, we investigate the effect of applying various assumptions on material properties of protein models in addition to the conventional hypothesis of constant mass density and constant Young's modulus. Finally, we compare these SES-based and EDCS based FE models with respect to each other in model generation and solution accuracy, and provide a guideline on the construction and the use of these FE models.

3.1 SES-based FE model

First, we construct seven SES-based FE models for each protein in the database using various probe radii ranging from 1.0 Å to 7.5 Å and a default vertex density of 1.0 vertex/Å^2 without any face reduction. The correlation coefficients between the MSF amplitudes obtained using ATM-NMA and FE-based NMA averaged over the proteins remain on a similar level between 0.80 and 0.86 while decreasing slightly with the probe size (Table 1 and Figure 4a). The solution accuracy of these FE models is higher than the one of more commonly used ENM whose average correlation coefficient with ATM-NMA is 0.77 when elastic linear springs are connected between alpha carbons only using the cutoff distance of 13 Å. Similar insensitivity of the solution to the probe radius is also observed in the mode overlap profile representing the accuracy of individual normal modes (Figure 4b). Monotonic decrease of mode overlap values with the mode number while retaining relatively high values at low frequencies is a common characteristic of coarse-grained models. While FE-based NMA solutions are comparable to ATM-NMA ones for most protein structures, we can observe significantly low correlations, less than 0.50, for few protein structures (Figure 5). Unusually high peaks, however, exist in the MSF profile of these structures computed using ATM-NMA, which neither appears in FE solutions nor ENM results. This seems an

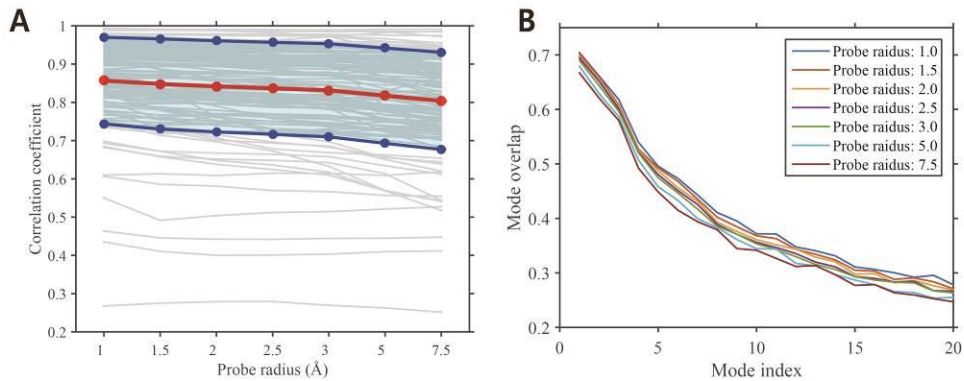

Figure 4. (A) Average correlation coefficients depending on probe radius. The red solid line represents the means and the blue solid line represents the standard deviations. (B) Mode overlap corresponding to several lowest modes with varying probe radius. Overlaps are determined by comparing each mode computed from FE analysis and ATM-NMA. Each solid line represents the averaged mode overlap over 135 proteins in each dataset of different probe radius.

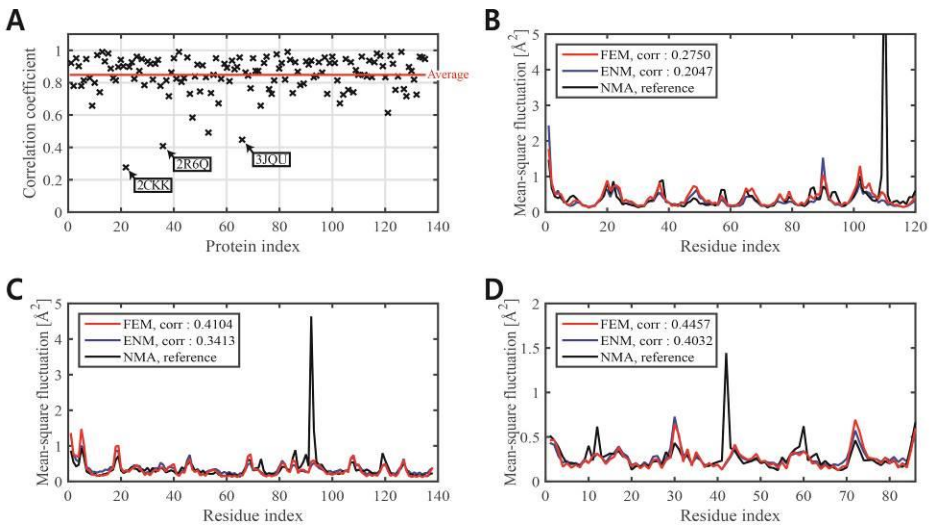

Figure 5. (A) The distribution of correlation coefficients between FEM and NMA. The probe radius 1.5Å and the vertex density 1.0 are used for FEM with no face reduction. Three proteins which show lowest correlation coefficients are labeled. MSF profiles of (B) 2CKK, (C) 2R6Q, and (D) 3JQU obtained by FEM, ENM, and NMA.

Probe radius [Å]	1.0	1.5	2.0	2.5	3.0	5.0	7.5	ANM
Correlation coefficient	0.8569	0.8481	0.8420	0.8367	0.8316	0.8174	0.8038	0.7728
Standard deviation	0.1132	0.1179	0.1192	0.1203	0.1218	0.1245	0.1271	0.1431

Table 1. Correlation coefficients depending on varying probe radius. The correlation coefficients are computed between MSFs obtained from FE analysis and ATM-NMA, and the values are averaged over 135 proteins in each dataset of different probe radius. Correlation coefficient computed between ANM and ATM-NMA is also stated for additional information.

artifact of ATM-NMA. If we exclude these extraordinary peaks, correlation coefficients rise to their normal values higher than 0.73.

