Evolution of off-lattice model proteins under ligand binding constraints

Erik D. Nelson and Nick V. Grishin

Phys. Rev. E **94**, 022410 — Published 15 August 2016

DOI: 10.1103/PhysRevE.94.022410

Evolution of model proteins under functional constraints

Erik D. Nelson* and Nick V. Grishin

Howard Hughes Medical Institute, University of Texas Southwestern Medical Center,

6001 Forest Park Blvd., Room ND10.124, Dallas, Texas 75235-9050

∗ E-mail: nelsonerikd@gmail.com

Abstract

We investigate protein evolution using an off–lattice polymer model evolved to imitate the behavior of small enzymes. Model proteins evolve through mutations to nucleotide sequences (including insertions and deletions) and are selected to fold and maintain a specific binding site compatible with a model ligand. We show that this requirement is, in itself, sufficient to maintain an ordered folding domain, and we compare it to the requirement of folding an ordered (but otherwise un–restricted) domain. We measure rates of amino acid change as a function local environment properties such as solvent exposure, packing density, and distance from the active site, as well as overall rates of sequence and structure change, both along and among model lineages in star phylogenies. The model recapitulates essentially all of the behavior found in protein phylogenetic analyses, and predicts that amino acid substitution rates vary linearly with distance from the binding site.

I. Introduction

Most proteins in nature are required to fold into a specific structure, or structure ensemble in order to perform their biochemical functions. These requirements impose strong constraints on the evolution of proteins and protein coding sequences. Mutations that interfere with folding or function are generally removed from a population by purifying selection, while neutral, or beneficial mutations can become fixed and accumulate over time. As a result of this interplay, functional requirements are reflected in the patterns of change within homologous protein sequences. Amino acid mutations that occur on the surface of a protein's folded structure, or are distant from its active site, tend to have a smaller effect on the fitness of a protein, and are fixed at a higher rate [1]. At the same time, local rates of change in protein sequence are correlated with local rates of change in protein structure [2–4].

Computer models of protein evolution, which approximate the functional requirements on protein sequences by, for example, folding polymer on a lattice, have been very successful in identifying the causes of evolutionary rate variation in protein sequences [4–11]. However, in most models directed at this problem, the folded structure of a protein is used as a proxy for functional constraints, and consequently changes in folded structure and the effects of specific functional requirements, such as binding to a target ligand, are neglected. In recent work, we began to explore the first of these problems using an off–lattice model in which polymers are evolved to maintain an ordered but otherwise un–restricted folding domain [12, 13]. We found that the model could recapitulate basic properties of protein evolution, such as maintenance of amino acid sequence complexity and solubility of folded structures, linear rates of amino acid change as a function of solvent exposure, and linear divergence of folded structures with the number of accepted mutations. Here, we build on this work, using a model in which polymers are evolved to mimic the behavior

of small enzymes in order to explore the effect of functional constraints on rates of change in protein sequence and structure.

In the revised model, polymers evolve as a result of mutations to model genes (including insertions and deletions) and are selected to re–configure a specific binding site structure compatible with a model ligand (Fig. 1). Again, we approximate population dynamics by a sequential–fixation process (i.e. the whole population is represented by a single sequence, an approximation that is valid when the time to fixation or loss of a mutation is shorter than the time between fixation events [14], and applies to organisms in the plant and animal kingdoms). We find that the requirement of binding a ligand is, in itself, sufficient to maintain an ordered folding domain [15], and we compare it to the requirement of folding an ordered but otherwise unrestricted domain treated in our earlier work. The model predicts that amino acid transition rates vary linearly with distance from the ligand binding site, and distinguishes between exposure, stress, and flexibility models of evolutionary rates [16, 17]. Evolutionary rates fluctuate significantly, both within and among model lineages, in contrast to the molecular clock (Poisson) hypothesis, but consistent with sophisticated analyses of phylogenetic data [18–20] and the predictions of earlier models [7–9]. Structural change is Lévy–like under both conditions, with long periods of structural stasis punctuated by shorter periods of structural change [13]. However, on average, the amount of change experienced under either condition is linear in the number of accepted mutations, in agreement with recent analyses of protein structure families [21, 22].

In the next section below, we describe the general features of our model and the procedures used to generate our data. Following this section, we describe our results and compare them to the predictions noted above.

II. Model

We approximate protein evolution by a discrete Markov (sequential–fixation) process, in which a single gene representing a population is subject to mutation, and mutated forms of the gene are accepted or rejected as the current state of the population according to one of the following conditions : In condition (i), gene products (model proteins) are required to re–configure an ordered but otherwise un–restricted nucleus, sufficient to support a small binding site against thermal fluctuations ; In condition (ii), gene products are required to re–configure a pre–defined binding site structure forming a small surface cavity compatible with a model ligand (see below).

In each iteration of the Markov process, each nucleotide position in a gene is subject to the possibility of replacement, insertion or deletion, except at positions that code for the binding sites of polymers evolved under condition (ii). Insertions and deletions (indels) are imposed in single codon units. The attempt frequencies for replacement and indel mutations are adjusted to reflect protein data : Nucleotide transitions are favored over transversions by a factor of 2, and multiple replacements within a gene are rare. Insertion and deletion mutations are equally probable, and together occur at about one tenth the rate of non–synonymous replacements. Multiple indel events and non–sense mutations are excluded from the Markov process entirely.

The fitness of a model gene is determined by folding $\mathcal{N} = 127$ replicas of its encoded polymer on a parallel computer and analyzing the resulting ensemble of structures. Folding is initiated from a random coil state below the folding transition temperature of a typical viable sequence evolved under condition (i). The time allowed for folding is determined by the length of a polymer according to an estimate provided by Lin and Zewail [23]. The temperature is reduced substantially, the replicas are equilibrated for a short period, and a final ensemble of structures, Γ, is recovered. The polymer model and

the folding procedure are described in more detail in Appendix I.

In condition (i), a structural alignment is performed for each pair of structures $\mathbf{x}^{\mu}$ and $\mathbf{x}^{\nu\neq\mu}$ contained in Γ. Structures are aligned by rotation, translation, and reflection through the closest $2N/3$ pairs of corresponding monomer positions, where N is the number of monomers in a polymer. The alignments are used to select a smaller ensemble, $\Delta\Gamma^{\star}$, to define the dominant energy basin recovered by the folding procedure, and a reference fold, $\mathbf{x}^{\star}$, closest to the center of the ensemble $\Delta\Gamma^{\star}$ (Fig. 1). The fitness of a sequence is defined by the degree of structural order in the ensemble $\Delta\Gamma^{\star}$; Let $\|\mathbf{x}_j^{\star} - \mathbf{x}_j^{\mu}\|$ denote the structurally aligned distance between monomer positions $\mathbf{x}_j^{\mu}$ and $\mathbf{x}_j^{\star}$. To measure structural order, we compute the mean–square distance,

$$\lambda_j^2 = \left\langle \|\mathbf{x}_j^{\star} - \mathbf{x}_j^{\mu}\|^2 \right\rangle_{\mu} \tag{1}$$

averaged over structures $\mathbf{x}^{\mu} \in \Delta\Gamma^{\star}$. A monomer is considered ordered when $\lambda_j \leq \lambda^{\dagger}$, analogous to the Lindemann melting criterion [24–26], where $\lambda^{\dagger} \sim 0.15\, l$, and $l = 3.8$ Angstroms is the length of a polymer link. A sequence folding an ensemble with at least 15 ordered monomers is accepted, otherwise it is rejected. The parameter $\lambda^{\dagger}$, the alignment method, and the procedure for selecting the ensemble $\Delta\Gamma^{\star}$ are described in more detail in Appendix II.

In condition (ii), simulations are initialized by a sequence evolved under condition (i) that spontaneously forms an ordered, hydrophilic surface cavity compatible with a model ligand (here, a monomer or dimer). A model ligand is optimally 'docked' to properly formed binding sites in the initial ensemble in order to define the "active state" of a folded replica. In subsequent steps of the Markov process, a folded replica is considered "active" when the distances between monomers in the binding site (including the target ligand)

are each within 1 Angstrom of the average distances between corresponding monomers in properly formed binding sites in the initial ensemble. The fitness of a sequence is then defined by the number of folded replicas that satisfy this condition. If the number of active replicas is greater than $3\mathcal{N}/4$, the mutation is accepted, otherwise it is rejected. These procedures are described in more detail in Appendix III.

III. Results

To explore the behavior of the model under condition (ii), we generated 3 star phylogenies [27], each phylogeny consisting of 5 lineages evolved from one of 3 initial sequences (see Supplemental Material). Each initial sequence is selected to re–configure an ordered binding site defined by 3–4 weakly attractive, or repulsive amino acid types, similar to those found in the active sites of small enzymes [28]. To compare fitness conditions, we also generated 3 phylogenies under condition (i) with the same initial sequences. Polymers evolved under condition (ii) maintain an ordered nucleus, similar to that required explicitly under condition (i), enclosed in a "halo" of disordered hydrophilic loops. The typical length of a polymer in our sample is about 35 monomers, and the typical length of a disordered loop is between 1–3 monomers, as in Fig. 1.

