Deviation from the Residue-Based

Linear Free Energy Relationship Reveals

a Non-Native Structure in Protein Folding

Daisuke Fujinami, Seiichiro Hayashi, and Daisuke Kohda*

Division of Structural Biology, Medical Institute of Bioregulation, Kyushu University, Maidashi 3-1-1, Higashi-ku, Fukuoka 812-8582, Japan

Correspondence to Daisuke Kohda

Postal address: Medical Institute of Bioregulation, Kyushu University, Maidashi 3-1-1, Higashi-ku, Fukuoka 812-8582, Japan, E-mail: kohda@bioreg.kyushu-u.ac.jp

Highlights

- A collection of equilibrium and kinetic constants of structural changes of proteins
- Residue-based linear free energy relationship widely holds between the two constants
- rbLFER indicates the absence of non-ground state structures in transition states
- rbLFER is an experiment proof of the consistency principle of protein folding
- Deviations from the linear relation suggest special structures in transition states

Abstract

Multiprobe measurements, such as NMR and hydrogen exchange study, can provide the equilibrium constant K and kinetic rate constant k of the structural changes of a polypeptide on a per-residue basis. We previously found a linear relationship between residue-specific $\log K$ values and residue-specific $\log k$ values for the two-state topological isomerization of a 27-residue peptide. To test the general applicability of the residue-based linear free energy relationship (rbLFER), we performed a literature search to collect residue-specific equilibrium and kinetic constants in various exchange processes, including protein folding, coupled folding and binding of intrinsically disordered peptides, and structural fluctuations of folded proteins. The good linearity in a substantial number of log-log plots proved that the rbLFER holds for the structural changes in a wide variety of protein-related phenomena. Protein molecules quickly fold into their native structures and change their conformations smoothly. Theoretical studies and molecular simulations advocate that the physicochemical basis is the consistency principle and the minimal frustration principle: Non-native structures/interactions are absent or minimized along the folding pathway. The linearity of the residue-based free energy relationship demonstrates experimentally the absence of non-native structures in transition states. In this context, the hydrogen exchange study of apomyoglobin folding intermediates is particularly interesting. We found that the residues that deviated from the linear relationship corresponded to the non-native structure, which had been identified by other experiments. The rbLFER provides a unique and practical method to probe the dynamic aspects of the transition states of protein molecules.

KEYWORDS: hydrogen exchange, non-native structure, protein folding, transition state, two-state exchange

Abbreviations used:

CPMG, Carr-Purcell-Meiboom-Gill
 EXSY, exchange spectroscopy
 FRET, Förster resonance energy transfer
 GdnDCI, Guanidine deuterium chloride
 HSQC, heteronuclear single quantum coherence
 HX, hydrogen exchange

IDP, intrinsically disordered polypeptide

LFER, linear free energy relationship

rbLFER, residue-based LFER

RD, relaxation dispersion

REFER, rate-equilibrium free energy relationship.

Introduction

Protein molecules quickly fold into their thermodynamically stable and biologically active structures. Protein folding is a property acquired by evolution and is highly optimized for each amino acid sequence. Nobuhiro Gō proposed the ‘consistency principle’ theory in 1983[1]. The consistency principle of protein folding is the physicochemical basis for a perfectly smooth surface of the funnel-like energy landscape to materialize the smooth folding of a polypeptide chain. It advocates the absence of non-native structures or interactions in every moment on the folding paths from any random initial structure. In 1987, Bryngelson and Wolynes proposed a more realistic version of the consistency principle, based on the theory of spin glasses[2]. Through evolutionary selections, the surface of the funnel-like landscape has been sculpted to become smooth by minimizing the frustration of tertiary interactions, but remains rugged to the minimum necessary for satisfying the functionality requirements, such as enzymatic activities and allostery of proteins[3,4]. The consistency principle and the minimum frustration principle have been widely supported by many molecular simulations. In 2013, the refolding trajectories of small proteins obtained by all-atom MD simulations with explicit water solvent were analyzed from the perspective of folding mechanisms[5]. In the successful refolding paths, native contacts were formed at higher probabilities than non-native contacts, strongly supporting the viewpoint of the consistency principle.

NMR provides local information around nuclei at an atomic resolution. We used NMR to determine the residue-specific equilibrium constants and residue-specific kinetic rate constants of a polypeptide chain in a two-state dynamic exchange, and discovered a linear relationship in the double logarithmic plot of the two parameters[6]. This type of plot is called a REFER (rate equilibrium free energy relationship) plot, and the linear relationship in this plot is generally referred to as the Linear Free Energy Relationship (LFER). LFER is widely observed in many chemical and biological fields and used to investigate the transition states of reaction and exchange phenomena[7–12]. Importantly, all known types of LFER are composed of data points obtained from single probe measurements. Data points, (K , k), are derived from different molecules by independent measurements, in which a perturbation is a difference in the chemical structure (including amino acid sequence), or from one molecule by repeated measurements under different perturbed

conditions, such as variations in the solvent. The interpretation of the results always requires the assumption that the perturbations do not affect the transition paths. In contrast, the LFER we found is derived from many amino acid residues of one polypeptide chain under a single set of conditions. The “perturbation” is the different positions of residues as the NMR probe. In this respect, we refer to this new type of LFER as residue-based LFER (rbLFER). Obviously, rbLFER is not affected by unintended changes of the transition paths.

