http://dx.doi.org/10.7566/JPSJ.82.114801

Local Hydrophobicity in Protein Secondary Structure Formation

Ming-Chya WU[1,2*] and Tian Yow TSONG[3]

[1]*Research Center for Adaptive Data Analysis, National Central University, Chungli 32001, Taiwan*
[2]*Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan*
[3]*College of Biological Sciences, University of Minnesota, Minneapolis, MN 55455, U.S.A.*

(Received July 10, 2013; accepted September 18, 2013; published online October 18, 2013)

In this paper, the effect of local hydrophobicity (LHP) in protein secondary structure formation is investigated using time series analysis approach. The LHP around a residue in a protein is defined as the sum of hydrophobicity (HP) of the surrounding residues within an effective range in a three-dimensional structure. HP and LHP as functions of the linear amino acid sequence are considered as time series, and are decomposed into a number of intrinsic mode functions (IMFs) using the empirical mode decomposition method. Correlation analysis of IMFs of HP and LHP of the wild-type (WT) proteins shows that the relative strength among IMF pairs is associated with the length scales of secondary structures. Examining the variations of secondary structures in mutants from the WT protein as a result of LHP changes, we propose that LHP is a useful parameter to describe secondary structure formation in proteins.

KEYWORDS: local hydrophobicity, empirical mode decomposition

1. Introduction

Proteins assume specified three-dimensional structures for biological activity and functional specificity from inter-and intra-molecular interactions. Among others, the hydrophobic interaction is an important component in stabilizing protein folded conformation.[1–6] It is defined as the free energy of transfer from water to a nonpolar liquid.[7] The ratio of the saturated concentrations $[X]$ of a molecule in the nonploar liquid and in the water at equilibrium gives the partition coefficient $K = [X]_{\text{water}}/[X]_{\text{nonpolar liquid}}$. The free energy transferred from the water to the nonploar liquid is given by $\Delta G = -RT \ln K$ with the gas constant R and the absolute temperature T, which is a measure of the hydrophobicity (HP) of the molecule. Because the HP values measured in various ways differ substantially, there are several representative scales.[8–10]

When a polypeptide folds, a residue in it experiences gradual changes in environment constructed by both the solvent and neighboring residues, rather than only water. Thus, the HP contributions from neighboring residues should also be taken into account. As the formation of hydrophobic core is essential for protein structure stability,[6] the concept of local hydrophobicity (LHP) has been introduced in a variety of folding models.[4,11,12] The LHP around a particular residue in a protein chain is defined as a sum of HP values of residues in its vicinity. Depending on the definition of vicinity, the estimations of LHP differ. Using the definition that the LHP h_i of the i-th residue is the sum of HP values of two nearest-neighbor residues on both sides of the sequence, i.e.,

$$h_i = \sum_{j=1,2} (s_{i-j} + s_{i+j}), \tag{1}$$

where s_j is the HP value of the j-th residue, Kanehisa and Tsong found that LHP stabilizes secondary structures in globular proteins.[4] However, the definition of LHP in Eq. (1) does not take into account the contributions of HP from other residues close to the i-th residue in three-dimensional space. In this paper, we revise the definition of Eq. (1) and define LHP as a sum of HP values of residues within an effective range r, as shown in Fig. 1. For a protein chain, the LHP σ_i of the i-th residue is given by

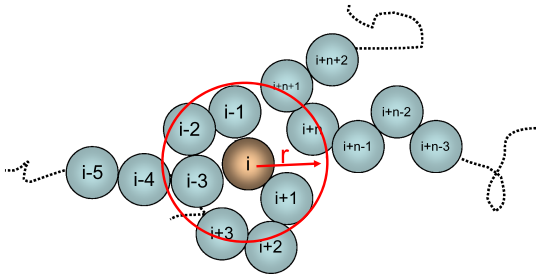

Fig. 1. (Color online) Schematic drawing for the definition of local hydrophobicity in this paper. One sphere represents one residue.

$$\sigma_i(r) = \sum_{\langle i,j\rangle_r} s_j, \tag{2}$$

where $\langle i,j\rangle_r$ indicates that the distance r_{ij} between pair residues i and j is within the effective range r. The residue $i = j$ is not included following the definition of the sequence-based LHP.[4] For simplicity, r is taken as $r = \alpha r_0$ with an integer α and $r_0 = 3.5$ Å is the averaged radius of 20 amino acids.[7] The configuration in Eq. (1) is included in Eq. (2) for $\alpha = 2$. Here we use $\alpha = 4$ to include more neighboring residues. In the context of LHP, Eq. (2) is more reasonable than Eq. (1) while the calculation of LHP in Eq. (2) can be implemented only when the three-dimensional structure of the protein is available.