Due to the inclusion of indels in the model, it is necessary to align the sequences along each lineage. Amino acid transition probabilities and rates are determined by counting the number of transitions of a given type along columns of an alignment (alignments are exact here except when indels occur within coding regions). Transition (exchange) probabilities are in relatively good agreement with protein phylogenetic data (Fig. 2), consdering that many infrequent transitions contributing to protein phylogenies are excluded here by the genetic code. Transition rates are linearly correlated with exposed surface area (Fig. 3),

in agreement with well known results for proteins [29–31]. To compute exposed surface area, monomers are viewed as interpenetrating spheres, each coated with a large number of equally spaced points [32]. The fraction of exposed surface area, $\delta\mathcal{A}_j/\mathcal{A}$, is measured as the fraction of points coating a monomer that are not enclosed in another sphere in the reference structure of a given sequence [12]. Transition rates, $\omega(\delta\mathcal{A})$, are measured as the number of transitions from monomers with exposure $\delta\mathcal{A}/\mathcal{A}$ divided by the amount of time (i.e., iterations of the Markov process) that monomers with exposure $\delta\mathcal{A}/\mathcal{A}$ are exposed to mutation. Fig. 3 describes the results obtained for lineages evolved under condition (i) ; The results for condition (ii) are similar.

In recent work, Echave et al. have suggested that local packing density, or stress can provide a more accurate description of amino acid transition rates [1]. In Fig. 4, we compute transition rates, $\omega(\mathcal{Q})$, as a function of the weighted contact measure of packing density,

$$\mathcal{Q}_i = \sum_{j \neq i} |\mathbf{x}_i - \mathbf{x}_j|^{-1} \tag{2}$$

for polymers evolved under condition (i). The linear form of the data for $\omega(\mathcal{Q})$ in Fig. 4 supports the stress model of transition rates proposed by Huang et al. [17] ; A plot of $\omega(1/\mathcal{Q})$ for the same set of lineages yields a curvilinear plot, in contrast to the linear relationship expected for rate dependence on structural flexibility (not shown). Again, similar results are obtained under condition (ii). Including the data for both conditions, transition rates exhibit slightly stronger correlations with $\mathcal{Q}$ than with $\delta\mathcal{A}/\mathcal{A}$, in agreement with empirical results of Yeh et al. [16] (see Supplemental Material).

Finally, in Fig. 5 we compute transition rates $\omega(\mathcal{R})$ as a function of distance $\mathcal{R}$ from the center of the binding complex. The linear increase of $\omega(\mathcal{R})$ with $\mathcal{R}$ is consistent with the anzatz used by Dean et al. [33] to identify causes of rate variation among sites in protein

structures (this prediction was verified expirically by Jack et. al during publication [34]).

It is well known that evolutionary rates fluctuate within protein lineages, in contrast to the molecular clock (Poisson) hypothesis [18–20]. This problem, known as "over–dispersion", or the "over–dispersed clock", has inspired relaxed methods of phylogeny re–construction in which mutation rates are allowed to vary among the branches of phylogenetic trees [35]. Bastolla et. al and Wilke have argued that over–dispersion in proteins can be explained in large part by the requirement of folding a protein into its functional structure [7–9]. Since our model includes a number of features neglected in these early models (such as explicit folding, fold change, and explicit functional requirements), it is of interest to re–examine this problem.

Fig.s 6–8 describe the structure of a typical lineage evolved under condition (ii), where $n^l(\tau)$ denotes the number of mutations accepted along a lineage l at time τ (measured in iterations of the Markov process), $P(\mathcal{T} \geq \tau)$ is the counter–cumulative distribution of waiting times $\mathcal{T}$ between accepted mutations (including indels), and λ is the nuclear Lindemann parameter, defined as

$$\lambda^2 \, = \, \frac{1}{N_{\|}} \left\langle \|\mathbf{x}^{\star} \, - \, \mathbf{x}^{\mu}\|^2 \right\rangle_{\mu} \tag{3}$$

where $N_{\|}$ is the number of monomers compared in structure alignments (here, $N_{\|} = 2N/3$ as in Eq. (1)). As is evident in Fig. 6, acceptance rates vary significantly along the lineage. Waiting times fit closely to an asymptotic power law (i.e. Pareto–like) distribution,

$$P(\mathcal{T} \geq \tau) \, \simeq \, (1 \, + \, \tau/\tau_m)^{-\alpha} \tag{4}$$

in contrast to the Poisson (exponential) distribution predicted for a molecular clock (Fig. 7). At the same time, the "nuclear" Lindemann parameter, $\lambda(\tau)$, remains within the

range of values $\lambda \stackrel{\sim}{<} 0.2\,l$ obtained by Zhou and Karplus [24] in all–atom simulations of folded proteins (Fig. 8). Similar results are obtained under condition (i).

Bastolla et. al have argued that the number of viable amino acid mutations available to a protein fluctuates as it drifts through sequence space, leading to bursts of mutations and periods of relative stasis along protein lineages, sufficient to explain over–dispersion in protein data. To examine these effects, we measure the structure of $n^l(\tau)$ using methods from time series analysis.

In Fig. 9, we describe the structure of $n^l(\tau)$ for each lineage using a simple phase diagram developed by Goh and Barabasi [36] The parameter $\boldsymbol{\beta}$ in this diagram describes the distribution of waiting times ;

$$\boldsymbol{\beta} \;=\; (\varrho \;-\; 1)\,/\,(\varrho \;+\; 1) \tag{5}$$

where $\varrho = \sigma/\mu$ is the coefficient of variation, $\mu = \langle \mathcal{T} \rangle$ is the mean, and $\sigma^2 \;=\; \langle \mathcal{T}^2 \rangle \;-\; \langle \mathcal{T} \rangle^2$ is the variance of the distribution of waiting times along a lineage. By construction, $\boldsymbol{\beta}$ is confined to the interval $\boldsymbol{\beta} \in [-1, 1]$; $\boldsymbol{\beta} = -1$ corresponds to a δ–function distribution, $\boldsymbol{\beta} = 0$ to an exponential distribution, and $\boldsymbol{\beta} = 1$ to the limit of a "fat–tailed" distribution as in Eq. (4) ; The parameter $\boldsymbol{\beta'}$ denotes the correlation function,

$$\boldsymbol{\beta'} \;=\; \langle (\mathcal{T}_i \;-\; \mu)(\mathcal{T}_{i+1} \;-\; \mu) \rangle \;/\; \sigma^2 \tag{6}$$

which describes the tendency for intervals of similar length to cluster along a lineage, where $\mathcal{T}_i$ denotes the waiting time between mutations i and $i + 1$. The shaded region of the diagram roughly indicates the phase space available to a compound Poisson process in which waiting times are selected at random from exponential distributions $\exp(-\lambda_1 \tau)/\lambda_1$ and $\exp(-\lambda_2 \tau)/\lambda_2$ with different rates, $\lambda_2 \neq \lambda_1$ (i.e., as might be expected on account

of the stratification of acceptance rates according to burial, or packing density). Most of the data lies outside this region, indicating a more complex process.

To describe longer range correlations among waiting time intervals, we measure fluctuations in the "height" function [37],

$$\mathcal{Y}_j = \sum_{i=1}^{j} \mathcal{T}_i - \langle \mathcal{T} \rangle \tag{7}$$

by the height auto–correlation function,

$$\langle \Delta \mathcal{Y}^{\mathsf{q}} \rangle = \langle |\mathcal{Y}_j - \mathcal{Y}_{j+k}|^{\mathsf{q}} \rangle_j \tag{8}$$

for $\mathsf{q} \leq 2$. The height auto–correlation function $\langle \Delta \mathcal{Y}^{\mathsf{q}} \rangle$ measures fluctuations in the amount of time accumulated along intervals of length k accepted mutations against the mean value, $k \langle \mathcal{T} \rangle$. For an uncorrelated process in which, for example, waiting times are selected at random from a Pareto–like distribution, fluctuations scale as $\langle \Delta \mathcal{Y}^{\mathsf{q}} \rangle \propto k^{\mathsf{q}\mathsf{H}}$ with $\mathsf{H} = 1/2$ where H is the Hölder roughness exponent [37]. The smooth scaling of fluctuations with k indicates the nesting of intervals with more rapid activity, as in a fractal, or self–affine pattern ; More rapid scaling (i.e., a Hölder exponent $\mathsf{H} > 1/2$) indicates persistent correlations between increment lengths, and, consequently, in the fraction of neutral mutations available to an evolving sequence ("multi–fractal" scaling, $\mathsf{H} = \mathsf{H}(\mathsf{q})$, is thought to indicate "multi–affine" structure). For $\mathsf{q} = 1$, we find that fluctuations exhibit uniform scaling over the range $k \stackrel{\sim}{<} 64$ yielding exponents $\mathsf{H}(1) \simeq 0.6 - 0.9$. For $\mathsf{q} = 2$, the data becomes rugged, and it is no longer accurate to describe the waiting time series in terms of an exponent ; To avoid this problem, we average the correlation functions obtained for different lineages. Fig. 10 describes the average over lineages evolved under condition (ii) for $\mathsf{q} = 1$ and $\mathsf{q} = 2$; In this case, we obtain

exponents $\mathsf{H}(1) \simeq 0.8$ and $\mathsf{H}(2) \simeq 0.7$ by fitting to the initial range of the data, $k \stackrel{\sim}{<} 64$. These results, in particular, the rollover in the data for $k \stackrel{\sim}{>} 64$, are consistent with those of Bastolla et. al, who used a similar approach to measure fluctuations in the number of neutral mutations available to an evolving sequence [8]. Results for lineages evolved under condition (i) are similar to those in Fig. 10.