Residue-based LFER was discovered in the NMR study of a 27-residue bioactive peptide, nukacin ISK-1[6]. The two sets of ^1H - ^{15}N amide cross peaks in the ^1H - ^{15}N HSQC (heteronuclear single quantum coherence) spectrum indicated the existence of two states in solution, termed states A and B[13]. The interconversion rate was slow on the second time scale, reflecting the high energy barrier, 16 $k_B T$. The peak intensity ratio of the two cross peaks of each residue in the HSQC spectrum provides an equilibrium constant, K . Unexpectedly, we found that the K values varied from one residue to another[14]. Therefore, we determined the exchange rates, k (forward) and k' (backward), using EXSY (exchange spectroscopy)[6]. We found that both the residue-specific k and k' values also varied. Subsequently, we examined the relationship between the residue-specific thermodynamic and kinetic terms and found a linear relationship between $\log k$ and $\log K$ in the REFER plot. In principle, the equilibrium constant and the kinetic rate constants are independent, but the residue-based LFER suggested a hidden relationship between the transition state and the two ground states over the entire polypeptide chain. Our present interpretation is that the physicochemical basis of the smooth transition between the two states emerges as the linear relationship in the REFER plot of nukacin ISK-1.

Here, we performed a literature search to collect residue-specific equilibrium and kinetic constants of proteins and generated REFER plots. A substantial number of the REFER plots exhibited good linearity, which is, in our opinion, a direct result of the consistency principle of protein conformational changes. Among the successful cases, the HX (hydrogen exchange) study of apomyoglobin folding intermediates is particularly interesting[15]. The majority of the amino acid residues are on a straight line in the REFER plot, but some residues deviate from the line. We found that these outlier residues precisely correspond to the non-native structures in the apomyoglobin folding intermediates, which were first discovered a decade-and-a-half ago[16,17]. The excellent agreement between the two independent analyses demonstrates that rbLFER can provide the structural aspects of the dynamic properties of protein molecules.

Results

Residue-based LFER examples

Extensive literature searches led to the collection of reports on residue-specific equilibrium constants and residue-specific kinetic constants of proteins (Table 1). We recovered the data from the tables and figures and generated the corresponding REFER plots (Fig. 1). The techniques used and the phenomena under consideration are quite diverse. The first category (including nukacin ISK-1) is the EXSY NMR analyses of slow exchanges between two sets of cross peaks (Fig. 1A and 1B)[6,18,19], and the second one is the relaxation dispersion (RD) NMR analyses of the binding of an IDP (intrinsically denatured polypeptide) to a target protein (Fig. 1C)[20,21]. In the case of an IDP, only one set of cross peaks was observed due to the fast exchange rates, but the information on the exchange process can be extracted from the averaged cross peaks. The same situation is applied to the structural fluctuations of native-state proteins studied by RD NMR (Fig. 1D)[22–27]. The HX experiment also provides information on the structural fluctuations of native proteins (Fig. 1E)[28–31]. Note that the NMR in the HX studies is used for the quantitation of the proton occupancy and does not directly observe the HX phenomena. Indeed, mass spectrometry is applicable in place of NMR. Overall, the majority of the REFER plots showed residue-based linear relationships.

Here, we discuss the interpretation of the slopes of the REFER plots, by focusing on the forward direction (right-pointing arrows) depicted by the blue lines and their associated data points (Fig. 1). The slope represents the structural and energetical similarity between the starting ground state and the transition state, and hence the value must be between 0 and 1. In the case of structural fluctuations of the native states (Fig. 1D and 1E), the blue lines are nearly horizontal, and the slopes are almost zero in many instances. This situation indicates the high similarity of the transition state to the native state. This is a convincing result, considering that the interactions that stabilize the native state must be disrupted at the beginning of the fluctuation. The least-square lines with negative slopes and slopes greater than 1 are difficult to interpret. These cases are depicted by dashed lines. In some cases, particularly for large proteins, the data points are scattered, and no least-square lines are shown. Understandably, the REFER plots based on the literature data, including those with interpretable slopes, must be properly assessed in the future.

Alternative representation of residue-based LFER

The physical interpretation of the REFER plot is clear: the two axes, $\log K$ and $\log k$, are proportional to the changes in free energy terms. However, the two axes are not independent due to the equation, $\log K = \log k - \log k'$, which tends to exaggerate the linearity of the REFER plot (see the next section). Therefore, the $\log k$ vs. $\log k'$ plot is suitable for the assessment of rbLFER. Figure 2 shows the type classifications of $\log k$ vs. $\log k'$ plots, and Figure 3 shows the $\log k$ vs. $\log k'$ plots of the real examples listed in Table 1. The classification types are defined according to the distribution pattern of the data points. Type N is referred to as the negative correlation between $\log k$ and $\log k'$, which leads to the blue least-