Insensitivity of FE-based NMA solutions to the probe size results from the fact that overall three-dimensional shape of the protein rather than its structural detail dominates the low frequency modes contributing mostly to the conformational dynamics. This flexibility of choosing the probe size can be useful in modeling high-molecular-weight proteins for which the generation of the molecular surface is often challenging. As the protein size increases, the number of reentrant surfaces formed when the probe is in contact with multiple Van der Waals spheres increases enormously, leading frequently to a considerable amount of surface defects on the molecular surface including self-intersecting triangles and non-manifolds that hinder the creation of a clean and closed surface mesh. If we use a bigger probe in this case, a smoother molecular surface can be obtained without surface defects. Even though we may lose certain structural features such as gaps and cavities in the model, they usually do not affect the solution as long as their characteristic length is much smaller than the overall size of the protein. Viral capsids are such high-molecular-weight protein structures with a complex three-dimensional shape particularly due to a large number of spikes. For example, a clean molecular surface can be obtained using the probe radius of 5.0 Å for the capsid structure of a poliovirus (Protein Data Bank ID: 1DGI), which is not possible using smaller probes (Figure 6).

However, great care must be exercised when modeling small protein structures. Although the solutions are almost insensitive to the probe radius on average, a drastic drop in solution accuracy is observed for some proteins when

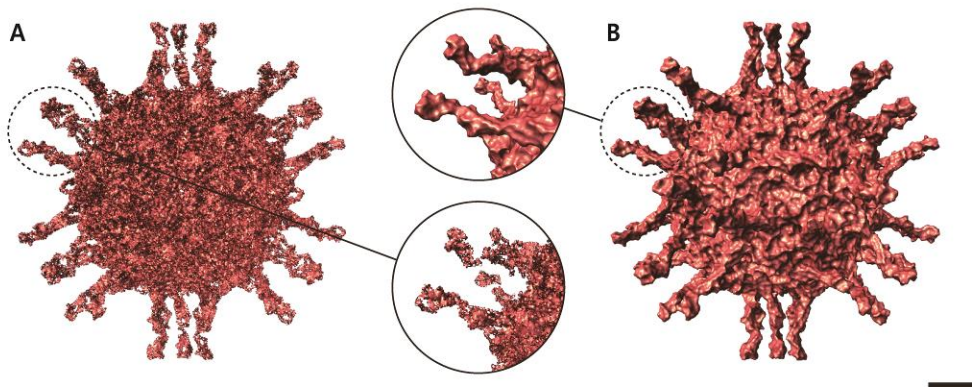

Figure 6. Solvent-excluded surface of poliovirus (Protein Data Bank ID: 1DGI) with (A) probe radius of 1.5 Å and (B) probe radius of 5.0 Å. The MSMS program used in generating surfaces fails to describe the whole structure of poliovirus with probe radius 1.5 Å leaving several parts unfilled and defected. In contrast, probe radius 5.0 Å generates a full watertight surface. Scale bar represents 50 Å.

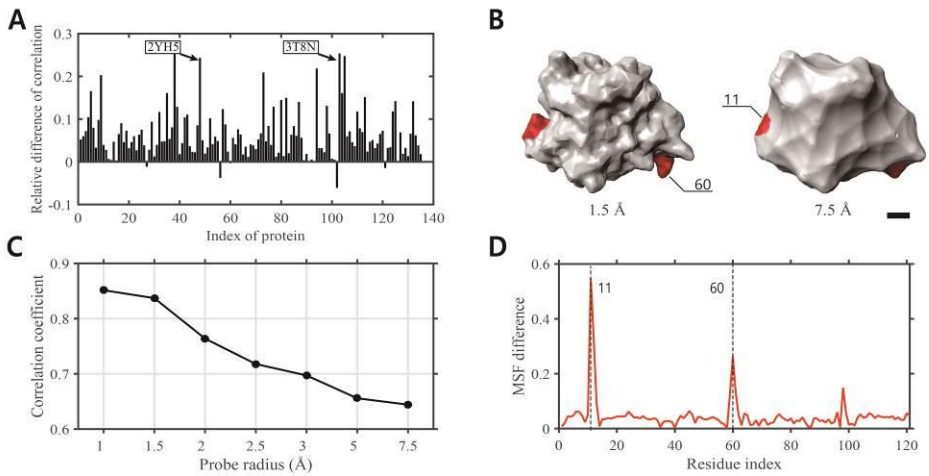

Figure 7. (A) Relative difference of correlation coefficients of SES models with 1.5 Å probe radius and 7.5 Å probe radius. The relative difference is computed as $(\rho_{1.5}/\rho_{7.5}-1)$. The positive value means that the 1.5 Å model is more accurate than the 7.5 Å model, and two specific proteins which show the largest difference are labeled. (B) SES of C-terminal domain of bamC protein (Protein Data Bank ID: 2YH5) modeled with probe radius 1.5 Å and 7.5 Å. (C) Correlation coefficient of 2YH5 depending on different probe radius. (D) Difference in MSFs of 1.5Å model and 7.5Å model. Residues where the peaks occur are marked on the figure (B) which may suffer from deterioration in MSF quality due to the change in probe radius. Scale bar represents 5.0 Å.