Finally, to compare variations in acceptance rate among model lineages to phylogenetic data, we compute the index of dispersion, or ratio of the variance to the mean number of mutations

$$I(\tau) \, = \, \left\langle (n^l(\tau))^2 \right\rangle_l \, - \, \left\langle n^l(\tau) \right\rangle_l^2 \, / \, \left\langle n^l(\tau) \right\rangle_l \tag{9}$$

in star phylogenies [18]. In Fig. 11, we describe the results of this calculation for a pair of phylogenies evolved from the same initial sequence under (A) condition (i) and (B) condition (ii). In each panel of Fig. 11, we measure $I(\tau)$ for whole sequences, and for amino acid positions that are homologous (aligned) to positions in the ancestral sequence – dashed lines roughly indicate the point at which the number of accepted mutations per position, per lineage is $\left\langle n^l/N \right\rangle_l \sim 1$. Typically, only a few deletions occur along any lineage within the range considered in the figure, and consequently the number of homologous positions remains roughly constant. However, both phylogenies contain lineages in which insertions are acquired in rapid succession ; The restriction to homologous positions is intended to exclude replacement mutations at these (inserted) positions, and more accurately resembles the situation encountered in protein alignments. Interestingly, $I(\tau)$ remains essentially unaltered by this restriction in phylogenies evolved under condition (ii). In general, we obtain index values in the range $I(\tau) \sim 1 - 7$ on time scales for which $\left\langle n^l(\tau) \right\rangle_l \stackrel{\sim}{<} 20$, in agreement with sophisticated estimates for protein phylogenies [18–20] ; For synonymous mutations, we obtain $I(\tau) \sim 1 - 2$.

To conclude our study, we measure the average rate of structural change along model lineages, and we analyze the patterns of structural increments, or flights between reference structures, using methods similar to those in Eq. (8). We first present our results for lineages evolved under condition (i) and then summarize our results for condition (ii).

In order to measure structural distance, it is first necessary to establish a homology between monomers in structures compared along a given lineage ; Let $\mathbf{s}(\tau)$ denote the (gapped) sequence at time τ in an alignment, and let $\mathbf{x}(\tau)$ denote its corresponding reference structure (for simplicity, we omit the superscript on reference structures in the discussion below). A pair of monomers in $\mathbf{x}(\tau)$ and $\mathbf{x}(\tau')$ are considered homologous when their positions in $\mathbf{s}(\tau)$ and $\mathbf{s}(\tau')$ are aligned. To compute distance, structures are first aligned through homologous positions using the methods described above. The distance $\Delta x(\tau, \tau')$ between structures $\mathbf{x}(\tau)$ and $\mathbf{x}(\tau')$ is defined as the root mean–square distance between the closest $N_{\|} = 20$ aligned monomers (similar to the number of ordered monomers maintained in the folded ensembles of evolved sequences), analogous to the procedure used by Illergard et. al. to measure structural drift in protein families [21]. Similar results are obtained using the Hamming distance between contact matrices formed by compared monomers in each structure [21].

In Fig. 12, we plot the distance from the ancestral fold, $\Delta x(0, \tau)$, for a specific lineage evolved under condition (i). The step–like pattern of the data is typical of lineages evolved under both conditions (see below). The distribution of structural flights between adjacent accepted mutations, $\Delta x(n, n+1)$, along a lineage is Lévy–like on average, resembling a normal distribution with an extended tail (not shown). In Fig. 13, we plot the average distance between structures separated by k accepted mutations, $\langle \Delta x(n, n+k) \rangle_n$, both for the lineage in Fig. 12 and for the average of $\langle \Delta x(n, n+k) \rangle_n$ over lineages evolved under

condition (i). Both sets of data fit accurately to a power–law,

$$\langle \Delta x \rangle \simeq \mathsf{p} + \mathsf{q}\, k^{\alpha} \tag{10}$$

with $\alpha \simeq 1$. Linear scaling is obtained on a slight reduction in the number of monomers compared in structure alignments. For $N_{\|} = 18$, the lineage average converges with the results obtained by Illergard et al. for protein families after adjustment by a constant factor to compensate for the larger number of positions compared in protein alignments (the corrected data lies almost on top of the lineage average in Fig. 13, and is omitted for clarity). For fewer compared monomers, $N_{\|} \leq 17$, structural change remains linear, but it occurs at a slower rate (i.e., a smaller value of q in Eq. 10).

To quantify patterns of structural change in $\Delta x(0, \tau)$, we compute the correlation function Eq. (8), with structural increments, $\Delta x(n, n+1)$, replacing temporal increments in Eq. (7). If structural change is Lévy–like, smaller increments will tend to cluster along a lineage, leading to exponents $\mathsf{H} > 1/2$. For comparison, we repeat this calculation for increments, $\Delta x(0, n+1) - \Delta x(0, n)$, measured against the ancestral structure. In this case, distance increments are signed, and consequently changes in the height function will tend to compensate, or anti–correlate within intervals of structural stasis, leading to exponents $\mathsf{H} < 1/2$. For lineages evolved under condition (i), we find strong positive and negative correlations, with $\mathsf{H} \sim 0.8$ and $\mathsf{H} \sim 0.2$ respectively along individual lineages, consistent with Lévy–like dynamics.

The results for condition (ii) are very similar to those obtained for condition (i) except that intervals of structural stasis are usually longer, and consequently, structures tend to evolve more slowly. The average of $\langle \Delta x(n, n+k) \rangle_n$ over lineages evolved under condition (ii) is still described accurately by Eq. (10) (with linear scaling under the conditions

described above), however, the overall change in distance along a lineage is less than half the value obtained for condition (i), and for a number of lineages, the ordered nucleus of the ancestral fold is completely conserved. Together, the results for conditions (i) and (ii) are consistent with those of Shakhnovich et al. and Pascual–Garcia et al. ; Functional requirements on proteins within protein structure families vary [38], leading to an apparent increase in the measured rate of structural change compared with proteins grouped under a specific functional class [22].

As a final remark, we note that temporal and structural increments along a lineage are, in general, uncorrelated ; Large changes in structure can occur suddenly, as the result of a single mutation, or smoothly, in concert with a burst of mutations. Conversely, a small change in structure resulting from an insertion or deletion can lead to a significant increase in acceptance rate that persists for most of a lineage. There seems to be no simple pattern to how these events occur, except that the majority of change takes place on the surface of a folded structure, and away from the binding site. Thus, what apparently links the behavior of the model with proteins is the tendency for amino acids to organize into protein–like complexions [39] under the pressure to fold an ordered nucleus.

The authors would like to thank one anonymous referee for suggestions which led to the calculations in Fig.s 4–6. This work was supported in part by the National Institutes of Health (GM094575 to NVG)

Appendix I

The polymer model is a chain of point monomers that interact as low resolution amino acids via spherically symmetric potentials. Interactions along the chain are described by potentials of the form,

$$U^{\kappa}(r) \;=\; \frac{\kappa}{2}\,(r-l)^2 \tag{11}$$

where r is the distance between monomers, l is the equilibrium length of a link, and κ is a constant (see below).

Interactions between non–adjacent monomers along the chain are constructed from the unit Morse potential,

$$\mu(r) \;=\; \exp(-2\alpha(r-l)) \;-\; 2\exp(-\alpha(r-l)) \tag{12}$$

The attractive minimum of the Morse potential occurs at $r = l$. Let

$$\mu^{r\leq l}(r) \;=\; \vartheta(l-r)\,\mu(r) \tag{13}$$

and

$$\mu^{r\geq l}(r) \;=\; \vartheta(r-l)\,\mu(r) \tag{14}$$

denote the components of the Morse potential in either side of the minimum, where ϑ is the unit step function. The potentials for attractive and repulsive amino acid interactions are constructed as,

$$U^{\epsilon'\leq 0}(r) \;=\; \epsilon\,\mu^{r\leq l}(r) \;+\; (\epsilon \;+\; \epsilon')\,\vartheta(l-r) \;-\; \epsilon'\,\mu^{r\geq l}(r) \tag{15}$$

and

$$U^{\epsilon' \geq 0}(r) \;=\; \epsilon\, \mu^{r \leq l}(r) \;+\; \epsilon\, \vartheta(l-r) \;+\; \epsilon' \, \exp(-\alpha(r-l)) \tag{16}$$

respectively (Fig. 14).

Each potential consists of an excluded volume part, $\epsilon\, \mu^{r \leq l}(r) \;+\; \epsilon\, \vartheta(l-r)$, modulated by the parameter ϵ, and a sequence dependent part, modulated by the parameter ϵ' ; The parameter ϵ' takes on different values,

$$\epsilon' \;=\; \epsilon\, E_{\mu\nu} / E_{\circ} \tag{17}$$

depending on the amino acid types involved in an interaction, where $E_{\mu\nu}$ is the energy of a contact between amino acids μ and ν defined by the empirical parameters in reference [40], and $E_{\circ} = \langle |E_{\mu\nu \geq \mu}| \rangle$ is the average strength of an interaction (the empirical parameters are obtained by re–scaling the Miyazawa–Jernigan parameters [41] using threonine as a reference solvent [40]). The potentials for unit strength attractive and repulsive interactions are plotted in Fig. 14.