square lines with the slope of $0 < \rho < 1$ in the REFER plot (cf. equation 9 in the previous paper[6]). In extreme cases, when the distribution of data points in the $\log k$ vs. $\log k'$ plot has a flattened shape, the least square lines have the slope of 0 or 1. According to the orientation of the oval-shaped distribution, vertical and horizontal, their types are defined as V and H, respectively. Since either $\log k$ or $\log k'$ is rather constant, the two log terms are uncorrelated in types V and H. Consequently, the Pearson's correlation coefficient R is zero in the $\log k$ vs. $\log k'$ plot, and one of the two least-square lines with a zero slope has an almost zero R^2 -value in the REFER plot. This fact indicates that the R and R^2 values in the two plots are not always good indicators of rLFER. Instead, 95% confidence ellipses are drawn to quantify the flat distributions (Fig. 3A-E). The closer flatness of the distribution to 1 reflects the higher linearity of the rLFER. Next, we consider the case of positive correlations between $\log k$ and $\log k'$. If the degree of variation of $\log k'$ is larger than that of $\log k$, the type is P, and if the degree of variation of $\log k'$ is smaller than that of $\log k$, the type is P'. Because the slopes of the blue least-square lines become negative or greater than 1 according to the type, the slopes are not physically meaningful. Such anomalous slope values occur if the exchange process cannot be described by a simple two-state model. For example, the ^{15}N relaxation dispersion studies of various mutants of the Fyn SH3 domain revealed the presence of about 1% of the transient intermediate state I in the three-state model, $\text{N} \rightleftharpoons \text{I} \rightleftharpoons \text{U}$ [22,23,26]. Even though the percentage is small, the assumption of the two-state model is not strictly valid, and thus the $\log k$ vs. $\log k'$ plots do not always show meaningful rLFER properties (Fig. 1D). Another possible cause is unintended measurement biases. The final type is nr (no relation). This is probably due to large measurement errors or accidental problems. For example, in the case of drk SH3N (Fig. 1B), the NMR spectra were recorded at pH 6. Under near-neutral conditions, the large exchange rate of amide protons with bulk water might reduce the measurement accuracy, despite the use of special NMR pulse sequences[18]. In the nukacin ISK-1 experiments, we used a low pH (pH 3.5) to slow down the water exchange and the P analysis to identify problematic data points affected by the exchange with water[6].

Risk of overestimation of linearity

The imbalance in the numbers of experimental values to be obtained and parameters to be determined is a serious problem in the accurate and precise determinations of the residue-specific equilibrium and residue-specific kinetic rate constants. As for ^1H - ^{15}N NMR, resonance- and state-specific NMR parameters, such as the relaxation rates, R_1 and R_2 , of ^1H and ^{15}N nuclei, must be considered in addition to the objective equilibrium and kinetic rate constants. Global fitting is a solution, but only the average value over many residues is available. Alternatively, an appropriate assumption can be introduced to reduce the number of fitting parameters. For example, some parameters are supposed to remain unchanged under different measurement conditions (Table 1). Therefore, we have to pay attention to the

risk that unexpected adverse effects caused by obligatory assumptions could generate an artificial linear relationship in REFER plots.

As already mentioned, the two axes of the REFER plot are not independent. Let ρ be the slope and α be the offset of rbLFER,

$$\log k = \rho \log K + \alpha. \quad (1)$$

Let ε be the deviation factor of k due to the biases and errors,

$$k^* = \varepsilon k \text{ and } K^* = \varepsilon k / k' = \varepsilon K \quad (2)$$

$$\log k^* - \log \varepsilon = \rho (\log K^* - \log \varepsilon) + \alpha. \quad (3)$$

Equation 3 indicates that the data point moves by $(\log \varepsilon, \log \varepsilon)$. This means that, if ρ has a value close to 0.5, the data point remains on the line and poses a risk of linearity overestimation.

In summary, the influences of measurement-specific biases and measurement errors must be considered seriously. The moderate linearity in the REFER plots does not simply prove a direct connection between the equilibrium constants and kinetic constants. The examination of the $\log k$ vs. $\log k'$ plot may help to identify such misinterpretations. Ideally, a special insight from the REFER plot can be tested by the results obtained from other experiments. The retrospective analysis of apomyoglobin folding intermediates in the next section provides a clear illustration of this point.

HX experiment of apomyoglobin

Information on the structural fluctuations of proteins can be obtained by monitoring the hydrogen/deuterium exchange of backbone amide protons with bulk water. The exchange mechanism consists of two processes: a two-state exchange of structural conversion and the exchange of the isotopes[32].

where NH(closed) is a folded state in which amide protons are protected from an exchange, and NH(open) is an open state in which exchange occurs. The H/D exchange rate, k_{int} , is highly dependent on the pH of the solution. Usually, a single high pH pulse is used in the EX1 regime ($k_{int} \gg k_{cl}$) to simplify the analysis. Alternatively, a wide pH range of the labeling pulse can be used. Since the pH dependence of k_{int} is well

known, k_{op} and k_{cl} can be determined without assuming the exchange mechanism, either EX1 ($k_{int} \gg k_{cl}$) or EX2 ($k_{int} \ll k_{cl}$). After a hydrogen/deuterium exchange reaction, NMR is used to measure the proton occupancy of each residue in an acidic solution. Mass spectrometry is also used after the fragmentation of proteins by protease digestion.

Sperm whale myoglobin is a popular model protein for understanding protein folding[33,34]. Myoglobin is a globular protein consisting of eight α -helices, designated A to H. The apo form of the protein has almost the same structure as the heme-bound holo form. Within the initial burst phase of apomyoglobin refolding, two kinetic intermediates, designated as I_a and I_b , are sequentially formed[35]. In the state I_a structure, the major portions of helices A, G, and H and part of helix B are established. Subsequently, parts of helices C, D, and E are formed and added to the already-existing helices in the state I_b structure (see Fig. 5C)[33,34]. Wright's group performed quenched-flow hydrogen exchange experiments using a continuous-flow mixer to determine the residue-specific kinetic parameters, k_{op} and k_{cl} , of the folding intermediates[15]. The k_{op} and k_{cl} values were each assumed to remain unchanged in the two labeling pulse durations of 0.4-4.0 ms and 6.0-9.6 ms, and simultaneous numerical fitting was performed to obtain a more accurate estimation of the rate constants. Consequently, k_{op} and k_{cl} are averaged values of I_a and I_b , respectively. We collected the k_{op} and k_{cl} data and associated errors from the literature[15] and constructed the REFER plot. We found that the (K , k_{op}) and (K , k_{cl}) data points were modestly aligned around straight lines (Fig. 4A). Upon closer inspection, several outlier data points were visually identified. Outlier residues (green and magenta) were removed to redraw the least-square lines, and the residues on the updated least-square lines formed a cluster of type N in the $\log k$ vs. $\log k'$ plot (Fig. 4B and 4C). The slope ρ of the blue least-square line is 0.23. This is in contrast to the HX studies of the native states of other proteins, which provided almost zero slopes (Fig. 1D and 1E). This reflects the differences in the shapes of the energy surfaces. The intermediate states of protein folding are in a shallow basin, but the native state is located at the bottom of a deep basin (see Fig. 5B).