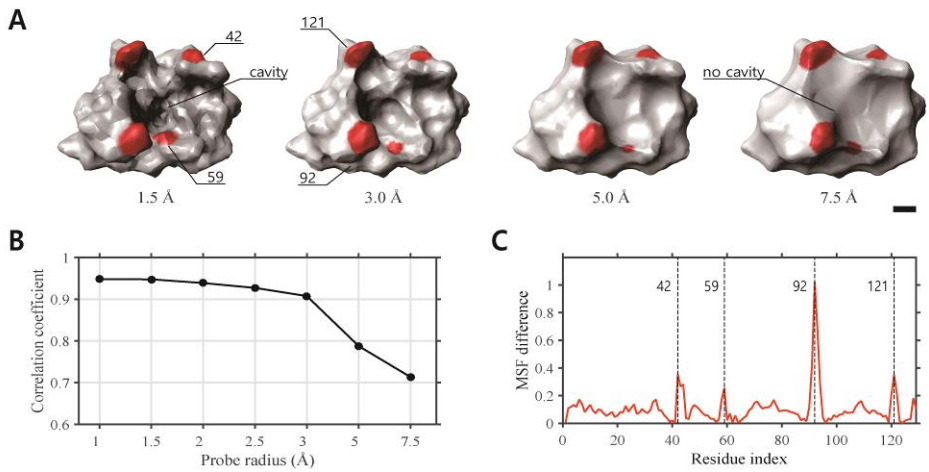

Figure 8. (A) SES of ketosteroid isomerase (Protein Data Bank ID:3T8N) modeled with probe radius 1.5 Å, 3.0 Å, 5.0 Å, and 7.5 Å. (B) Correlation coefficient of 3T8N depending on different probe radius. (C) Difference in MSFs of 1.5Å model and 7.5Å model. Residues where the peaks occur are marked on the figure (B) which may suffer from deterioration in MSF quality due to the change in probe radius. Scale bar represents 5.0 Å.

modeled as the probe becomes larger. For example, the correlation coefficient drops more than 20 percent for seven proteins when the probe radius of 7.5 Å is used instead of usual 1.5 Å (Figure 7a). These proteins contain certain characteristic features such as deep cavities or highly concave regions whose size is non-negligible compared to the overall protein size and that disappear from the model when built using a large probe, which deteriorates the solution accuracy significantly. For example, the C-terminal domain of bamC protein (Protein Data Bank ID: 2YH5) and ketosteroid isomerase (Protein Data Bank ID: 3T8N) are such proteins with those structural features (Figures 7B and 8A). The correlation coefficient goes down significantly from 0.85 to 0.65 for the C-terminal domain of bamC protein and from 0.95 to 0.71 for ketosteroid isomerase when the probe radius used for each model changes from 1.5 Å to 7.5 Å (Figures 7C and 8B). Considerable differences in MSF are observed at residues 11 and 60 for the C-terminal domain of bamC protein as they are located near highly concave regions whose molecular shape is sensitive to the probe size (Figure 7D). Similarly, in ketosteroid isomerase, the highest MSF difference between the models occurs at residues 42, 59, 92, and 121 that are adjacent to a deep cavity in the center (Figure 8C). This large cavity vanishes when the probe radius of 7.5 Å is used (Figure 8A), resulting in substantial reductions of MSF at its nearby residues.

In addition to the probe size, the vertex density and the face reduction ratio can further control the resolution of FE models. To investigate their effects on NMA results, we build seven FE models for each protein in the database using three vertex densities (1.0, 2.0, and 3.0 vertex/Å^2) without any face reduction and four face reduction ratios (1/2, 1/4, 1/8, and 1/16) with the vertex density of 1.0 vertex/Å^2. The probe radius of 1.5 Å is used for all FE models constructed for this

investigation. The average number of faces comprising SES ranges from 903 to 30,000, resulting in 3,322 to 227,432 degrees of freedom of FE models (Table 2).

Correlation coefficients and mode overlap plots clearly show that the results are almost insensitive to the vertex density and the face reduction ratio (Figure 9). This observation is somewhat counterintuitive because it is generally expected to obtain more accurate solutions with finer FE models in finite element analysis. Only a minor improvement in the mean correlation coefficients can be seen as the number of degrees of freedom increases (Table 2). This result arises also from the fact that the overall shape of proteins primarily determines the low-frequency normal modes. Hence, as long as the global shape is maintained, the fineness or coarseness of FE models hardly affects FE-based NMA solutions. This property is useful in analyzing large proteins as we may use a significantly coarsened FE model without substantial loss of solution accuracy. To illustrate, we analyze a crystal structure of chaperonin GroEL (Protein Data Bank ID: 1XCK) using several FE models whose number of degrees of freedom is considerably varying from 5,706 to 309,090 (Figure 10). The correlation coefficients between the finest, reference model and the other reduced models remain higher than 0.90 despite this drastic model reduction up to 98%.

However, it is noteworthy that the eigenvalues are dependent on the resolution of FE models (Figure 11). As the mesh size increases or the number of degrees of freedom decreases, the FE model becomes stiffer because the deformation is confined to a smaller conformational space, elevating eigenvalues in consequence if we use the same Young's modulus regardless of resolution (Figure 11A). In other words, the effective Young's modulus of a protein leading to the

highest correlation of FE-based NMA results with ATM-NMA ones decreases with the mesh size (Figure 11B). Therefore, it is highly recommended to use a sufficiently fine FE model if aiming to precisely calculate the effective Young's modulus of a protein. Otherwise, the use of coarse FE models would be sufficient and more efficient.

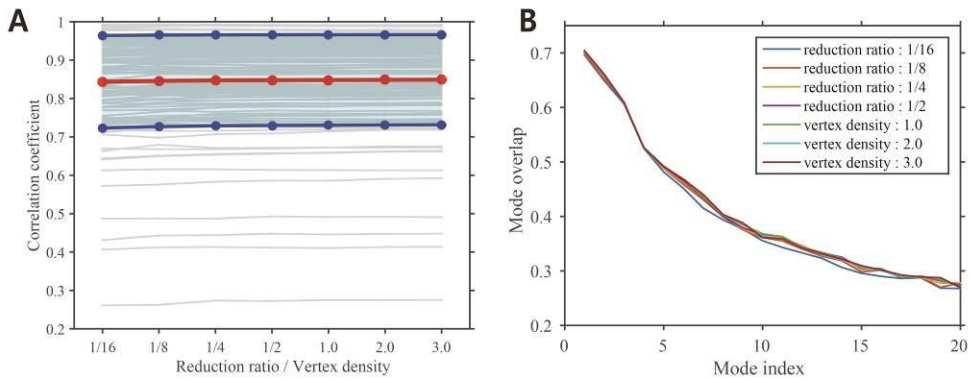

Figure 9. (A) Average correlation coefficients depending on reduction ratio and vertex density. The red solid line represents the means and the blue solid line represents the standard deviations. (B) Mode overlap corresponding to several lowest modes with different vertex density and reduction ratio. Each solid line represents the averaged values over 135 proteins in each dataset of different vertex density and reduction ratio. Overlaps are determined by comparing each mode computed from FE analysis and ATM-NMA.