To investigate the effect of LHP in secondary structure formation, we study the relation between LHP and distinct secondary structures of typical proteins based on the structure files released at the Protein Data Bank (PDB).[13] We first calculate HP and LHP of a protein and consider they were "time series". Such analogy is achieved by regarding the sequential residue number as "time" and the HP and LHP values as functions of the time sequence. In this way, sophisticated time series analysis approaches are applicable. The HP and LHP time series are then decomposed by the empirical mode decomposition (EMD)[14] into a number of

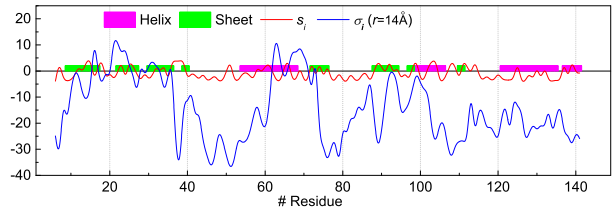

Fig. 2. (Color online) Hydrophobicity HP (s_i) and local hydrophobicity LHP (σ_i) vs residue number for SNase (PDB code 1EY0), estimated by the Kyte–Doolittle scale.[8] Elements of secondary structure are indicated by colored bands.

intrinsic mode functions (IMFs), in which each IMF is characterized by a distinct length scale, i.e., number of residues. We finally analyze the correlations between IMFs of HP and IMFs of LHP in the same length scale, and compare them with secondary structures in the protein. Based on this correlation, we propose a scheme to estimate secondary structure contents in mutants from wild-type (WT) proteins, and compare the results with experimental data of mutagenesis analysis.

The rest of this paper is organized as follows. In Sect. 2, the time series analysis approach to HP and LHP are introduced. In Sect. 3, we present the results of HP and LHP decompositions of a number of proteins, and estimations of secondary structure contents in these proteins based on the correlation analysis of LHP and secondary structures. The results are compared with experimental data. Finally, we conclude shortly in Sect. 4.

2. Methods

Let us first consider the HP and LHP of a typical protein, shown in Fig. 2. Here, the values of HP and LHP are estimated by the Kyte–Doolittle scale,[8] and we set $\alpha = 4$ in Eq. (2). HP and LHP as functions of residue sequential number are considered as a time series, denoted as $x(t)$. We assume that HP and LHP are composed of distinct components characterized by different length scales and certain components are associated secondary structure formation in the stages of nucleation, hierarchical assembly and stabilization of protein folding. To determine these components, we exploit the EMD method.[14]

2.1 *Decompositions of HP and LHP*

The EMD method has been developed on the assumption that any time series consists of simple intrinsic modes of oscillations. The adaptive decomposition scheme utilizes the actual time series to construct the decomposition base rather than decomposing it into a prescribed set of base functions. The decomposition is achieved by iterative "sifting" processes for extracting modes by identification of local extremes and subtraction of local means. The iterations are terminated by a criterion of convergence.[14] Under the procedures of EMD[14–20] a time series $x(t)$ is decomposed into n IMFs c_i's and a residue r_n,

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t). \tag{3}$$

Ideally, the "frequency" content of each IMF is not overlapped with others such that the characteristic frequencies of all components are distinct. In time domain, higher order modes have longer "period" in terms of the number of residues. This defines the components in different length scales. The IMFs are symmetric with respect to the local zero mean and have the same numbers of zero crossings and extremes, or a difference of 1. All the IMFs are orthogonal to each other for infinite long time series.[14] For data of finite length, the decomposition is *unique* by minimizing the orthogonality index κ

$$\kappa = \frac{\sum_{i\neq j,t} c_i(t)c_j(t)}{\sqrt{\sum_{i,t} c_i^2(t) \sum_{j,t} c_j^2(t)}}. \tag{4}$$

2.2 *Correlation between LHP and secondary structures*

The correlation between HP and LHP is defined as the inner product of two "vectors" (IMFs) c_k(HP) and c_k(LHP) in the same order k,

$$\eta(k) = \frac{c_k(\mathrm{HP}) \cdot c_k(\mathrm{LHP})}{|c_k(\mathrm{HP})||c_k(\mathrm{LHP})|}, \tag{5}$$

which is normalized between -1 to 1. In general, the correlation between HP and LHP is stronger in higher order k, due to the fact that the background HP does not change in folding, while the weaker correlation in lower modes suggests a subtle structure adaption in folding. A protein undergoes a self-arrangement such that the LHP calculated from its tertiary structure follows the folding information imprinted in its amino acid sequence.[21] The stronger correlation between certain IMFs of HP and LHP reveals that the length scales of these IMFs are associated with length scales of the secondary structures in a given protein. This inference can be verified from the correlation between these IMFs of LHP and secondary structures, and the correlation can be visually inspected from the profiles of the IMFs and distributions of the secondary structures (see Sect. 3 for details).