Figure 5 schematically shows the folding funnel and the energy diagram of apomyoglobin. The curved lines in the funnel depict refolding paths starting from unfolded states. The horizontal axis is the reaction coordinate of protein folding from 0 to 1. This axis can be regarded as the time axis, and an apomyoglobin molecule moves from left to right along a path. In the early time window (0.4 – 9.6 ms) from the start of refolding (green area in Fig. 5A), the intermediate states I_a and I_b are sequentially formed, and each amide group is in an exchange between the closed and open states, $I_{close} \rightleftharpoons I^\ddagger \rightleftharpoons I_{open}$, along the folding path (Fig. 5B). Note that the transition state $I^\ddagger$ in the HX exchange reaction is different from the transition state T in the entire refolding process, $U \rightarrow N$.

The outlier residues were mapped on the N-state structure of apomyoglobin. All outlier residues (L104, F106, A134, L135, E136, L137, R139, D141, I142, and A143) are located at the interface between helix G and helix H (Fig. 5D, magenta). It is quite interesting that the Wright group showed that the intermediate I_b was a mixed state of two conformations, one of which contained the translocation of helix H by one helical turn toward its N-terminus relative to helix G (Fig. 5C), by using the HX study of amino acid mutants[16] and the combination of the HX study with fluorescence quenching and FRET (Förster resonance energy transfer) measurements[17]. The translocated helix H is not present in the folded state N, and thus they referred to the translocated form of helix H as a non-native structure. They also proposed that state I_b is in an energetic frustration at the B-G and G-H helix interfaces to facilitate the dynamic exchange of the two conformations. Strikingly, the region highlighted by the outlier residues precisely coincides with the amino acid residues involved in the non-native helix translocation (Fig. 5D, magenta)

Discussion

In a two-state exchange process, in principle, the equilibrium constant and one of the two kinetic rate constants are independent, and the other kinetic rate constant is uniquely determined. Here, we showed that proteins have the property of the residue-based LFER (Fig. 1). The linear double logarithmic relationships between residue-specific equilibrium constants and the two kinetic rates of the same residue indicate that the thermodynamic and kinetic energy terms are closely related throughout a polypeptide chain. As a building block of a polymer, each amino acid residue of a protein participates in the conformational changes in a coordinated fashion to achieve smooth structural transformations. We consider this special feature to reflect a property of proteins acquired through evolution.

The residue-based LFER is an experimental demonstration of the consistency principle. This declaration can be proved by contradiction. Here, we use the term ‘non-native’ in the sense that it is absent in the two ground states including the unstructured state. If non-native structures or interactions are present in the transition state, then the involved amino acid residues can affect the free energy of the transition state independently of the ground states and the position of the corresponding data point will deviate from the least square line. In rare instances, however, the data points are aligned by chance. Thus, in a precise mathematical sense, the role of rbLFER is more of a statistical null hypothesis test than a rigorous mathematical proof of the consistency principle. In other words, it is almost impossible to explain the physicochemical basis of the linear relationships in Fig. 1 by the assumption that the transition states contain non-native structures or interactions.

The consistency principle represents an ideal protein. Real proteins do possess some non-ideality due to minimal frustrations. An ideal protein would provide (K, k) data points on a perfectly straight line

in the REFER plot, whereas a real protein provides (K, k) data points scattered around a straight line, even after perfect measurement error removal. Here, we found the deviations from the least-square lines in the REFER plot for the apomyoglobin folding intermediates (Fig. 4). The distribution of the outlier residues in the three-dimensional structure is in good agreement with the non-native translocation of helix H in intermediate I_b , which was previously identified by other methods (Fig. 5D). The prevention of the formation of the translocated form by an engineered disulfide bond increased the overall folding rate by two-fold[17], and the stabilization of the translocated form by mutations in helix F decreased the folding rate down to 20%[34]. These results are in agreement with the consistency principle: a non-native structure on the folding pathway delays the overall folding process.

The quenched-flow HX experiment of apomyoglobin examined the structural conversion between I_{close} and I_{open} through the transition state $I^\ddagger$ along the folding path (Fig. 5B). For better precision, the two exchange processes, $I_a \rightleftharpoons I^\ddagger \rightleftharpoons I_{op}$ and $I_b \rightleftharpoons I^\ddagger \rightleftharpoons I_{op}$, were jointly analyzed to provide their averaged picture[15]. First, we focus on the structural fluctuation of I_b . The I_{op} state connected with I_b is more or less similar to I_a . We assume that $I^\ddagger$ contains a greater fraction of the translocated helix H than that in the ground state I_b . The increase in the fraction of the translocation is the 'non-native structure' that is absent in the two ground states, I_b and I_{op} . Another interpretation is the elevated exchange rate of the helix translocation. In this case, the increase in the rate is the 'non-native interaction' that is absent in the two ground states. These non-native structures/interactions violate the consistency principle and generate the outlier residues in the REFER plot. In contrast to I_b , I_a contains fewer secondary structures and thus does not contribute much to the deviation of the REFER plot. In sum, the rbLFER of the HX study detected non-native structures/interactions on the folding pathway by observing the transition state $I^\ddagger$ (Fig. 5B).