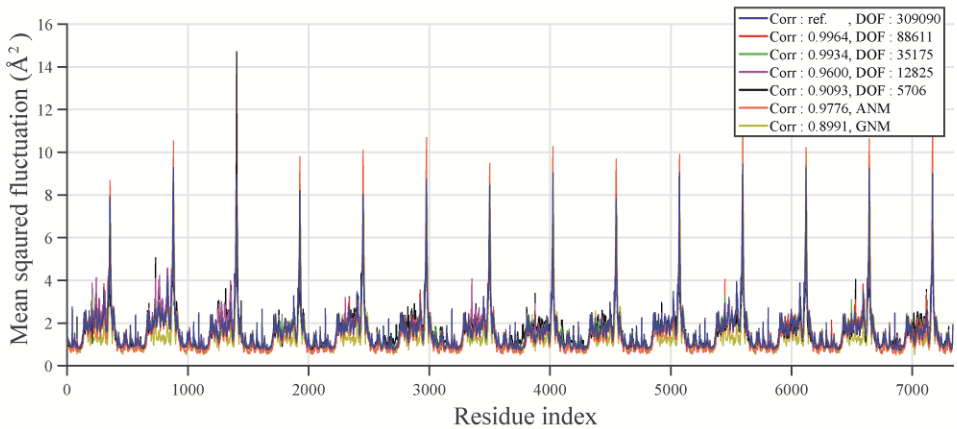

Figure 10. MSFs of chaperonin GroEL (Protein Data Bank ID: 1XCK) modeled with various degree of freedoms. The correlation coefficient of each model with respect to the most refined model is labeled. MSF obtained from ANM and GNM is plotted additionally.

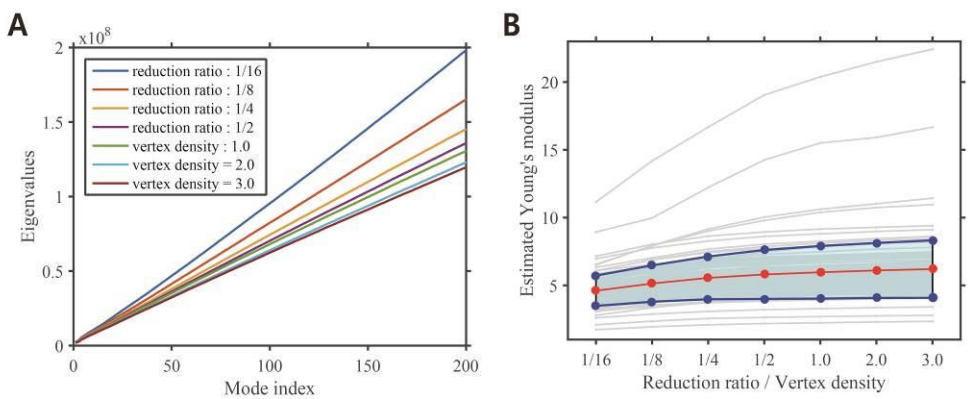

Figure 11. (A) Averaged eigenvalues corresponding to 200 lowest modes with different vertex density and reduction ratio. (B) Estimated Young's modulus with different vertex density and reduction ratio.

Parameters	Reduction ratio				Vertex density		
	1/16	1/8	1/4	1/2	1.0	2.0	3.0
Number of faces	903	1,782	3,666	7,167	14,769	21,621	30,000
Degree of freedom	3,322	8,078	20,671	43,815	100,748	152,842	227,432
Correlation coefficient	0.8436	0.8463	0.8478	0.8484	0.8486	0.8492	0.8495
Standard deviation	0.1205	0.1193	0.1184	0.1180	0.1179	0.1175	0.1173

Table 2. Numbers of faces, degrees of freedom, and Correlation coefficients depending on different reduction ratio and the vertex density. The number of triangular faces in the molecular surface and the degree of freedom of finite element protein model are averaged over 135 protein models. The correlation coefficients are computed between MSFs obtained from FE analysis and ATM-NMA, and the values are averaged over 135 proteins in each dataset of different degrees of freedom.

3.2 EDCS-based FE model

Here, we perform a parametric study for the EDCS-based FE model by constructing and analyzing the models using five resolutions every 1.0 Å from 4.0 Å to 8.0 Å and four grid spacing options (resolution/5, resolution/3, resolution/2, and resolution/1) at each resolution, leading to 20 FE models in total for each protein in the database. We exclude the effect of face reduction ratio explored for the SES-based FE model because the same conclusion can be drawn for the EDCS-based one as well.

Results clearly indicate that NMA solutions are almost insensitive to the resolution of simulated electron densities as long as a sufficiently fine grid is used (Figure 12 and Table 3). Only a slight decrease of the mean correlation coefficient and mode overlap is observed as lowering the resolution or increasing the grid spacing. This property offers great flexibility in constructing a computationally efficient FE model without hurting solution accuracy. If the grid spacing close to the selected resolution is used, however, we cannot vouch for the reliability of FE-based NMA solutions anymore. For example, individual curves of the correlation coefficient fluctuate widely rather than monotonically decrease with the resolution (Figure 12A). It is highly recommended to use grid spacing smaller than half the resolution of electron densities to obtain a reasonable NMA solution.