2.3 *Estimation of secondary structures in mutants*

Based on the correlation between LHP and secondary structures, we setup a quantitative description of the relation. We define the correlation coefficient ρ_i as the overlap between LHP σ_i, a function of HP $f(s_i)$, and a measure of secondary structures w_i for the i-th residue,

$$\rho_i = \sigma_i f(s_i) w_i. \tag{6}$$

The function $f(s_i)$ is introduced such that residues with similar LHPs can be in different secondary structures. However, the function form of $f(s_i)$ is inexplicit here and can be determined only when roles of all residues in the protein are known. For a given protein, the correlation coefficient ρ_i is fixed and the function $f(s_i)$ is roughly fixed for the proteins which are reversible in folding-unfolding. The folding of sub-domains of such mutants undergoes a similar nucleation as that of WT except the sub-domain where point mutated residues locate. We then use Eq. (6) to determine secondary structures of a mutant via evaluating σ_i and w_i of WT. The relationships among w_i's for different secondary structures can be estimated from the LHP for the first order approximation, i.e., all residues in a specified secondary structure have the same σ values. We calculate the average value of σ for different secondary structure using $\langle\sigma\rangle =$

$(\sum \sigma_s)/N_s$, with N_s the number of residues in a specified secondary structure. Under the first order approximation, there exists $\langle\sigma\rangle\langle w\rangle \approx$ const. and from which the relation among average values $\langle w\rangle$ of helix, sheet and coil can be obtained. This provides a route to determine changes of secondary structures from the tendency of changes of w (see Sect. 3 for details).

For some proteins that have similar $\langle\sigma\rangle$ values for secondary structures, the tendency of secondary structure changes cannot be determined from them directly. This usually happens for large proteins, e.g., WT Src Tyrosine Kinase (2SRC, 450 amino acid residues) has similar $\langle\sigma\rangle$ for helices and sheets (-0.02 for helices and -0.65 for sheets). For this case, one possible way to determine the tendency is using the values of correlated IMFs discussed in Sect. 2.2 instead of $\langle\sigma\rangle$. Further, we remark that though the relative values of $\langle\sigma\rangle$ are associated with the degree of exposure of helices and sheets to surface, they are not necessary equivalent due to the inhomogeneous hydrophobicity scale. For example, the helices and sheets of WT Prions (1QM2, 104 amino acid residues) have similar degree of exposure, while their $\langle\sigma\rangle$ values are distinct (-11.31 for helices and -4.38 for sheets). A sophisticated strategy to determine the tendency for various cases requires further investigations.

Next, under the second order approximation (i.e., different residues in a specified secondary structure have distinct σ values), for the j-th residue being mutated, we have

$$\sigma_i w_i f(s_i) \approx \sigma_i' w_i' f'(s_i), \tag{7}$$

for $i \neq j$, where the prime denotes that σ_i, w_i, and $f(s_i)$ are calculated for a mutant by replacing the C_α atom of the j-th residue in WT protein PDB file. By expanding $\sigma_i' = \sigma_i + \Delta\sigma_i$ and $w_i' = w_i + \Delta w_i$, we have

$$\frac{\Delta w_i}{w_i} \approx \frac{f(s_i)}{f'(s_i)} \frac{\sigma_i}{\sigma_i + \Delta\sigma_i} - 1. \tag{8}$$

We restrict our demonstration to single and double substitutions of residues such that $f(s_i) \approx f'(s_i)$ is valid for mutants following a similar early nucleation. Equation (8) then becomes

$$\frac{\Delta w_i}{w_i} \approx -\frac{\Delta\sigma_i}{\sigma_i + \Delta\sigma_i}. \tag{9}$$

σ_i and w_i are relative quantities for the non-normalized form of σ_i in Eq. (2). One can always work on positive σ_i and w_i by properly shifting their values. If $\Delta\sigma_i$ is positive, then Δw_i is negative and $w_i' < w_i$. For a specified protein, the increase of LHP σ_i at the i-th residue results in a structure change from a secondary structure to another and vice versa, depending on the residue being mutated and the tolerance for the change of LHP of the secondary structure. Consequently, this can serve a scheme to determine mutant secondary structure contents from WT PDB structure by analyzing LHP. The strategy is to use a few number of mutants to establish rules from LHP analysis and then use the rules to evaluate other mutants. The analysis can be based on experimental data or numerical simulations carried out in the same chemical condition.

3. Results and Discussions

In this section, we present the application of our methods to a number of classic proteins. Each protein is considered as an independent case study.

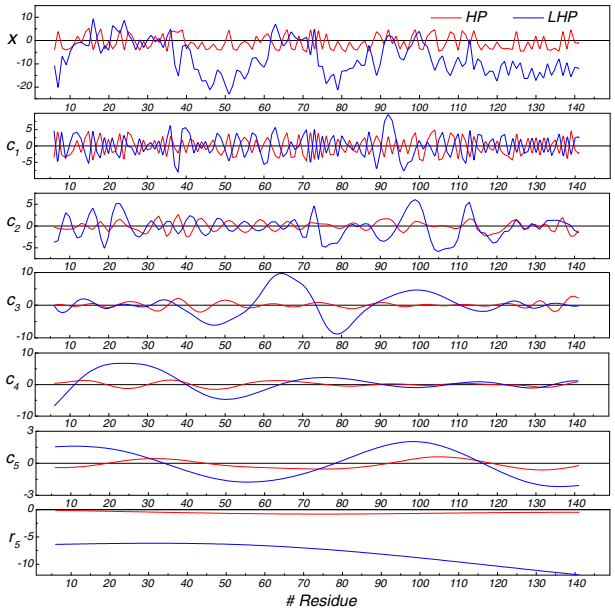

Fig. 3. (Color online) Empirical mode decomposition of the time series of hydrophobicity HP (s_i) and local hydrophobicity LHP (σ_i) ($r = 14$ Å) for SNase. Each time series is decomposed into 5 IMFs ($c_1, \ldots, c_5$) and 1 residue (r_5).