To describe polymer kinetics, we integrate the Langevin equation using the method of van Gunsteren and Berendsen [42], with monomer mass $m = 1.66 \cdot 10^{-22}\,\mathrm{g}$, friction coefficient $\gamma = 10\,\mathrm{ps}^{-1}$, and integration time step, $\Delta t = 0.01\,\mathrm{ps}$. The parameters used to define the potentials are $l = 3.8$ Angstroms, $\kappa = 11\, k_B T_0$, $\alpha = 2.1$ Angstroms^{-1}, and $\epsilon = 2\, k_B T_0$, where k_B is Boltzmann's constant and $T_0 = 302.15$ Kelvin°.

Folding is initiated from a random coil state below the folding transition temperature of a typical evolved sequence, which we estimate as $T_f \sim 1.25\, T_0$ from specific heat data. The time allowed for folding is determined by the length of the polymer according to the

estimate of Lin and Zewail [23],

$$t_f \, = \, N \left(\frac{3}{e} \right)^N \, \Delta t_f \tag{18}$$

where $\Delta t_f \, = \, 10$ ps roughly describes the timescale for positional exchanges among monomers on the surfaces of polymer nuclei. Following this step, the replicas are equilibrated for a short time $t_q = t_f/3$ at temperatures $T_1 = 218.2$ Kelvin° and $T_2 = 134.3$ Kelvin°.

Appendix II

To apply condition (i), each structure $\mathbf{x}^\mu \in \Gamma$ is considered as a possible reference structure, and each of the remaining structures, $\mathbf{x}^{\nu \neq \mu}$, are aligned to $\mathbf{x}^\mu$ by rotation, translation, and reflection. Because the surfaces of folded polymers are disordered, we compute structural distance by using an iterative procedure which ultimately aligns the closest $2N/3$ pairs of monomers : Let $\mathbb{A}$ denote the sequence positions of monomers compared in an alignment, initially including all positions. In each iteration of the alignment procedure, structures are aligned to minimize the squared distance

$$|\mathbf{x}^\mu \, - \, \mathbf{x}^\nu|^2_{\mathbb{A}} \, = \, \sum_{j \in \mathbb{A}} \left(\mathbf{x}^\mu_j \, - \, \mathbf{x}^\nu_j \right)^2 \tag{19}$$

The index of the most distant monomer pairing in $\mathbb{A}$ is then removed, and the process is repeated until $2N/3$ optimally aligned pairs of monomers remain.

Let $\|\mathbf{x}^\mu \, - \, \mathbf{x}^\nu\|$ denote the distance obtained in the last iteration of this procedure – i.e., the distance measured by the closest $2N/3$ pairs of monomers. To measure the accuracy of a multiple alignment with a given structure, $\mathbf{x}^\mu$, we compute the Lindemann

parameter,

$$\lambda(\mathbf{x}^{\mu}) \;=\; \left[\frac{3}{2N} \frac{4}{3\mathcal{N}} \sum_{\nu} \|\mathbf{x}^{\mu} \,-\, \mathbf{x}^{\nu}\|^2 \right]^{1/2} , \tag{20}$$

or average distance between compared monomers, where the sum is restricted to the closest $3\mathcal{N}/4$ structures in Γ, which we denote by $\Delta\Gamma^{\mu}$. The reference structure, $\mathbf{x}^{\star}$, and the corresponding ensemble, $\Delta\Gamma^{\star}$, are defined by the multiple alignment that leads to the minimal value of $\lambda(\mathbf{x}^{\mu})$.

Let $\|\mathbf{x}_j^{\star} - \mathbf{x}_j^{\nu}\|$ denote the distance between a pair of monomers in an iterative alignment (but not restricted to the set of compared monomers). To measure the degree of order for a particular sequence position, we compute the monomeric Lindemann parameter,

$$\lambda_j \;=\; \left[\frac{4}{3\mathcal{N}} \sum_{\nu} \|\mathbf{x}_j^{\star} \,-\, \mathbf{x}_j^{\nu}\|^2 \right]^{1/2} . \tag{21}$$

where the sum is restricted to structures in $\Delta\Gamma^{\star}$. A sequence position (monomer) is considered ordered when $\lambda_j \,\leq\, \lambda^{\dagger}$. Normally, the value of $\lambda^{\dagger}$ is considered constant, however, here the radius of a folded polymer can change along a lineage, which affects the inherent accuracy of an alignment. To account for this effect, we define the melting point threshold by a function $\lambda^{\dagger}(N)$ that scales with the radius of gyration of a collapsed polymer [43],

$$\alpha\, \lambda^{\dagger}(N) \;=\; -4.54 \,+\, 2.36 \left(\frac{2N}{3} \right)^{1/3} . \tag{22}$$

Here, the factor of $2/3$ accounts for the number of monomers compared in structure alignments, and the parameter α is selected so that $\lambda^{\dagger}(30) \,=\, 0.16\, l$. The other constants in this expression are identical to those suggested by Maiorov and Crippen to define the threshold for meaningful comparisons in protein alignments. [44].

Appendix III

To apply condition (ii), it is necessary to dock the target ligand onto the binding site structures re-configured by replicas in the folding procedure. In order to accomplish this, the folded structure of a replica is enclosed in a spherical shell consisting of $\sim 10^4$ evenly distributed points [32]. We then measure, and record the energy of the target ligand (for the moment, a single monomer) at each point on this shell. In this procedure, interactions with monomers in the binding site group are considered attractive, and are described by unit Morse potential, $\mu(r)$, while interactions with monomers not included in the binding site group are described by the repulsive core of the Morse potential, $\mu^{r\leq l}(r)$. The radius of the shell is reduced, and the energies are re–computed at each point, iteratively, until the shell lies inside the folded replica. The structure of the binding site complex (i.e. binding cavity plus ligand) is determined from this sweep as the configuration with minimal energy, and a docked complex is considered active when it meets the conditions described in the text. Compound ligands are treated in a similar way, with each monomer in the ligand docked to its respective binding partners individually. To compute Lindemann parameters for folded ensembles, the structures are aligned by the procedure used for condition (i).

References

[1] J. Echave, S. Spielman, and C. O. Wilke, Nat. Rev. Genet. **17**, 109 (2016).

[2] L. Mirny and E. Shakhnovich, J. Mol. Biol. **308**, 123 (2001).

[3] J. L. England and E. I. Shakhnovich, Phys. Rev. Lett. **90**, 218101 (2003).

[4] G. Tiana, N. V. Dokholyan, R. A. Broglia, and E. I. Shakhnovich, J. Chem. Phys. **121**, 2381 (2004).

[5] J. D. Bloom, C. O. Wilke, F. H. Arnold, and C. Adami, Biophys. J. **86**, 2758 (2004).

[6] J. D. Bloom, A. Raval, and C. O. Wilke, Genetics **175**, 255 (2007).

[7] U. Bastolla, H. E. Roman, and M. Vendruscolo, J. Theor. Biol. **200**, 49 (1999).

[8] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, Phys. Rev. Lett. **89**, 208101 (2002).

[9] C. O. Wilke, BMC Genetics **5**, 25 (2004).

[10] A. E. Lobkovsky and E. V. Koonin, Proc. Natl. Acad. Sci. USA **107**, 2983 (2010).

[11] A. E. Lobkovsky and E. V. Koonin, PLoS Comp. Biol. **7**, e1002302 (2011).

[12] E. D. Nelson and N. V. Grishin, Phys. Rev. E **90**, 062715 (2014).

[13] E. D. Nelson and N. V. Grishin, Phys. Rev. E **91**, 060701 (2015).

[14] D. M. McCandlish and A. Stoltzfus, Q. Rev. Biol. **89**, 225 (2014).

[15] S. Saito, M. Sasai, and T. Yomo, Proc. Natl. Acad. Sci. USA **94**, 11324 (1997).

[16] S. Yeh *et al.*, Mol. Biol. Evol. **31**, 135 (2013).

[17] T. Huang, M. Marcos, J. Hwang, and J. Echave, BMC Evol. Biol. **14**, 78 (2014).

[18] J. H. Gillespie, Proc. Natl. Acad. Sci. USA **81**, 8009 (1984).

[19] J. H. Gillespie, Mol. Biol. Evol. **6**, 636 (1989).

[20] D. J. Cutler, Mol. Biol. Evol. **17**, 1647 (2000).

[21] K. Illergard, D. H. Ardell, and A. Elofsson, Proteins **77**, 499 (2009).

[22] A. Pascual-Garcia *et al.*, Proteins **78**, 181 (2010).

[23] M. M. Lin and A. H. Zewail, Proc. Natl. Acad. Sci. USA **109**, 9851 (2012).

[24] Y. Zhou, D. Vitkup, and M. Karplus, J. Mol. Biol. **285**, 1371 (1999).

[25] H. Jang, C. K. Hall, and Y. Zhou, Biophys. J. **82**, 646 (2002).

[26] R. A. la Violette and F. H. Stillinger, J. Chem. Phys. **83**, 4079 (1985).

[27] M. Kimura, *The neutral theory of evolution* (Cambridge University Press, New York, NY, 1983).

[28] G. J. Bartlett, C. T. Porter, N. Borkakoti, and J. M. Thornton, J. Mol. Biol. **324**, 105 (2002).

[29] E. A. Franzosa and X. Xia, Mol. Biol. Evol. **26**, 2387 (2009).

[30] E. A. Franzosa and X. Xia, PLoS One **7**, e46602 (2012).

[31] D. C. Ramsey, M. P. Scherrer, T. Zhou, and C. O. Wilke, Genetics **188**, 479 (2011).

[32] M. Tegamark, ApJ Lett. **470**, L81 (1996).