Some proteins experience a drastic drop in the correlation coefficient as the resolution increases. When the grid spacing is set to be one fifth the resolution, for

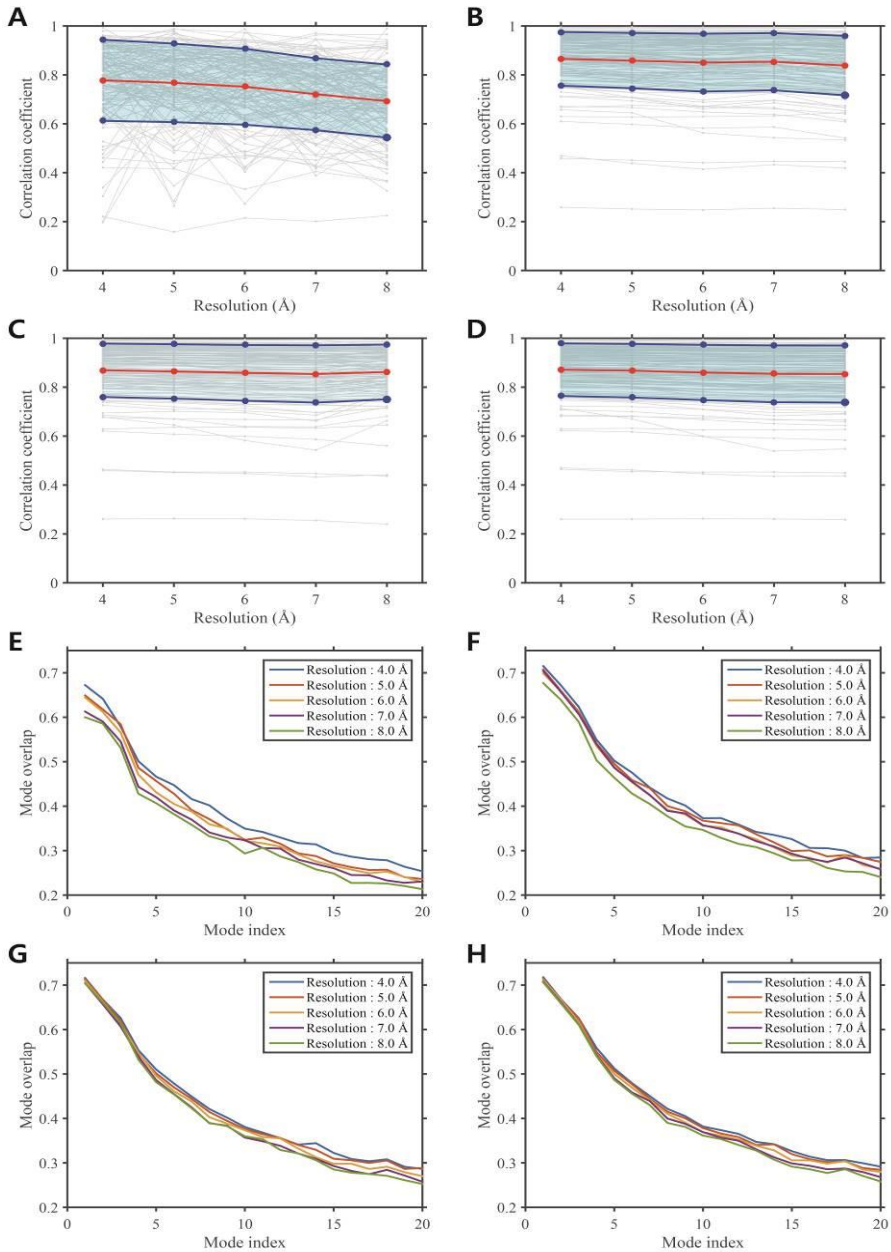

Figure 12. Correlation coefficients depending on the resolution with the grid spacing of (A) resolution/1, (B) resolution/2, (C) resolution/3, and (D) resolution/5. The red solid line represents the means and the blue solid line represents the standard deviations. The averaged mode overlap depending on the resolution with grid spacing of (A) resolution/1, (B) resolution/2, (C) resolution/3, and (D) resolution/5.

Grid spacing	Resolution				
	4Å	5Å	6Å	7Å	8Å
Resolution/5	0.8714	0.8675	0.8602	0.8546	0.8540
Resolution/3	0.8692	0.8644	0.8587	0.8541	0.8624
Resolution/2	0.8656	0.8586	0.8503	0.8437	0.8380
Resolution/1	0.7780	0.7674	0.7516	0.7210	0.6929

Table 3. Correlation coefficients depending on the grid spacing and the resolution. The correlation coefficients are computed between MSFs obtained from FE analysis and ATM-NMA, and the values are averaged over 135 proteins in each dataset of different degrees of freedom.

example, correlation coefficient of a certain protein is observed to be dropped from 0.71 to 0.55 which is about 0.23% loss (Figure 12 D). By closely observing the shape of this certain protein structure, we could derive the similar insight as the case when we had investigated the SES-based model with different probe radius. The protein contains large hollow region in the center of body which might be significant for the dynamics. This characteristic shape is well depicted when high resolution is used but disappears as lowering the resolution. Although the solutions are insensitive to the resolution in overall, a care must be taken when using low-resolution EDCS whether the characteristic shapes such as cavity and hollow region are properly depicted.

It is conventional to assume constant mass density and Young's modulus in FE-based NMA, but it is also possible to use spatially varying material properties. While it is not clear how to vary these properties in the SES-based FE model, we may hypothesize in the EDCS-based FE model that the material properties follow the electron density variation in space. To illustrate, we formulate and test three simple hypotheses on the spatial variation of the material properties: (1) Young's modulus is linearly proportional to the electron density while the mass density is constant, (2) the mass density is linearly proportional to the electron density while Young's modulus is constant, and (3) both Young's modulus and the mass density are linearly proportional to the electron density.