Conclusion

The linear relationship between $\log K$ (the logarithm of the equilibrium constant) and $\log k$ (the logarithm of the rate constant) in the REFER plot holds for structural changes of many proteins, according to our retrospective analysis (Fig. 1). The analytical methods include EXSY NMR, R_2 dispersion relaxation NMR, and HX measurements. The residue-based LFER is seen in the two-state slow exchange (Fig. 1A) and structural fluctuations (Fig. 1D and 1E) of monomeric small proteins, and in two-molecule systems, such as IDP-protein interactions (Fig. 1C). The substantial number of rbLFER examples indicates the applicability of the consistency principle to a wide variety of protein-related phenomena. The linearity of the REFER plot can be used as a measure of the deviation from ideal conditions for smooth protein structural changes. The reanalysis of the HX study of apomyoglobin[15] is particularly fruitful (Figs. 4 and 5). This unexpected outcome demonstrates that the rbLFER is not a special experiment to prove the consistency principle but is a practical method to obtain dynamic information of proteins. Finally, one must

bear in mind that measurement biases and errors could lead to the overestimation of linearity in the REFER plots. The development of a novel methodology is keenly anticipated for bias-free, accurate determinations of the residue-specific exchange parameters of polypeptide chains.

Methods

We performed a literature search to collect residue-specific equilibrium and residue-specific kinetic constants of proteins, mainly in the PubMed literature database (<http://www.ncbi.nlm.nih.gov/pubmed/>). The keywords included ‘two states’, ‘two sets of cross peaks’, ‘exchange spectroscopy’, ‘residue-specific’, LFER, etc., and their combinations. The linear regression analyses of the REFER plots and log k vs. log k' plots were performed in the Excel files. The protein cartoon was generated with the program PyMOL, version 2.4.2 (Schrödinger). The cartoon image of the apomyoglobin was generated using the PDB entry 2JHO.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science (JSPS, Japan), KAKENHI Grant Number JP21H02448, and the Mitsubishi Foundation (Japan) Research Grants in the Natural Sciences, Grant Number 202110017, to D.K.

Author Contributions

Conceptualization, D.F., and D.K.; Investigation, D.F., H.S., and D.K.; Writing – Original Draft, D.K.; Writing – Review & Editing, F.D., and S.H.; Funding Acquisition, D.K.

Declaration of Interests

The authors declare no competing interests.

Appendix A. Supplementary data

Supplementary data for this article include one figure (PDF).

Details of the REFER plot of the apomyoglobin folding intermediate (PDF)

References

- [1] N. Go, Theoretical Studies of Protein Folding, *Annu. Rev. Biophys. Bioeng.* 12 (1983) 183–210. <https://doi.org/10.1146/annurev.bb.12.060183.001151>.
- [2] J.D. Bryngelson, P.G. Wolynes, Spin glasses and the statistical mechanics of protein folding., *Proc. Natl. Acad. Sci. U. S. A.* 84 (1987) 7524–7528. <https://doi.org/10.1073/pnas.84.21.7524>.
- [3] D.U. Ferreira, E.A. Komives, P.G. Wolynes, Frustration in biomolecules, *Q. Rev. Biophys.* 47 (2014) 285–363. <https://doi.org/10.1017/S0033583514000092>.
- [4] D.U. Ferreira, E.A. Komives, P.G. Wolynes, Frustration, function and folding, *Curr. Opin. Struct. Biol.* 48 (2018) 68–73. <https://doi.org/10.1016/j.sbi.2017.09.006>.
- [5] R.B. Best, G. Hummer, W.A. Eaton, Native contacts determine protein folding mechanisms in atomistic simulations, *Proc. Natl. Acad. Sci. U. S. A.* 110 (2013) 17874–17879. <https://doi.org/10.1073/pnas.1311599110>.
- [6] D. Fujinami, S. Hayashi, D. Kohda, Residue-Specific Kinetic Insights into the Transition State in Slow Polypeptide Topological Isomerization by NMR Exchange Spectroscopy, *J. Phys. Chem. Lett.* 12 (2021) 10551–10557. <https://doi.org/10.1021/acs.jpclett.1c02387>.
- [7] A. Matouschek, A.R. Fersht, Application of physical organic chemistry to engineered mutants of proteins: Hammond postulate behavior in the transition state of protein folding, *Proc. Natl. Acad. Sci. U. S. A.* 90 (1993) 7814–7818. <https://doi.org/10.1073/pnas.90.16.7814>.
- [8] K.W. Plaxco, J.I. Guijarro, C.J. Morton, M. Pitkeathly, I.D. Campbell, C.M. Dobson, The folding kinetics and thermodynamics of the Fyn-SH3 domain, *Biochemistry.* 37 (1998) 2529–2537. <https://doi.org/10.1021/bi972075u>.
- [9] B. Song, J.H. Cho, D.P. Raleigh, Ionic-strength-dependent effects in protein folding: Analysis of rate equilibrium free-energy relationships and their interpretation, *Biochemistry.* 46 (2007) 14206–14214. <https://doi.org/10.1021/bi701645g>.
- [10] A.R. Fersht, A. Matouschek, L. Serrano, The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding, *J. Mol. Biol.* 224 (1992) 771–782. [https://doi.org/10.1016/0022-2836\(92\)90561-W](https://doi.org/10.1016/0022-2836(92)90561-W).
- [11] A.R. Fersht, S. Sato, Φ -Value Analysis and the Nature of Protein-Folding Transition States, *Proc. Natl. Acad. Sci. U. S. A.* 101 (2004) 7976–7981. <https://doi.org/10.1073/pnas.0402684101>.