3.1 *Staphylococcal Nuclease*

Staphylococcal nuclease (SNase) is a small, globular protein (149 residues) with a single tryptophan at position 140. The dimension of the denatured states of SNase and some of its mutants is compact, no matter by acid or GdmCl denaturation.[22] The structure of WT SNase has been solved by both X-ray crystallography and nuclear magnetic resonance. Here we use the structure file 1EY0 for the analysis. Snase has 4 helices and 8 sheets. Because of the flexibility of N-terminus and C-terminus, only coordinates of middle 136 residues out of the total 149 residues were determined.

The decompositions of the time series of HP and LHP are shown in Fig. 3, in which each time series is decomposed into 5 IMFs ($c_1, \ldots, c_5$) and 1 residual r_5. The orthogonality indices of these decompositions are $\kappa = 0.0102$ for HP and 0.0022 for LHP. The correlation of IMFs from the strongest to the weakest is: $k = 5, 6, 2, 4, 1, 3$, as shown in Fig. 4. The residual r_5 ($k = 6$) is a trend of the time series. There is substantially no trend in HP, but a systematic decrease of LHP along the sequence can be observed in Fig. 3, because the LHP of the residues close to the N-terminus is in average higher than those around the C-terminus.

On the basis of length scale, the IMFs c_4 and c_5 are of special interest because they are associated with the number of residues in typical secondary structures. Meanwhile, the IMF c_2 accounts the contribution to LHP from two nearest-neighboring residues, so there is a stronger correlation with HP. Due to the fact that s_i of residue i is not involved in σ_i, the IMF c_1 is associated with HP from one nearest-neighboring residue such that there is one-position displacement in sequence and a weaker correlation between HP and

Table I. Percentage contents of secondary structures in SNase and its mutants estimated by PDB structures, the far-UV CD spectra and the analysis of local hydrophobicity LHP.

Structure	Wild-Type		W140A		F61W/W140A		Y93W/W140A		E75G	
	CD	PDB	CD	LHP	CD	LHP	CD	LHP	PDB	LHP
Helix (%)	23.1	28.9	6.7	13.4	15.6	18.8	12.9	15.4	28.9	27.5
Sheet (%)	23.2	26.8	52.4	45.6	29.0	30.2	38.4	43.0	26.8	28.2
Others (%)	53.7	44.3	40.9	40.9	55.4	51.0	48.7	41.6	44.3	44.3

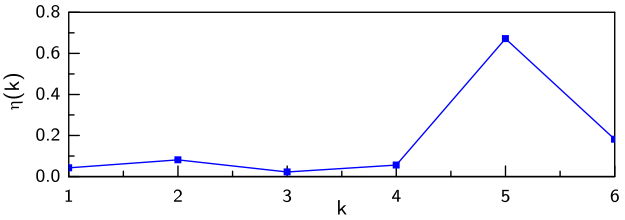

Fig. 4. (Color online) Correlation $\eta(k)$ between IMFs of HP and LHP in the same order k.

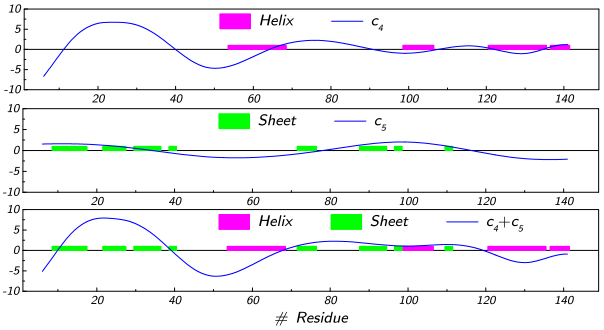

Fig. 5. (Color online) Secondary structures and IMFs c_4 and c_5 of local hydrophobicity LHP.

LHP. The weakest correlation in the IMF c_3 is a result of the balance of HP contributions in this length scale.

Figure 5 shows explicit correlations among profiles of IMFs c_4 and c_5 of LHP and the location of secondary structures in SNase. Here, helix structures have relatively lower LHPs and sheet structures have relatively higher LHPs. Thus, secondary structures are significantly correlated with the LHP calculated from the tertiary structure. We futher calculate the average values of LHP for secondary structures. The results are $\langle\sigma\rangle = -7.5$ for all residues, $\langle\sigma\rangle = -10.7$ for residues in coils and turns, $\langle\sigma\rangle = -8.4$ for helices, and $\langle\sigma\rangle = -2.3$ for sheets. According to $\langle\sigma\rangle\langle w\rangle \approx$ const., we have

$$\langle w\rangle(\text{turn or coil}) > \langle w\rangle(\text{helix}) > \langle w\rangle(\text{sheet}). \quad (10)$$

Next, we test the scheme proposed in Sect. 2.3 to estimate the secondary structure contents in mutants. We use the data from our mutation experiments carried out in the same chemical condition. Details of the experiments has been reported in a separated paper.[23] Table I lists the percentage contents of secondary structures for WT and mutants estimated from far-UV circular dichroism (CD) spectra. The secondary structure assignment of PDB data was verified by STRIDE.[24] The estimation by CD has a percentage error of at least 5% due to the non-independence of CD curves for pure secondary structures.