[33] A. M. Dean, C. Neuhauser, E. Grenier, and B. Golding, Mol. Biol. Evol. **19**, 1846 (2000).

[34] B. R. Jack, A. G. Meyer, J. Echave, and C. O. Wilke, PLoS Biol. **14**, e1002452 (2016).

[35] O. G. Pybus, PLoS Comp. Biol. **4**, e151 (2006).

[36] K. Goh and A. L. Barabási, Europhys. Lett. **81**, 48002 (2008).

[37] A. L. Barabási, P. Sépfalusy, and T. Vicsek, Physica A **178**, 17 (1991).

[38] B. E. Shakhnovich, N. V. Dokholyan, C. DeLisi, and E. I. Shakhnovich, J. Mol. Biol. **326**, 1 (2003).

[39] S. Hormoz, Sci. Rep. **3**, 2919 (2013).

[40] M. R. Betancourt and D. Thirumalai, Prot. Sci. **8**, 361 (1999).

[41] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[42] W. F. van Gunsteren and H. J. C. Berendsen, Mol. Phys. **45**, 637 (1982).

[43] A. Milchev, W. Paul, and K. Binder, J. Chem. Phys. **99**, 4786 (1993).

[44] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **235**, 625 (1994).

[45] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, *Atlas of protein sequence and structure* (Nat. Biomed. Res. Found., Washington, DC, 1972), pp. 345–352.

Figure Captions

Fig. 1. (Color online) (A) Folded structure, $\mathbf{x}^\star$, and (B) sample of the folded ensemble $\Delta\Gamma^\star$ recovered by a sequence evolved under ligand binding conditions (see Section II). Amino acids (monomers) are colored blue, light blue, blue–green, green, yellow, orange, and red, in order of increasing affinity to solvent. The binding site monomers and the target ligand (here, a single monomer) are colored black. Ordered binding sites evolve spontaneously under condition (i), and are selected to resemble the active sites of small enzymes for subsequent evolution under condition (ii) (see Section II).

Fig. 2. (Color online) Amino acid transition (exchange) probabilities, $p(\mu,\nu) = A_{\mu\nu}/\sum_\nu A_{\mu\nu}$, for sequences evolved under condition (i), where $A_{\mu\nu}$ is the number of transitions recorded between amino acids μ and ν. The initial state of a transition is indicated along the lower axis. Model values are indicated by filled red circles. Empirical values obtained from the data of Dayhoff et al. [45] are indicated by open blue circles. The value of $p(\mu,\nu)$ is indicated by the radius of the corresponding circle.

Fig. 3. (Color online) Amino acid transition rate, $\omega(\delta\mathcal{A})$, versus exposed surface area $\delta\mathcal{A}$ for sequences evolved under condition (i). $\omega(\delta\mathcal{A})$ is plotted in units of 10^{-4} accepted mutations per time step. The dashed line is a fit to the data in the region $\delta\mathcal{A}/\mathcal{A} \leq 0.8$.

Fig. 4. (Color online) Amino acid transition rate, $\omega(\mathcal{Q})$, versus local packing density $\mathcal{Q}$ for sequences evolved under condition (i). $\omega(\mathcal{Q})$ is plotted in units of 10^{-4} accepted mutations per time step ; $\mathcal{Q}$ is measured in units of Angstroms^{-2}. The dashed line is a fit to the region $0.4 \leq \mathcal{Q} \leq 1.2$.

Fig. 5. (Color online) Amino acid transition rate, $\omega(\mathcal{R})$, versus distance from the binding site, $\mathcal{R}$, for sequences evolved under condition (ii). $\omega(\mathcal{R})$ is plotted in units of 10^{-4} accepted mutations per time step ; $\mathcal{R}$ is measured in Angstroms. The dashed line is a fit to the region $\mathcal{R} \leq 12.5$.

Fig. 6. (Color online) Number of mutations, $n^l(\tau)$, accepted along typical lineage, l, evolved under condition (ii). Thick grey segments denote events in the tail of the waiting time distribution $P(\mathcal{T} \geq \tau)$.

Fig. 7. (Color online) Distribution of waiting times, $P(\mathcal{T} \geq \tau)$, for the lineage described in Fig. 6 (circles). The dashed line is a fit to the Pareto distribution in Eq. (4). The solid line is a fit to a Poisson (exponential) distribution. The tail of $P(\mathcal{T} \geq \tau)$ is indicated by the thick grey dashed line.

Fig. 8. (Color online) Plot of the Lindemann parameter, $\lambda(\tau)$, for the lineage described in Fig. 6. Thick grey segments denote events in the tail of the waiting time distribution $P(\mathcal{T} \geq \tau)$.

Fig. 9. (Color online) Phase diagram defined by Eq.s (5) and (6). Lineages evolved under condition (i) are represented by circles. Lineages evolved under condition (ii) are represented by squares. The shaded region roughly indicates the phase space available to a 2–state Poisson process.

Fig. 10. (Color online) Height auto–correlation function, $\langle \Delta \mathcal{Y}^1 \rangle$, for a typical lineage (squares), and averages of $\langle \Delta \mathcal{Y}^1 \rangle$ and $\langle \Delta \mathcal{Y}^2 \rangle$ over all lineages (circles) evolved under condition (ii). Dashed lines are linear fits to the data in the range $k \leq 64$ (the logarithms are base 10). Exponents, $\mathsf{H}(\mathsf{q})$, obtained from the fits are provided in the text.

Fig. 11. (Color online) Index of dispersion, $\mathcal{I}(\tau)$, for a typical pair of phylogenies evolved under (A) condition (i), and (B) condition (ii). The dotted line roughly indicates the point at which the average number of accepted mutations per position, per lineage is $\left\langle n^l/N \right\rangle_l \sim 1$. The lower curves (red) describe the restriction to homologous positions discussed in the text.

Fig. 12. (Color online) Distance from the ancestral fold, $\Delta x(0, \tau)$, for a typical lineage evolved under condition (i).

Fig. 13. (Color online) Average distance, $\langle \Delta x(n, n+k) \rangle_n$, between structures separated by k accepted mutations for the lineage in Fig. 12 (squares), and average of $\langle \Delta x(n, n+k) \rangle_n$ over all lineages (circles) evolved under condition (i). Dashed lines are fits to the data according to Eq. (10) yielding exponents, $\alpha \simeq 0.9$ and $\alpha \simeq 0.8$ respectively. Linear scaling of the lineage average is obtained by a slight reduction in the number of monomers compared in structural alignments, as discussed in the text.

Fig. 14. (Color online) Potential functions, $U^{\epsilon'}(r)$, for cross-chain interactions at unit core strength, $\epsilon = 1$, unit attraction, $\epsilon' = -1$ (dot–dashed line) and unit repulsion, $\epsilon' = 1$ (dashed line).