- [12] D.A. Kingery, S.A. Strobel, Analysis of enzymatic transacylase Brønsted Studies with application to the ribosome, *Acc. Chem. Res.* 45 (2012) 495–503. <https://doi.org/10.1021/ar100162b>.
- [13] D. Fujinami, A.-A. -Mahin, K.M. Elsayed, M.R. Islam, J. Nagao, U. Roy, S. Momin, T. Zendo, D. Kohda, K. Sonomoto, The lantibiotic nukacin ISK-1 exists in an equilibrium between active and inactive lipid-II binding states, *Commun. Biol.* 1 (2018) 150. <https://doi.org/10.1038/s42003-018-0150-3>.
- [14] D. Fujinami, H. Motomura, H. Oshima, A.-A. Mahin, K.M. Elsayed, T. Zendo, Y. Sugita, K. Sonomoto, D. Kohda, Mosaic Cooperativity in Slow Polypeptide Topological Isomerization Revealed by Residue-Specific NMR Thermodynamic Analysis, *J. Phys. Chem. Lett.* 11 (2020) 1934–1939. <https://doi.org/10.1021/acs.jpclett.9b03591>.
- [15] T. Uzawa, C. Nishimura, S. Akiyama, K. Ishimori, S. Takahashi, H.J. Dyson, P.E. Wright, Hierarchical folding mechanism of apomyoglobin revealed by ultra-fast H/D exchange coupled with 2D NMR, *Proc. Natl. Acad. Sci. U. S. A.* 105 (2008) 13859–13864. <https://doi.org/10.1073/pnas.0804033105>.
- [16] C. Nishimura, H.J. Dyson, P.E. Wright, Identification of native and non-native structure in kinetic folding intermediates of apomyoglobin, *J. Mol. Biol.* 355 (2006) 139–156. <https://doi.org/10.1016/j.jmb.2005.10.047>.
- [17] P.C. Aoto, C. Nishimura, H.J. Dyson, P.E. Wright, Probing the non-native H helix translocation in apomyoglobin folding intermediates, *Biochemistry.* 53 (2014) 3767–3780. <https://doi.org/10.1021/bi500478m>.
- [18] N.A. Farrow, O. Zhang, J.D. Forman-Kay, L.E. Kay, Comparison of the Backbone Dynamics of a Folded and an Unfolded SH3 Domain Existing in Equilibrium in Aqueous Buffer, *Biochemistry.* 34 (1995) 868–878. <https://doi.org/10.1021/bi00003a021>.
- [19] M. Tollinger, N.R. Skrynnikov, F.A.A. Mulder, J.D. Forman-Kay, L.E. Kay, Slow dynamics in folded and unfolded states of an SH3 domain, *J. Am. Chem. Soc.* 123 (2001) 11341–11352. <https://doi.org/10.1021/ja011300z>.
- [20] M. Arai, K. Sugase, H.J. Dyson, P.E. Wright, Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding, *Proc. Natl. Acad. Sci. U. S. A.* 112 (2015) 9614–9619. <https://doi.org/10.1073/pnas.1512799112>.
- [21] K. Sugase, H.J. Dyson, P.E. Wright, Mechanism of coupled folding and binding of an intrinsically disordered protein, *Nature.* 447 (2007) 1021–1025. <https://doi.org/10.1038/nature05858>.
- [22] D.M. Korzhnev, X. Salvatella, M. Vendruscolo, A.A. Di Nardo, A.R. Davidson, C.M. Dobson, L.E. Kay, Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR, *Nature.* 430 (2004) 586–590. <https://doi.org/10.1038/nature02655>.
- [23] D.M. Korzhnev, P. Neudecker, A. Zarrine-Afsar, A.R. Davidson, L.E. Kay, Abp1p and Fyn SH3 domains fold through similar low-populated intermediate states, *Biochemistry.* 45 (2006) 10175–10183. <https://doi.org/10.1021/bi0611560>.
- [24] D. Létourneau, M. Bédard, J. Cabana, A. Lefebvre, J.G. Lehoux, P. Lavigne, STARD6 on steroids: Solution structure, multiple timescale backbone dynamics and ligand binding mechanism, *Sci. Rep.* 6 (2016) 1–16. <https://doi.org/10.1038/srep28486>.