According to Eq. (9), we need information of 136 values of $\Delta w_i/w_i$. As an example of rough estimations, here we evaluate $-\Delta\sigma_i/|\langle\sigma\rangle|$. Because in the folding of SNase, N-terminal β core and C-terminal domain form first and then combine to form the final structure,[12] we analyze mutagenesis in two domains. LHPs of mutants are calculated using the WT structure with residue substitutions. We first calculate σ_i at each residue for WT and mutants, from which $\Delta\sigma_i$ for each residue was obtained. Next, we determine the tolerances of σ for different secondary structures by fitting $\Delta\sigma_i/|\langle\sigma\rangle|$ with F61W/W140A. The results are 5% of the average of LHP for sheets (i.e., $|\Delta\sigma|_s \sim 0.375$) and 25% for helices (i.e., $|\Delta\sigma|_h \sim 1.875$). Then, according to Eq. (10) and from the data of F61W/W140A, we derive the algorithms: (1) If $\Delta\sigma_i > |\Delta\sigma|_s$, the portion of helix at the i-th residue becomes a sheet while sheet structure remains a sheet. (2) If $\Delta\sigma_i < -|\Delta\sigma|_s$, the portion of helix at the i-th residue remains a helix while sheet structure becomes a helix. (3) If $\Delta\sigma_i < -|\Delta\sigma|_h$, the portion of helix at the i-th residue becomes a turn (loop). Besides, we require that each helix has at least four residues. Finally, we calculate $\Delta\sigma_i$ and determine the structure changes of W140A, F61W/W140A, Y93W/W140A, and E75G in Fig. 6. Here, the secondary structures of WT are shown in upper panel and the structures predicted by LHP analysis are shown in lower panel. For example, in Fig. 6(a), $\Delta\sigma_i$ at 128–141 are larger than $|\Delta\sigma|_s$ and a helix becomes a sheet. By counting the number of residues in helices and sheets after such manipulations, we estimate the percentage changes of helix and sheet contents. The results are shown in Table I. The structure of E75G has been determined to be the same as WT SNase.[25] Our estimation from LHP is consistent with the PDB data. The differences between the estimation from CD spectra and the present work are within 8%.

Furthermore, we have also used the FoldIndex[26] for prediction. FoldIndex is a folding degree predictor that estimates the local and general probability for a sequence to fold under specified conditions of hydrophobicity and window size.[26] For WT SNase and its mutants, the folding degree in order is W93W/W140A (−0.072) > W140A (−0.073) > WT (−0.079) = E75G (−0.079) > F61W/W140A (−0.081), showing the same trend of the folding fractions of helix and sheet predicted by LHP analysis.

3.2 *Lysozyme*

Next, we apply our methods on the hen egg white lysozyme and its mutants. We use the PDB entry 2LYZ as the structure file of WT lysozyme, in which there are 129 amino acid residues. To ensure that the mutants under study

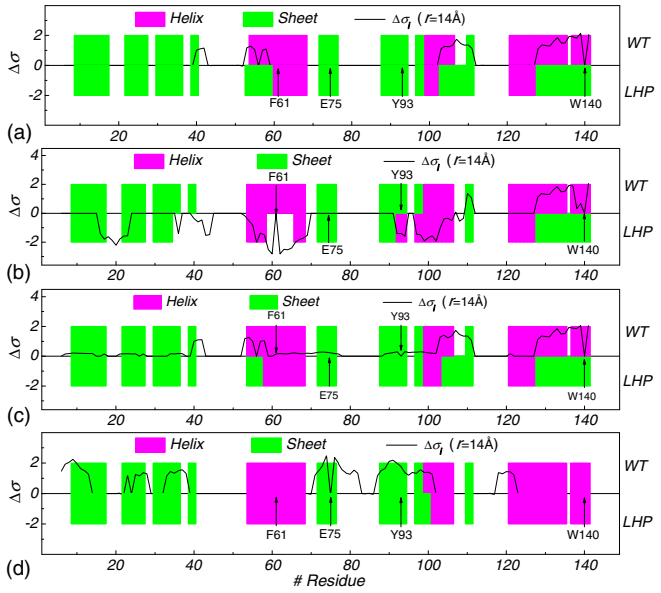

Fig. 6. (Color online) Changes of local hydrophobicity LHP ($\Delta\sigma_i$) of mutants (a) W140A, (b) F61W/W140A, (c) Y93W/W140A, and (d) E75G. The upper panel labels secondary structures in WT SNase, and the lower panel labels predictions from $\Delta\sigma_i$.