A

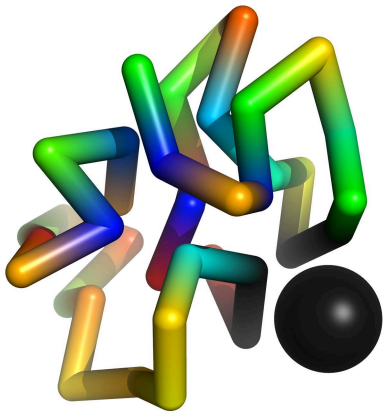

Figure 1A 25Jul2016

B

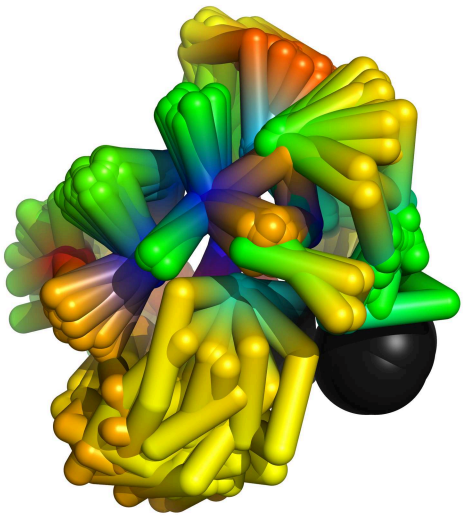

Figure 1B 25Jul2016

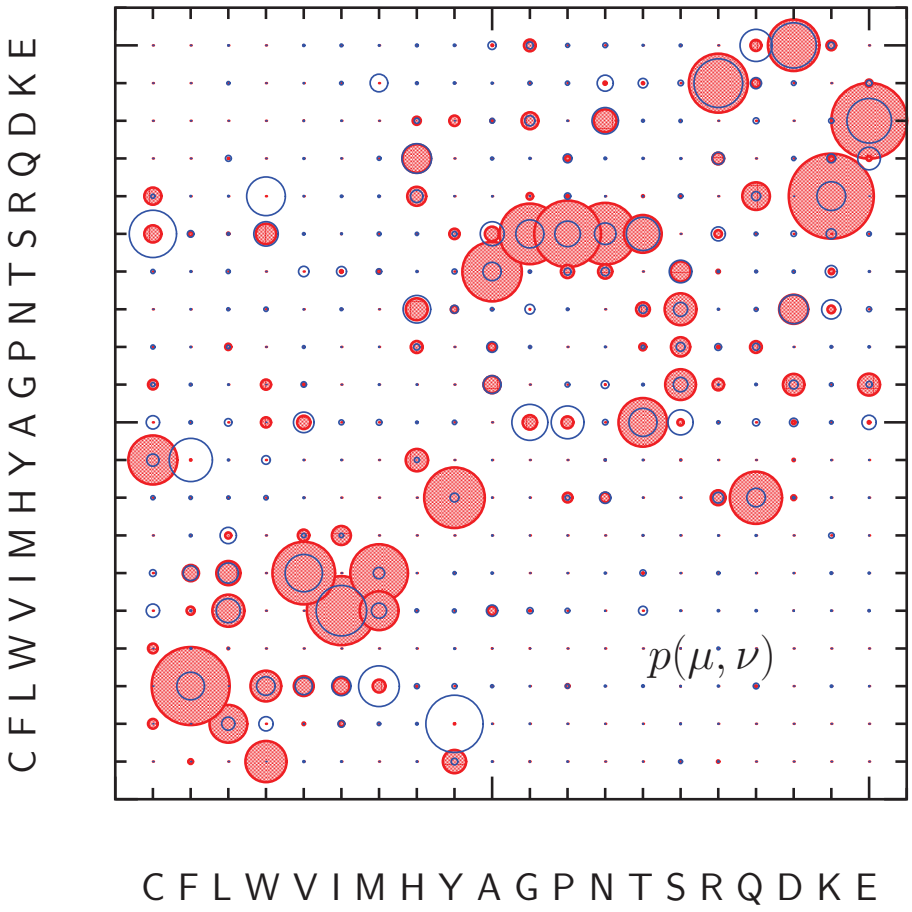

Figure 2 25Jul2016

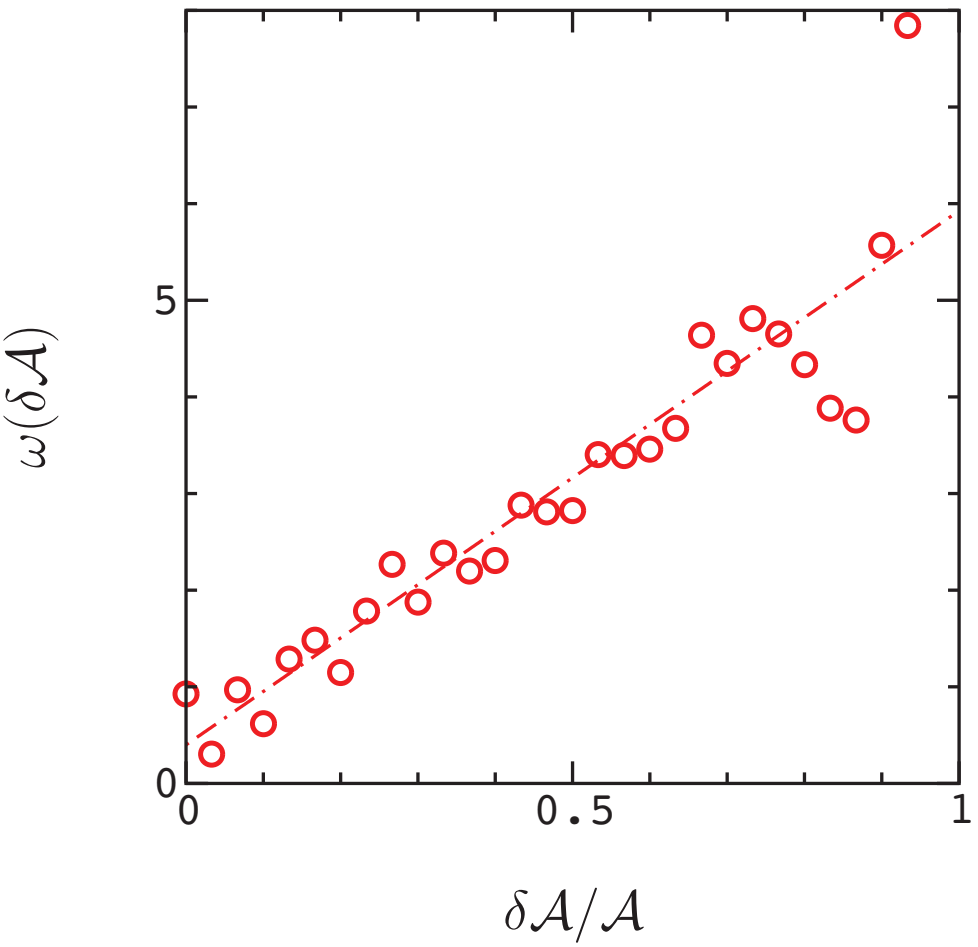

Figure 3 25Jul2016