- [25] X. Liu, L.C. Golden, J.A. Lopez, T.R. Shepherd, L. Yu, E.J. Fuentes, Conformational Dynamics and Cooperativity Drive the Specificity of a Protein-Ligand Interaction, *Biophys. J.* 116 (2019) 2314–2330. <https://doi.org/10.1016/j.bpj.2019.05.008>.
- [26] D.M. Korzhnev, V.Y. Orekhov, L.E. Kay, Off-resonance R^{1ρ} NMR studies of exchange dynamics in proteins with low spin-lock fields: An application to a fyn SH3 domain, *J. Am. Chem. Soc.* 127 (2005) 713–721. <https://doi.org/10.1021/ja0446855>.
- [27] Y. Xiao, T. Lee, M.P. Latham, L.R. Warner, A. Tanimoto, A. Pardi, N.G. Ahn, Phosphorylation releases constraints to domain motion in ERK2, *Proc. Natl. Acad. Sci. U. S. A.* 111 (2014) 2506–2511. <https://doi.org/10.1073/pnas.1318899111>.
- [28] T. Sivaraman, C.B. Arrington, A.D. Robertson, Kinetics of unfolding and folding from amide hydrogen exchange in native ubiquitin, *Nat. Struct. Biol.* 8 (2001) 331–333. <https://doi.org/10.1038/86208>.
- [29] H.M. Rodriguez, A.D. Robertson, L.M. Gregoret, Native state EX2 and EX1 hydrogen exchange of Escherichia coli CspA, a small β-sheet protein, *Biochemistry.* 41 (2002) 2140–2148. <https://doi.org/10.1021/bi011347x>.
- [30] C.B. Arrington, L.M. Teesch, A.D. Robertson, Defining protein ensembles with native-state NH exchange: Kinetics of interconversion and cooperative units from combined NMR and MS analysis, *J. Mol. Biol.* 285 (1999) 1265–1275. <https://doi.org/10.1006/jmbi.1998.2338>.
- [31] C.B. Arrington, A.D. Robertson, Microsecond to minute dynamics revealed by EX1-type hydrogen exchange at nearly every backbone hydrogen bond in a native protein, *J. Mol. Biol.* 296 (2000) 1307–1317. <https://doi.org/10.1006/jmbi.2000.3536>.
- [32] M.M.G. Krishna, L. Hoang, Y. Lin, S.W. Englander, Hydrogen exchange methods to study protein folding, *Methods.* 34 (2004) 51–64. <https://doi.org/10.1016/j.ymeth.2004.03.005>.
- [33] H.J. Dyson, P.E. Wright, How does your protein fold? Elucidating the apomyoglobin folding pathway, *Acc. Chem. Res.* 50 (2017) 105–111. <https://doi.org/10.1021/acs.accounts.6b00511>.
- [34] C. Nishimura, Folding of apomyoglobin: Analysis of transient intermediate structure during refolding using quick hydrogen deuterium exchange and NMR, *Proc. Japan Acad. Ser. B Phys. Biol. Sci.* 93 (2017) 10–27. <https://doi.org/10.2183/pjab.93.002>.
- [35] M. Jamin, R.L. Baldwin, Two forms of the pH 4 folding intermediate of apomyoglobin, *J. Mol. Biol.* 276 (1998) 491–504. <https://doi.org/10.1006/jmbi.1997.1543>.
- [36] A.A. Di Nardo, D.M. Korzhnev, P.J. Stogios, A. Zarrine-Afsar, L.E. Kay, A.R. Davidson, Dramatic acceleration of protein folding by stabilization of a nonnative backbone conformation, *Proc. Natl. Acad. Sci.* 101 (2004) 7954–7959. <https://doi.org/10.1073/pnas.0400550101>.

Table 1. Proteins used for the generation of REFER plots

Protein/Domain	Comment	Method ^a	Assumption	Ref.	Source
Nukacin ISK-1	27 res, 3 mono S bonds	¹ H-EXSY		[6]	Tables S1 and S3
Drk SH3N	59 res	¹⁵ N-EXSY		[18]	Figs. 3 and 6
Myb32-KIX:MLL28 ^b	32 res Myb (IDP) and 87 res KIX + 28 res MLL	¹⁵ N-RD (CPMG) + titration	two states	[19]	Table S2
pKID-KIX	34 res pKID (IDP) and 87 res KIX	¹⁵ N-RD (CPMG) + titration	two states ^c	[20]	Fig. 5
FynSH3 G48M	59 res, 288 K	¹⁵ N-RD (CPMG)	two states ^d	[21]	Table 1
FynSH3 G48V	59 res, 283 K	¹⁵ N-RD (CPMG)	two states ^d	[22]	Fig. 1b
FynSH3 mutant	59 res, A39V/N53P/V55L, 308 K	¹⁵ N-RD (CPMG)	two states ^d	[22]	Fig. 1c
Abl1 SH3 mutant	59 res, E7L/V21K/N23G/G48V, 313 K	¹⁵ N-RD (CPMG)	two states ^d	[23]	Fig. 1b
STARD6	220 res, apo form	¹⁵ N-RD (CPMG)		[23]	Fig. 1a
Tiam2 PDZ mutant	90 res, M978L/E979K/F982L/V987L	¹⁵ N-RD (CPMG)		[24]	Fig. 4
Fyn SH3 G48M	59 res, 298 K	¹⁵ N-RD (R1ρ)	two states	[25]	Table 2
2P-ERK2 kinase	356 res, dual-phosphorylated form	Methyl ¹³ C-RD (CPMG)		[26]	Table 1
Ubiquitin	76 res, pH 7-9, 1.5 and 1.8 M guanidine deuterium chloride (GdnDCl)	HX/NMR	k _{op} and k _{cl} are independent of pH, EX1 mechanism	[27]	Table S2
CspA	70 res, pH 6-11	HX/NMR + NMR (saturation transfer)	k _{op} and k _{cl} are independent of pH, EX1+EX2 mechanism	[28]	Table 1
OMTKY3 (turkey ovomucoid third domain)	56 res, pH 6-10	HX/NMR&ESI-MS	k _{op} and k _{cl} are independent of pH, EX2 mechanism	[29]	Table 2
OMTKY3	56 res, pH 10-12	Quenched-flow HX/NMR	k _{op} and k _{cl} are independent of pH, EX1 mechanism	[30]	Table 1
apomyoglobin	153 res, apo form, pH 7-11	Quenched-flow HX/NMR	k _{op} and k _{cl} are independent of pH and have identical values at the refolding times of 0.4 and 6 ms. EX1+EX2 mechanism	[31]	Table S1

^a Relaxation dispersion (RD) experiments using either the CPMG (Carr-Purcell-Meiboom-Gill) pulse train or a constant radiofrequency spin lock field (R1ρ).

^b The minor binding site for the pKID peptide on the KIX domain was masked by the second IDP peptide derived from the MLL (mixed-lineage leukemia) protein.

^c In the original report, the authors concluded that the three-state model was appropriate for the coupled binding and folding of the pKID peptide, Free $\rightleftharpoons$ Intermediate (encounter complex) $\rightleftharpoons$ Bound. Here, the exchange between the free state and the encounter complex was analyzed.