are in the same chemical condition, we select mutants in an experiment, deposited by the same group.[27] There are five mutants with solved structures available at PDB, including G49A(1FN5), G67A(1FLU), G71A(1FLW), G102A(1FLY), and G117A(1FLQ). The EMD of HP and LHP of lysozyme yields 5 IMFs ($c_1, \ldots, c_5$) and 1 residue (r_5), as shown in Fig. 7(a). The orthogonality test gives $\kappa = -0.00003$ for HP and $\kappa = -0.0148$ for LHP. IMFs c_4 and c_5 of HP and LHP are more correlated than other component pairs. IMFs c_4 and c_5 of LHP show correlations with secondary structures as shown in Fig. 7(b). For the case of lysozyme, sheet structures share higher LHP than helix structures. The average of LHP is $\langle\sigma\rangle = -3.4$ for all residues, $\langle\sigma\rangle = -5.4$ for residues in random coils and turns, $\langle\sigma\rangle = 2.8$ for residues in helices, and $\langle\sigma\rangle = -7.0$ for residues in sheets. The relatively lower value of average LHP for sheet structures is due to the fact that all sheet structures are surface-exposed region in lysozyme. Thus, we have

$$\langle w\rangle(\text{sheet}) > \langle w\rangle(\text{turn or coil}) > \langle w\rangle(\text{helix}). \quad (11)$$

The tolerances of σ for different secondary structures determined by fitting $\Delta\sigma_i/|\langle\sigma\rangle|$ with G49A are 5% of the average of LHP for sheets (i.e., $|\Delta\sigma|_s \sim 0.35$) and 25% for helices (i.e., $|\Delta\sigma|_h \sim 0.70$). Our estimation for secondary structure contents in G49A as shown in Fig. 7(c) is 49.6% for helices, 8.5% for sheets, and 41.9% for others, with respect to 51.2% for helices, 6.4% for sheets, and 42.6% for others calculated directly from the PDB data (1FN5). Table II summarises the estimations percentage contents of secondary structures in the five mutants of lysozyme. These results show that the differences between the LHP estimations and the data from PDB structure files are within 5%.

3.3 *Ubiquitin*

As a further example, we analyze the structure file of ubiquitin with the PDB code 1UBQ, in which there are 76

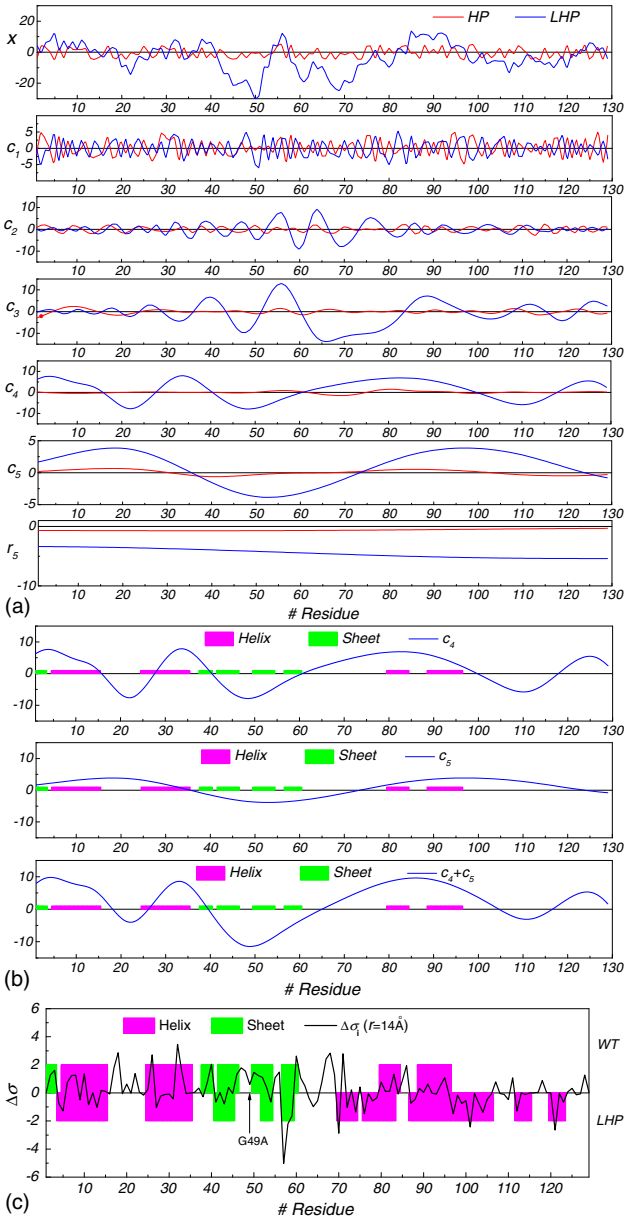

Fig. 7. (Color online) (a) Empirical mode decompositions of HP and LHP of hen egg white lysozyme (2LYZ). There are 5 IMFs ($c_1, \ldots, c_5$) and 1 residue (r_5). (b) IMFs c_4 and c_5 of LHP and secondary structures in lysozyme. (c) Changes of local hydrophobicity LHP ($\Delta\sigma_i$) of the mutant G49A. The upper panel labels secondary structures in WT, and the lower panel labels predictions from $\Delta\sigma_i$.