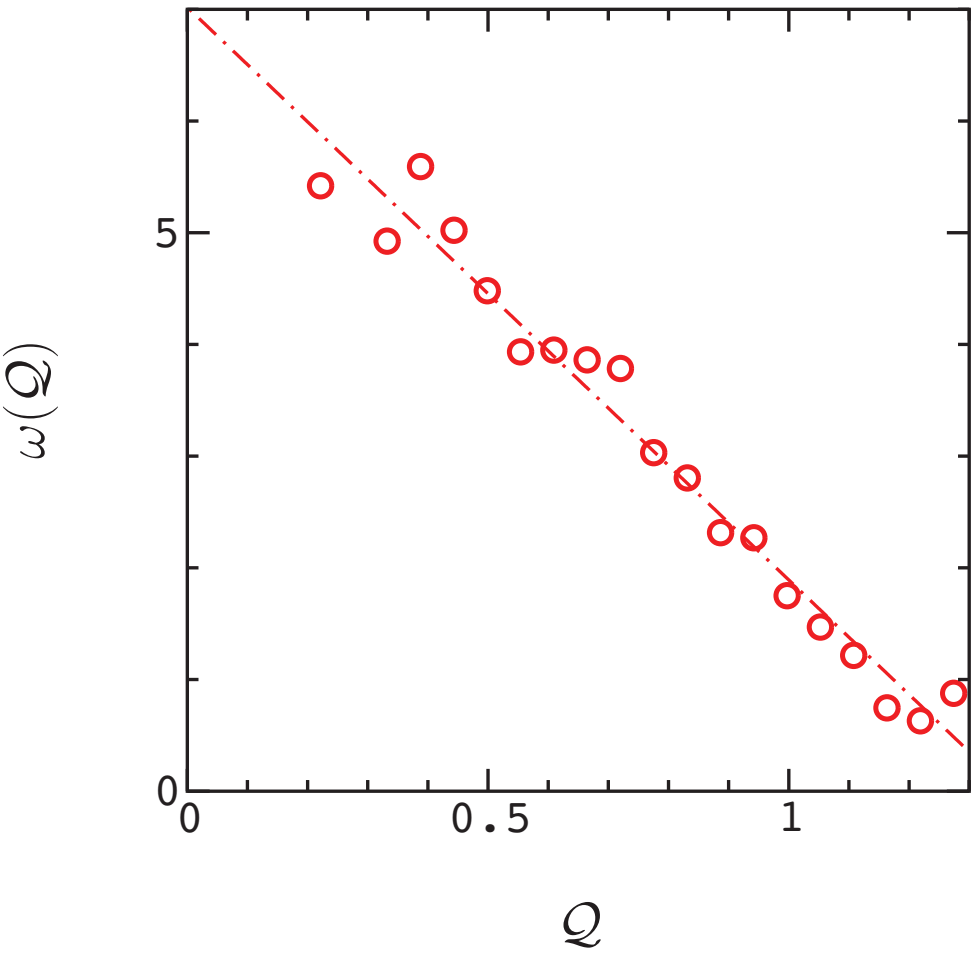

Figure 4 25Jul2016

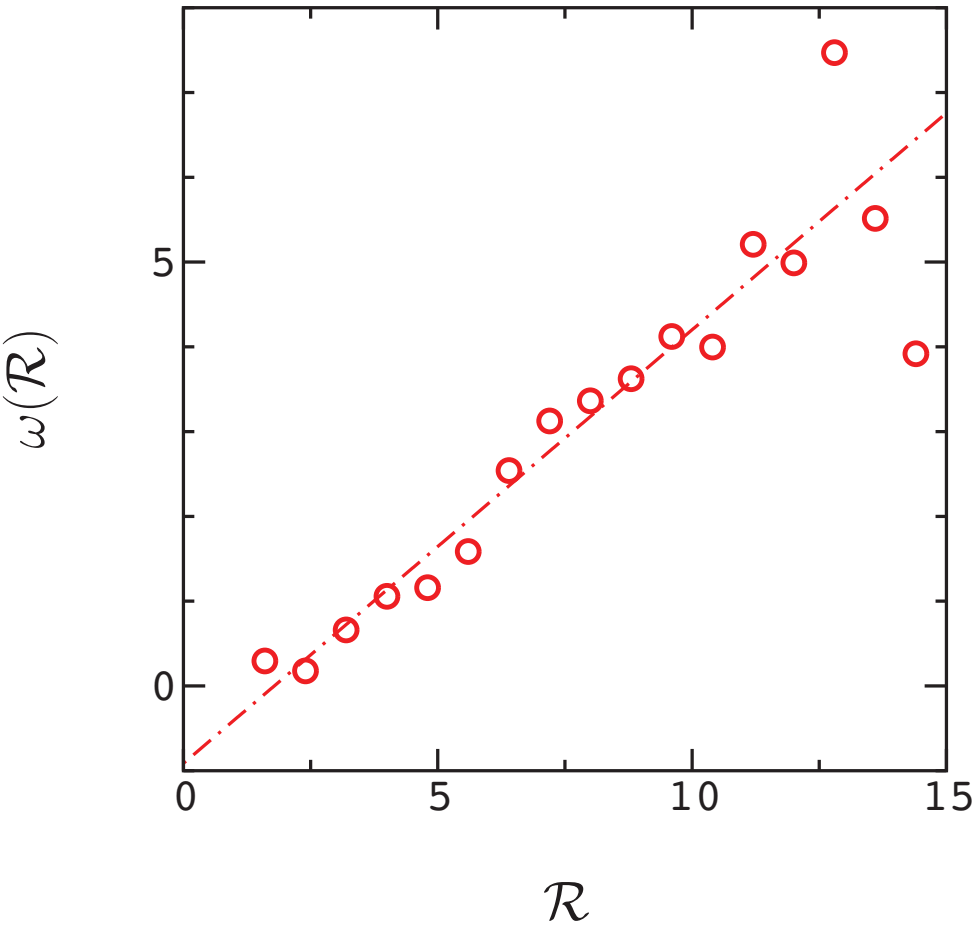

Figure 5 25Jul2016

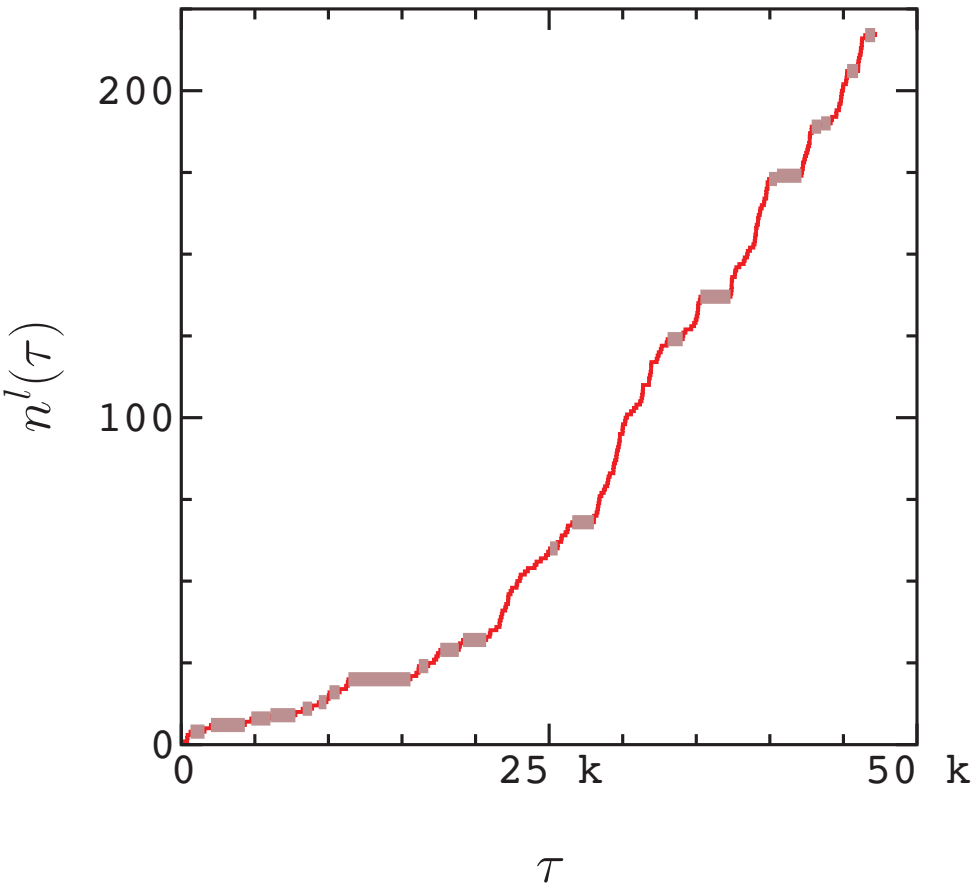

Figure 6 25Jul2016

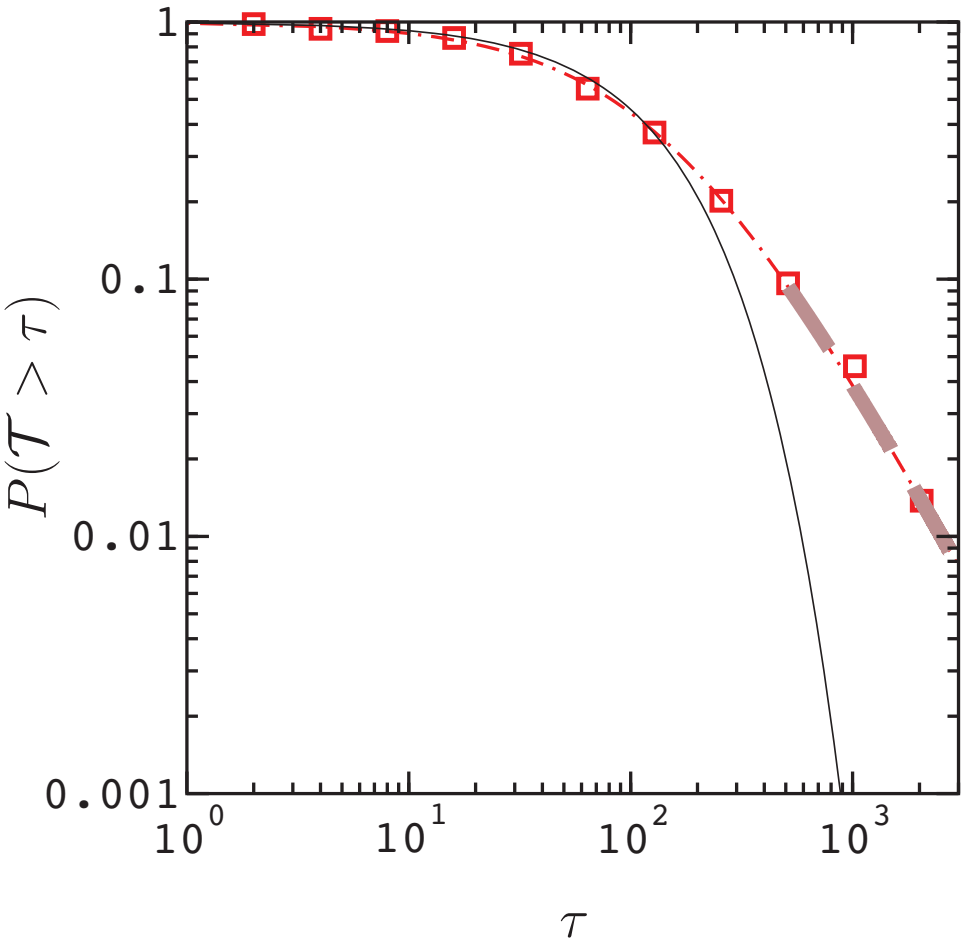

Figure 7 25Jul2016