^d In the original reports, the authors concluded that the three-state model was appropriate for the exchange processes. However, we used the kinetic constants calculated assuming a two-state model to generate the REFER plots.

Fig. 1. Residue-based REFER plots based on the literature data listed in Table 1. (A) Nukacin ISK-I. (B) N-terminal SH3 domain from *Drosophila* drk protein (drk SH3N). The drk SH3N is in an exchange between unstructured state U and native folded state N. There are different data sets from two reports. (C) Myb32 and pKID are intrinsically disordered polypeptides that bind to the KIX domain. (D) Structural fluctuations of native states were investigated by the relaxation dispersion (RD) NMR method. The SH3 domains are

derived from the Fyn and Abp1 proteins. The mutations in the SH3 domains markedly increased the folding rate despite their destabilization of the folding state, and have suitable properties for the RD studies[36]. The STARD (steroidogenic acute regulatory-related lipid transfer domain) from the STARD6 protein. The PDZ (PSD-95/Dlg/ZO-1) domain from the Tiam2 (T cell lymphoma invasion and metastasis 2) protein. A dual-phosphorylated (2P-ERK2) form of the ERK2 (extracellular signal-regulated kinase 2) protein. (E) Structural fluctuations of native states investigated by the hydrogen exchange method. There are two measurement conditions for ubiquitin and OMTKY3 (turkey ovomucoid third domain 3). In (A)-(E), the least-square lines and data points associated with the forward direction are colored blue, and those associated with the backward direction are orange. The least-square lines with interpretable slopes between 0 and 1 are depicted as solid, colored lines, whereas those with uninterpretable slopes less than 0 or greater than 1 are depicted as dashed black lines. No least-square lines are drawn if the correlations are considered insignificant. The pairs of arrows show the interconversion processes between two states and the structural fluctuations of native-state proteins. The concentric circle represents a basin-shaped energy landscape of the structural fluctuations around the native state, N , in equilibrium with the open state, N_{op} .

Fig. 2. Relationships between $\log k$ and $\log k'$ as the basis of rbLFEr. Several types are defined according to the data point distribution. Type N shows a negative correlation, and types V and H show flattened distributions of data points with zero correlations. The vertically flattened distribution provides two least-square lines with the slopes ρ of 0 (blue) and -1 (orange), and the horizontally flattened distribution provides those with the slopes ρ of 0 (orange) and 1 (blue) in the REFERENCE plots. Types P and P' are positive correlations between $\log k$ and $\log k'$. Type nr has no relation between $\log k$ and $\log k'$, and the REFERENCE plot may show a weak, artificial correlation trend.

Fig. 3. Datapoint distributions in the $\log k$ vs. $\log k'$ plots. In each panel, protein/domain name, category type defined in Fig. 2, a 95% confidence ellipse, and the flatness of distribution are shown. The flatness of distribution is defined by $f = 1 - b/a$, where a is the long axis and b is the short axis of the confidence ellipses. The axis ranges are set equally for the proper interpretation of the flatness of the confidence ellipses.

Fig. 4. REFER plot and log k vs. log k' plot of apomyoglobin folding intermediate. (A) REFER plot using all observed residues with estimated fitting uncertainties[15]. The data points and least square lines of the log k_{op} vs. log K plot and the log k_{cl} vs. log K plot are blue and orange, respectively. (B) Replot. Outlier residues (green and magenta) were removed to redraw the least-square lines (blue and orange). See Fig. S1 for details. The inset shows the position of the transition state $I^\ddagger$ on the reaction coordinate of the HX reaction. (C) Log k vs. log k' plot. The yellow dots are the residues contributing to the least-square lines, and the purple dots are the outlier residues in (B). The yellow group belongs to type N. The yellow dots are enclosed by a 95% confidence ellipse.

Fig. 5. Non-native structure in the apomyoglobin folding intermediate. (A) Schematic of the folding funnel from the unfolded state U to the native state N . The thick and thin curved lines in the funnel represent individual folding paths from two initial states. The time window (0.4-9.6 ms) probed by the HX method is colored green. The conformational space containing the translocated helix H is colored magenta. The movement of the trajectory lines in and out of the magenta region illustrates the helix H translocation. (B) Energy diagram of the folding process. The concentric circles represent a basin-shaped energy landscape of the structural fluctuations in a microscopic thermodynamic equilibrium with the corresponding open state through the specific transition states, $I^\ddagger$ and $N^\ddagger$. T denotes the transition states of the entire folding process. (C) Sequential formation of the eight helices in apomyoglobin[33,34]. The kinetic intermediates I_a and I_b were observed at the 0.4 ms and 6 ms time points, respectively. The half-translucent cylinders represent partially formed α -helices. Helix F is not stably packed in the apo N state. The white arrow indicates the translocation of helix H. (D) Mapping of the outlier residues in the REFER plot on the N -state structure of myoglobin (PDB 2JHO). The native holo-structure is used as the best alternative to the intermediate structures. The residues on the least-square lines in Fig. 4B are colored cyan and the outlier residues are magenta. The other residues without information are colored white. These unprobed residues include parts of helices A, G, and H due to the full protection of the amide protons, and helix F and the loops due to the lack of local secondary structures in the intermediate states.

Fig. S1. Details of the REFER plot of the apomyoglobin folding intermediate. (A) Labeling of amino acid residues. (B) Error bars of the first least-square lines (blue and orange) and the second least-square lines (green and magenta). The second lines are almost parallel to the first lines, which might suggest a hidden mechanism for the formation of the non-native structure, but the large errors hamper proper interpretation.