amino acids residues. The EMD of HP and LHP of ubiquitin yields 4 IMFs ($c_1, \ldots, c_4$) and 1 residue (r_4), as shown in Fig. 8(a). The orthogonality indices are $\kappa = 0.0211$ for HP and $\kappa = 0.0463$ for LHP. Among the IMFs, IMFs c_3 and c_4 of HP and LHP are more correlated than other modes. IMFs c_3 and c_4 of LHP also show correlations with secondary structures [Fig. 8(b)]. For ubiquitin, sheet structures share higher LHP than helix structures. The average of LHP is $\langle\sigma\rangle = -2.3$ for all residues, $\langle\sigma\rangle = -5.7$ for residues in random coils and turns, $\langle\sigma\rangle = -7.8$ for residues in helices, and $\langle\sigma\rangle = 3.1$ for residues in sheets. Thus, the relation of w among different secondary structures is

$$\langle w\rangle(\text{helix}) > \langle w\rangle(\text{turn or coil}) > \langle w\rangle(\text{sheet}). \quad (12)$$

Table II. Percentage contents of secondary structures in the mutants of hen white lysozyme estimated by PDB structures and the analysis of local hydrophobicity LHP.

Structure	G49A(1FN5)		G67A(1FLU)		G71A(1FLW)		G102A(1FLY)		G117A(1FLQ)	
	PDB	LHP	PDB	LHP	PDB	LHP	PDB	LHP	PDB	LHP
Helix (%)	51.2	49.6	51.2	46.5	51.2	49.6	51.2	49.6	51.2	53.5
Sheet (%)	6.2	8.5	6.2	7.0	6.2	8.5	6.2	7.0	6.2	9.3
Others (%)	42.6	41.9	42.6	46.5	42.6	41.9	42.6	43.4	42.6	37.2

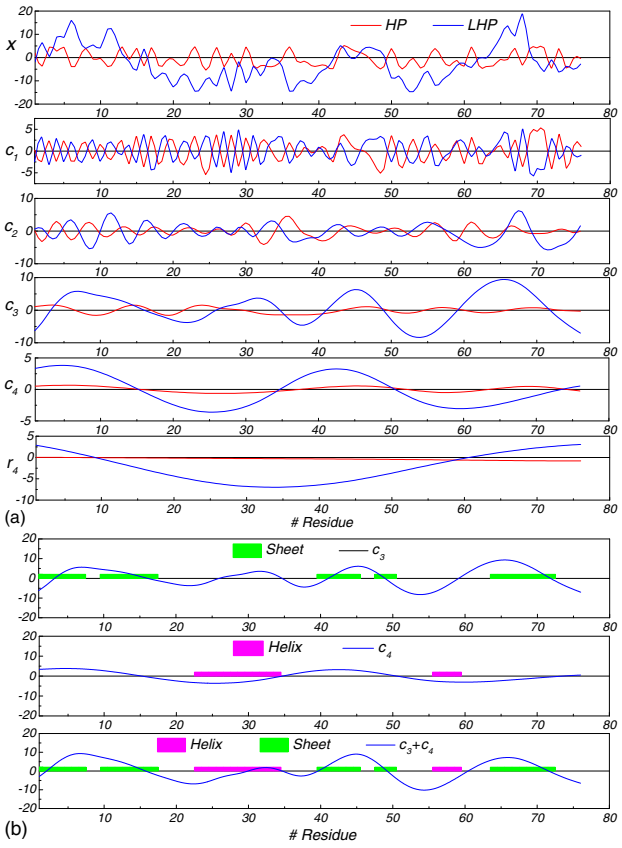

Fig. 8. (Color online) (a) Empirical mode decompositions of HP and LHP of ubiquitin. There are 4 IMFs ($c_1, \ldots, c_4$) and 1 residue (r_4). (b) IMFs c_3 and c_4 of LHP and secondary structures in ubiquitin.

The correlation between LHP and secondary structures again suggests that the proposed scheme in Sect. 2.3 is applicable in estimating secondary structure contents in the mutants of ubiquitin by considering the different ordering for w values of secondary structures. We skip the estimations here as the procedures and results are similar to those of SNase and lysozyme.

4. Conclusions

We have investigated the effect of LHP in protein secondary structure formation by analyzing the variations of secondary structures in mutants as a result of changes in LHP with respect to WT proteins. By defining LHP from three-dimensional protein structure files and considering HP and LHP as time series, we used time series analysis approach to analyze the relations between HP and LHP, and between LHP and secondary structures. Considering SNase, lysozyme, and ubiquitin as examples, we showed that there is correlation between IMFs of HP and LHP time series, and the correlation strength is associated with the length scales of IMFs defined via the EMD method. The correlated IMFs of LHP associated with the length scales of secondary structures are highly correlated with distribution of secondary structures. On the basis of such correlation, we tested the possibility to estimate percentage contents of secondary structures in mutants from changes of LHP. We demonstrated a rough algorithm to determine changes of secondary structures in mutants via changes of LHP in mutants. The results are quantitatively consistent with the results of far-UV CD spectra and PDB structure data of SNase and lysozyme.