Results demonstrate that the use of varying Young's modulus can improve the solution accuracy while the use of varying mass density has a negligible or negative effect on it (Figure 13 and Table 4). When the first hypothesis is used, 93 proteins out of 135 experience the increase of the correlation coefficient up to

15.4% (Figure 13A). If we use the second hypothesis, the change of the correlation coefficient is only 0.3% drop on average, 1.8% increase at best, and 3.7% drop at worst (Figure 13B). The result when using the third hypothesis is similar to the one with the first hypothesis (Figure 13C). It is concluded that we may achieve an improvement of NMA solutions by using spatially varying material properties, but it may not be very significant if we take an additional computational effort into consideration. Therefore, the use of constant material properties is in practice a reasonably good assumption to use in FE-based NMA. This result also explains why conventional FE-based NMA and ENM-based NMA produce similar normal mode solutions even though they differ in mass and stiffness variation.

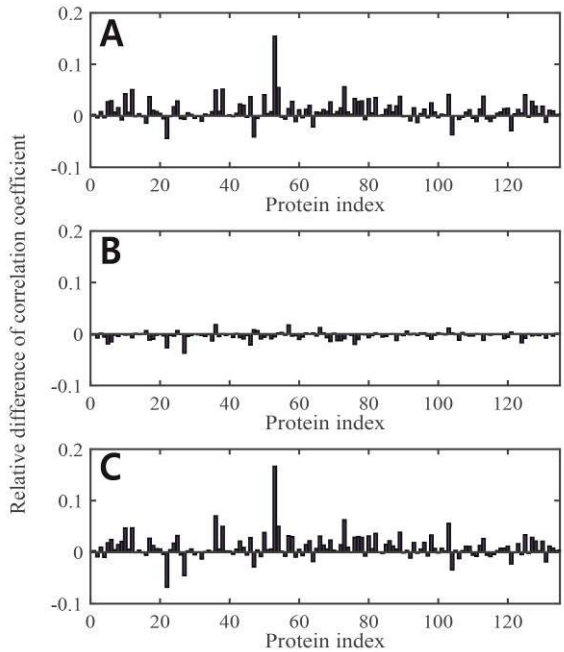

Figure 13. Relative difference of correlation coefficients of EDCS models with constant material properties and three different hypotheses on spatial variation of the material properties. (A) Young's modulus is proportional to the electron density and the mass density is constant. (B) Young's modulus is constant and the mass density is proportional to the electron density. (C) Both Young's modulus and the mass density are proportional to the electron density. The relative difference is computed as $\left(\rho_{\text{var}}/\rho_{\text{const}}-1\right)$, where ρ_{var} and ρ_{const} denote the correlation coefficient of FE models with varying material properties and FE models with constant material properties respectively.

Material properties	Both constant	Proportional Young's modulus	Proportional mass density	Both proportional
Correlation coefficient	0.8644	0.8721	0.8619	0.8720
Standard deviation	0.1112	0.1095	0.1124	0.1086

Table 4. Comparison of average correlation coefficients. The FE models with constant material properties and the ones with varying material properties are compared. Resolution 5.0Å and the grid spacing 1.67 are used in generating EDCS.

3.3 Comparison between SES-based and EDCS-based FE models

We have shown so far that both SES-based and EDCS-based FE models are quite robust to variation in model parameters and provide accurate NMA solutions close to ATM-NMA ones. More direct comparisons of two models reveal that the EDCS-based FE model exhibits better performance than the SES-based one. For instance, the EDCS-based FE models constructed using the resolution of 5.0 Å and the grid spacing of 1.7 Å (resolution/3) provide NMA solutions whose correlation coefficients are 1.9% higher on average than the ones obtained by the SES-based FE models generated using the probe radius of 1.5 Å and the vertex density of 1.0 vertex/$Å^2$ without face reduction (Figure 12). There are only 16 protein structures displaying lower correlation coefficient when the EDCS-based FE model is used. However, the decrease of correlation coefficients is less than 1.5% except for one structure whose MSF profile calculated using ATM-NMA possesses extraordinarily high peaks.

It is interesting that the correlation coefficient is improved by approximately 30% when analyzing a crystal structure of dynein light chain LC8 (Protein Data Bank ID: 3BRL) using the EDCS-based FE model (Figure 14A), which is mainly due to the highest difference in MSF amplitude at residue 64 (Figure 14B). The discrepancy of the molecular surface of two FE models near this residue explains the difference in their dynamics. A hollow region exists near residue 64 in the EDCS-based model (Figure 14D), but it is not shown in the SES-based model (Figure 14C). As a result, the dynamic motion of residue 64 becomes significantly

constrained by neighboring residues in the SES-based FE model.

This striking discrepancy originates from the fundamental difference in creating the molecular surface of two methods. SES consists of contact surfaces where the probe is in touch with the Van der Waals surfaces and reentrant surfaces that is a part of the probe surface toward the inside of a protein from contact lines (Figure 15A). While SES describes the atomic details of the molecular surface quite well when the probe actually touches the Van der Waals surfaces, it has limitation in describing concave or hollow regions with an entrance narrower than the probe diameter. In this case, the reentrant surface blocks the entrance and hence hollow regions disappear in the model. On the other hand, EDCS may describe the molecular surface in less detail, but concave or hollow regions can be better represented (Figure 15B) as it is solely generated from the electron densities of adjacent atoms superposed in space, which is not hindered by the narrowness of the entrance to hollow regions. Hence, the EDCS-based FE model seems better than the SES-based FE model particularly when analyzing a protein structure containing hollow regions with a narrow entrance as well as in general.