The successful estimations of secondary structure contents in mutants via LHP analysis implies that LHP plays an essential role in proteins. It is assumed that HP of residues carries portions of the folding information,[2,3] and at least partially determine early nucleations of folding. When segments of sequence form sub-domains, LHP instead of HP plays a crucial role in subsequent hierarchical assemblies. The LHP defined from three-dimensional structure allows us to estimate the influences of single and double point mutation in secondary structures. The multi-scale view derived from EMD manifests the effects of LHP in different folding processes. Consequently, while LHP itself is insufficient for predicting tertiary structures, the knowledge derived from LHP analysis is informative in determining stability of mutants on the basis of secondary structures.

Finally, we emphasize that the proposed method is based on protein 3D structure data. Thus, it is unapplicable when such data is not available. Another limit comes from the assumption of the non-significant change of mutuant structure from a reference solved structure, which can be WT or another solved mutant. The estimation of changes of secondary structure contents will be false when this assumption is unrealistic.

Acknowledgments

This work was supported by the National Science Council of the Republic of China (Taiwan) under Grant No. NSC 100-2112-M-008-003-MY3, and NCTS of Taiwan.

*mcwu@ncu.edu.tw

1) M. I. Kanehisa and T. Y. Tsong: J. Mol. Biol. **124** (1978) 177.
2) G. D. Rose and S. Roy: Proc. Natl. Acad. Sci. U.S.A. **77** (1980) 4643.
3) N. T. Southall, K. A. Dill, and A. D. J. Haymet: J. Phys. Chem. B **106** (2002) 521.
4) M. I. Kanehisa and T. Y. Tsong: Biopolymers **19** (1980) 1617.
5) K. A. Dill: Biochemistry **29** (1990) 7133.
6) S. Sun, R. Brem, H. S. Chan, and K. A. Dill: Protein Eng. **8** (1995) 1205.
7) T. E. Creighton: *Proteins: Structures and Molecular Properties*

(Freeman, New York, 1993) 2nd ed., pp. 153 and 160.
8) J. Kyte and R. F. Doolittle: J. Mol. Biol. **157** (1982) 105.
9) Y. Nozaki and C. Tanford: J. Biol. Chem. **246** (1971) 2211.
10) T. P. Hopp and K. R. Woods: Proc. Natl. Acad. Sci. U.S.A. **78** (1981) 3824.
11) H. S. Chan and K. A. Dill: Phys. Today **46** [2] (1993) 24.
12) T.-Y. Tsong, C.-K. Hu, and M.-C. Wu: Biosystems **93** (2008) 78.
13) H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne: Nucleic Acids Res. **28** (2000) 235.
14) N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu: Proc. R. Soc. London, Ser. A **454** (1998) 903.
15) M.-C. Wu, M.-C. Huang, H.-C. Yu, and T. C. Chiang: Phys. Rev. E **73** (2006) 016118.
16) M.-C. Wu and C.-K. Hu: Phys. Rev. E **73** (2006) 051917.
17) M.-C. Wu: Physica A **375** (2007) 633.
18) M.-C. Wu: J. Korean Phys. Soc. **50** (2007) 304.
19) M.-C. Wu, E. Watanabe, Z. R. Struzik, C.-K. Hu, and Y. Yamamoto: Phys. Rev. E **80** (2009) 051917.
20) M.-C. Wu: Europhys. Lett. **97** (2012) 48009.
21) C. B. Anfinsen: Biochem. J. **128** (1972) 737.
22) C.-Y. Chow, M.-C. Wu, H.-J. Fang, C.-K. Hu, H.-M. Chen, and T.-Y. Tsong: Proteins: Struct. Funct. Bioinformatics **72** (2008) 901.
23) H.-Y. Hu, M.-C. Wu, H.-J. Fang, M. D. Forrest, C.-K. Hu, T. Y. Tsong, and H. M. Chen: Biophys. Chem. **151** (2010) 170.
24) M. Heinig and D. Frishman: Nucleic Acids Res. **32** (2004) W500. The STRIDE server is at [http://webclu.bio.wzw.tum.de/stride/].
25) K. W. Leung, Y.-C. Liaw, S.-C. Chan, H.-Y. Lo, F. N. Musayev, J. Z. W. Chen, H.-J. Fang, and H.-M. Chen: J. Biol. Chem. **276** (2001) 46039.
26) J. Prilusky, C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman, and J. L. Sussman: Bioinformatics **21** (2005) 3435. FoldIndex webpage: [http://bip.weizmann.ac.il/fldbin/findex].
27) K. Masumoto, T. Ueda, H. Motoshima, and T. Imoto: Protein Eng. **13** (2000) 691.