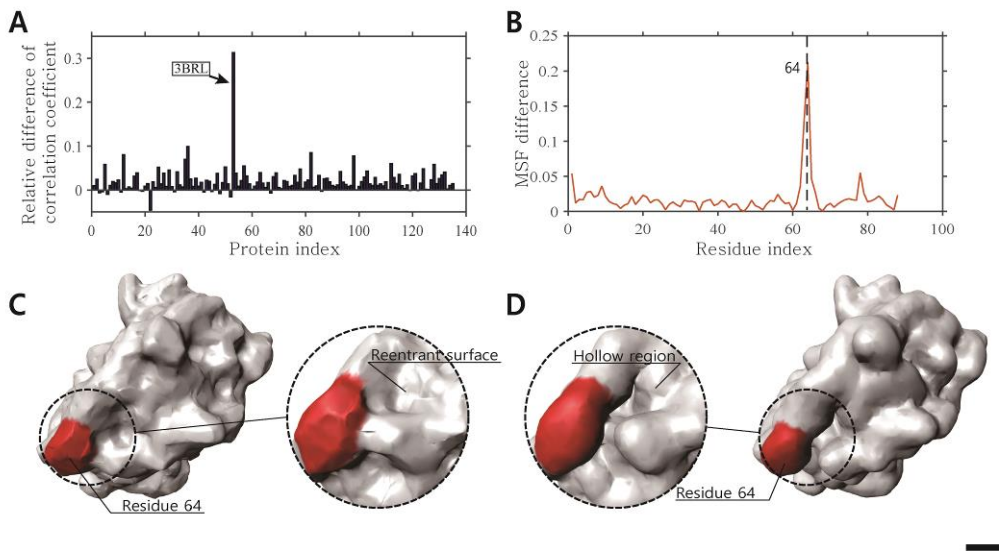

Figure 14. (A) Relative difference of correlation coefficients between SES models and EDCS models. The relative difference is computed as $\left(\rho_{EDCS}/\rho_{SES}-1\right)$. (B) Difference in MSFs of EDCS model and SES model. Residues where the peak occurs is colored red in figure (C) and (D). The LC8 S88E protein (Protein Data Bank ID: 3BRL) modeled with (C) SES and (D) EDCS. The difference of SES and EDCS are described in the figure. The Scale bar represents 5.0 Å.

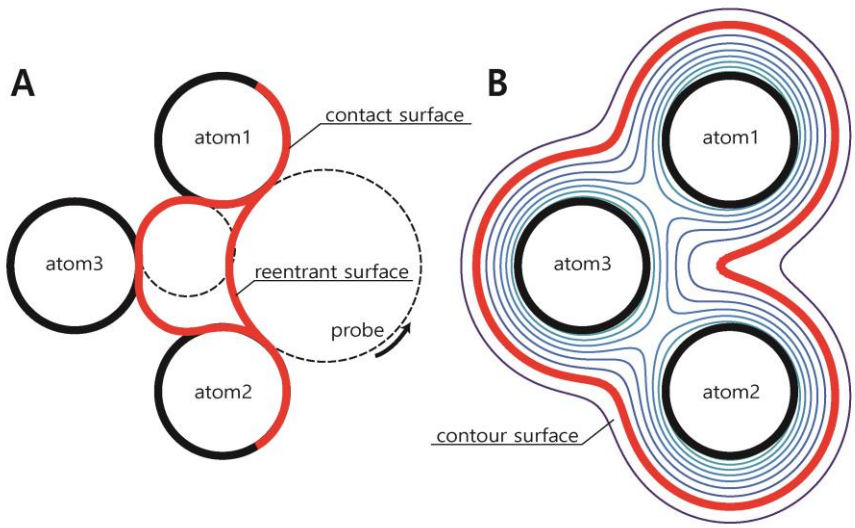

Figure 15. Comparison of surface definition. (A) Solvent excluded surface. (B) Electron density contour surface.

4. Conclusion

We present our effort to critically evaluate the performance of finite-element-based normal mode analysis for protein dynamics depending on various model parameters in generating the FE protein models. Two different types of surface definitions are used: a solvent-excluded surfaces and electron density contour surface. The probe radius, vertex density, and reduction ratio are the parameters used in computing SES. The resolution and the grid spacing are the parameters used in computing SES. As expected, the higher correlation are found in more sharp, distinct shapes (i.e. lower probe radius, lower resolution) and more discretized surface (i.e. higher vertex density, lower reduction ratio). Though the results show variance with input parameters, the correlation changes were generous having all results remain in tolerable range compared to ANM. However the change in the probe radius and resolution sometimes show large drop in correlation for some proteins, since those parameters affect the outer shape of the surface, especially when cavity or hole exist. Therefore, SES-based FE model has tolerance in using wide range of probe radius and vertex density (or face reduction) once the proteins' characteristic shapes are reasonable depicted. This might be useful in modeling proteins with various size and shapes.

On the other hand, EDCS has tolerance in the resolution once the enough grid spacing is provided which is usually finer than 'resolution/2'. Assigning the different material properties has positive effects on the solution accuracy, but the effect is not significant. In most cases, EDCS is more advantageous in protein dynamics than the SES since the EDCS depicts the characteristic shapes of protein structures better with less effort in generating the surface.

References

1. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. *Science* 1991;254(5038):1598-1603.

2. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. *Nature* 1977;267(5612):585-590.

3. Garcia AE, Harman JG. Simulations of CRP:$(cAMP)_2$ in noncrystalline environments show a subunit transition from the open to the closed conformation. *Protein Sci.* 1996;5(1):62-71.

4. De Groot BL, Hayward S, Van Aalten DMF, Amadei A, Berendsen HJC. Domain motions in bacteriophase T4 Lysozyme: a comparison between molecular dynamics and crystallographic data. *Proteins* 1998;31(2):116-127.

5. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.* 2002;9(9):646-652.

6. Brooks B, Karplus M. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci.* 1983;80:6571-6575.

7. Levitt M, Sander C, Stern PS. The normal modes of a protein: native bovine pancreatic trypsin inhibitor. *Int. J. Quantum Chem.* 1983;24:181-199.

8. Kitao A, Go N. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* 1999;9:164-169.

9. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* 1996;77:1905-1908.

10. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* 1997;2:173-181.

11. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 2001;80:505-515.

12. Kim MH, Kim MK. Review: elastic network model for protein structural dynamics. *JSM Enzymol. Protein Sci.* 2014;1:1001.

13. Bathe M. A finite element framework for computation of protein normal modes and mechanical response. *Proteins* 2008;70:1595-1609.

14. Kim DN, Altschuler J, Strong C, McGill G, Bathe M. Conformational dynamics data bank: a database for conformational dynamics of proteins and supramolecular protein assemblies. *Nucleic Acids Res.* 2011;39:D451-455.

15. Kim DN, Nguyen CT, Bathe M. Conformational dynamics of supramolecular protein assemblies. *J. Struct. Biol.* 2011;173:261-270.

16. Kim J, Kim JG, Yun G, Lee PS, Kim DN. Toward Modular analysis of supramolecular protein assemblies. *J. Chem. Theory Comput.* 2015;11:4260-4272.

17. Sedeh RS, Bathe M, Bathe KJ. The subspace iteration method in protein normal mode analysis. *J. Comput. Chem.* 2010;31:66-74.

18. Sakalli I, Schöberl J, Knapp EW. mFES: a robust molecular finite element solver for electrostatic energy computations. *J. Comput. Chem.* 2014;10:5095-5112.

19. Gibbons MM, Klug WS. Nonlinear finite-element analysis of nanoindentation of viral capsids. *Phys. Rev. E* 2007;75:031901.

20. Gibbons MM, Klug WS. Influence of nonuniform geometry on nanoindentation of viral capsids. *Biophys. J.* 2008;95:3640-3649.

21. Connolly ML. Analytical molecular surface calculation. *J. Appl. Cryst.* 1983;16:548-558.

22. Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* 1996;38:305-320.

23. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera─a visualization system for exploratory research and analysis. *J. Comput. Chem.* 2004;25:1605-1612.

24. Cignoni P, Callieri M. Corsini M. Dellepiane M, Ganovelli F, Ranzuglia G. Meshlab: an open-source mesh processing tool. In: Sixth Eurographics Italian Chapter Conference: Salereno, Italy, 2008;129-136.

25. Löhner R, Parikh P. Generation of three-dimensional unstructured grids by the advancing-front method. *Int. J. Numer. Meth. Fluids* 1988;8:1135-1149.

26. Peraire J, Peiro J, Formaggia L, Morgan K, Zienkiewicz OC. Finite element Euler computations in three dimesnions. *Int. J. Numer. Meth. Eng.* 1988;26:2135-2159.

27. Shenton DN, Cendes ZJ. Three-dimensional finite element mesh generation using Delaunay tessellation. *IEEE Trans. Magn.* 1985;21:2535-2538.

28. Watson DF. Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes. *Comput. J.* 1981;24:167-172.

29. Fischer H, Polikarpov I, Craievich AF. Average protein density is a molecular-weight-dependent function. *Protein Sci.* 2004;13:2825-2828.

30. Garland M, Heckbert PS. Surface simplification using quadric error metrics. In: Proceedings of the 24th annual conference on computer graphics and interactive techniques. Los Angeles, CA, August 1997;209-216.

31. Na H, Song G. Conventional NMA as a better standard for evaluating elastic network models. *Proteins* 2015;83:259-267.

32. Kondrashov DA, Van Wynsberghe AW, Bannen RM, Cui Q, Phillips GN Jr. Protein structural variation in computational models and crystallographic data.

Structure 2007;15:169-177.

33. Na H, Song G. The performance of fine-grained and coarse-grained elastic network models and its dependence on various factors. *Proteins* 2015;83:1273-1283.

Abstract (Korean)

단백질은 생명체가 기능을 함에 있어 중요한 역할을 하는 단위체로서, 그 기능은 주로 동적 특성에 의해 결정된다. 단백질의 기능을 이해하기 위해 동적 특성을 파악하는 것은 매우 중요하며, 때문에 동특성 해석을 위한 여러 시뮬레이션 기법들이 개발되어 있다. 분자동역학, 전원자 정규모드해석과 같은 고전적인 전원자 (all-atom) 기반 시뮬레이션은 단백질의 움직임을 상당히 정확하게 예측하여 단백질의 기능에 대한 중요한 정보를 제공한다. 하지만 이러한 전원자 기반 시뮬레이션 기법은 계산량이 상당하여 단백질의 종류에 따라 제한되는 경우가 많다. 이에 대한 대안으로 유한요소 기법을 사용할 수 있다. 유한요소 기법을 이용하면 모든 원자를 고려하지 않고 단백질의 형상만을 고려하기 때문에 효율적으로 단백질의 동적 특성을 파악할 수 있다. 이러한 계산적 축소에도 불구하고 상당한 정확도를 제공하기 때문에 상당히 응용 가능성이 높다. 하지만 아직 유한요소 모델을 어떻게 생성해야 하는가에 대한 고찰이 부족하다. 단백질의 유한요소 모델을 생성을 위한 접근법과 변수가 다양하게 존재하지만, 다양한 접근법이 결과에 어떠한 영향을 미치는가에 대한 분석은 미미한 실정이다. 본 연구에서는 분자 표면의 종류, 변수 등 다양한 접근법을 이용해 유한요소 모델을 생성하고, 이러한 변수가 결과에 어떠한 영향을 미치는지 분석, 고찰하여 적절한 단백질 유한요소 모델을 생성하는 방법을 제시하였다.