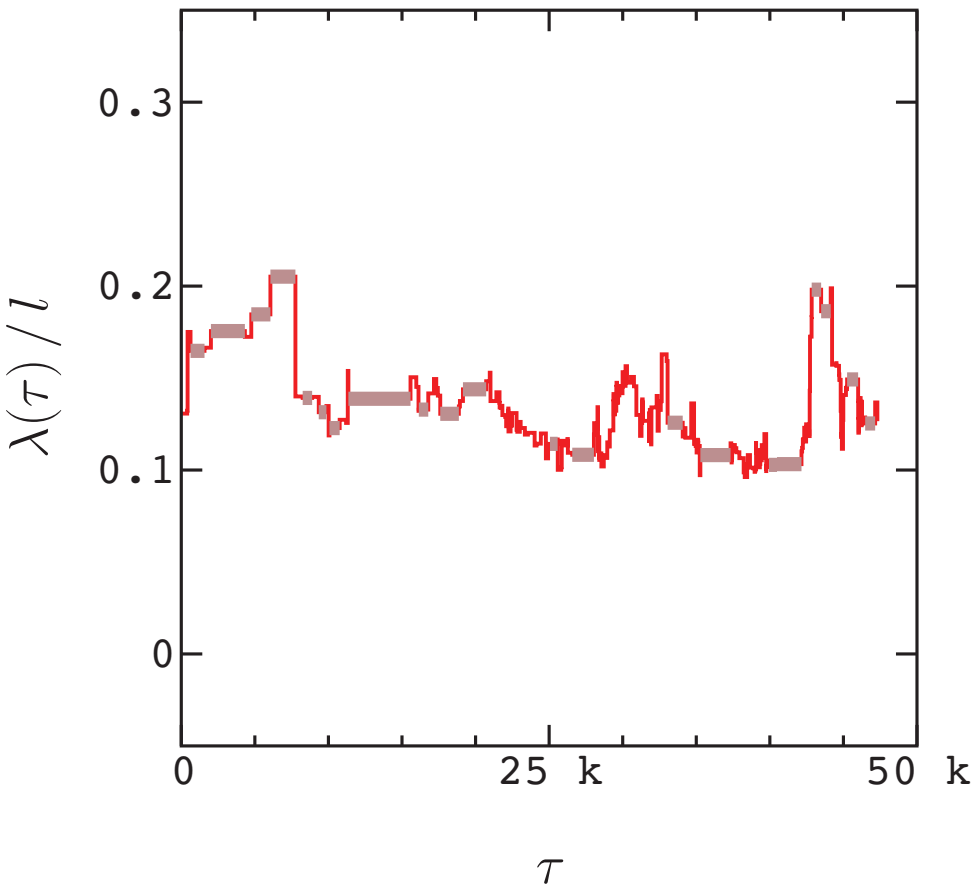

Figure 8 25Jul2016

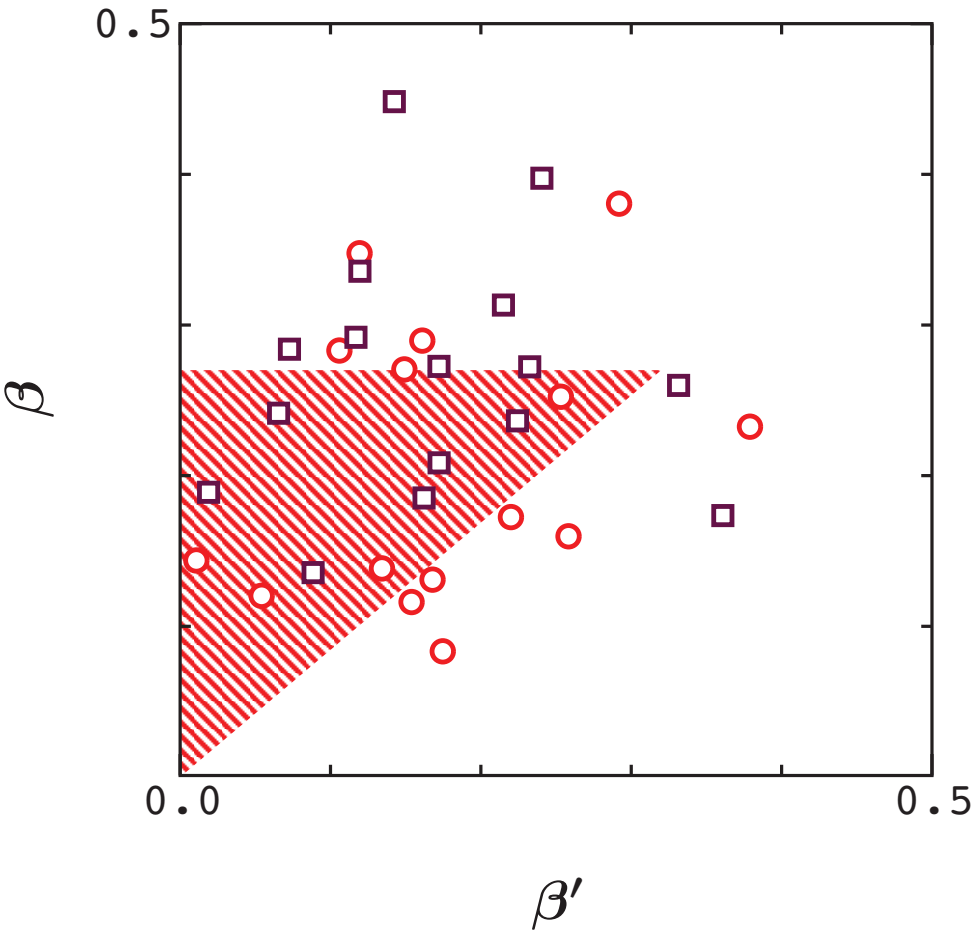

Figure 9 25Jul2016

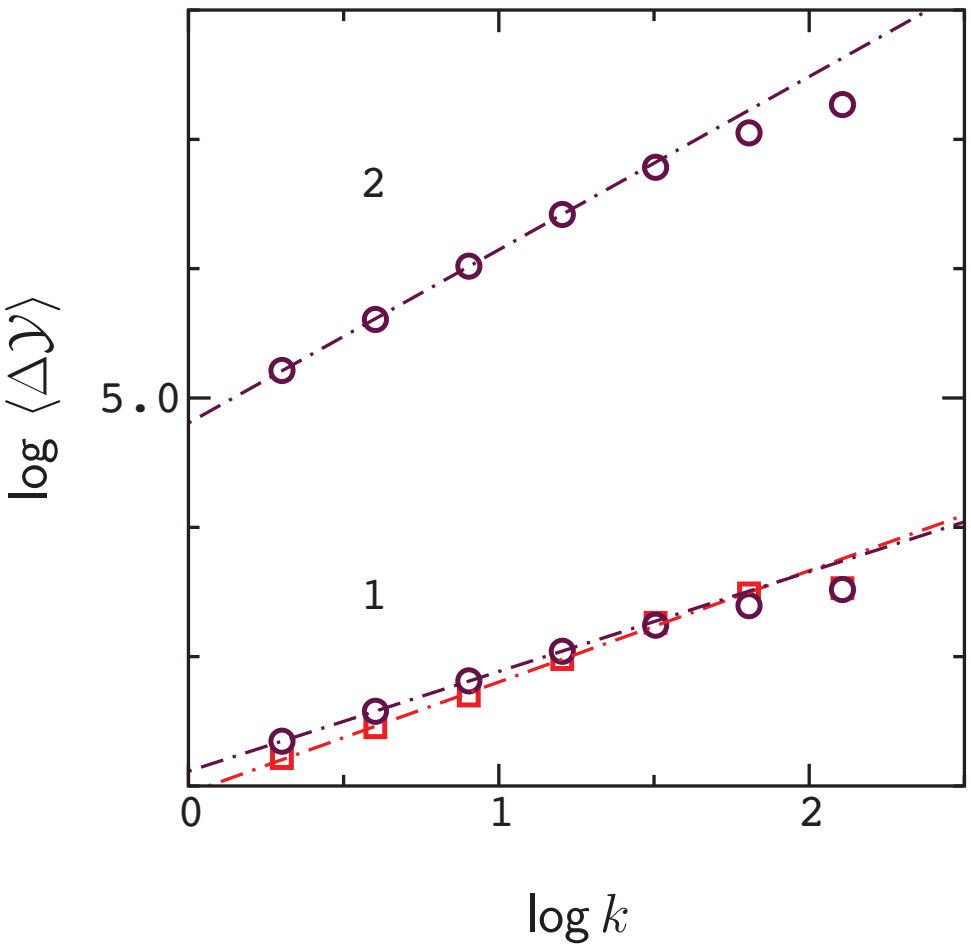

Figure 10 25Jul2016

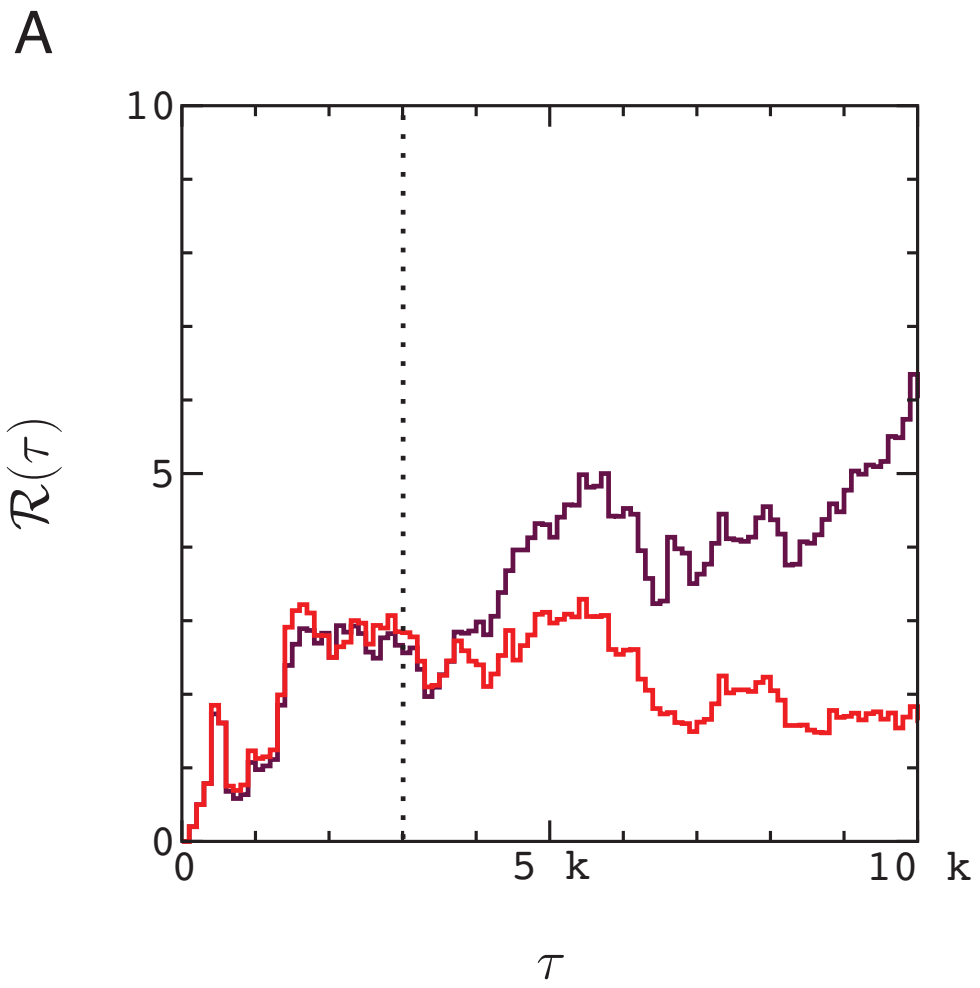

Figure 11A 25Jul2016

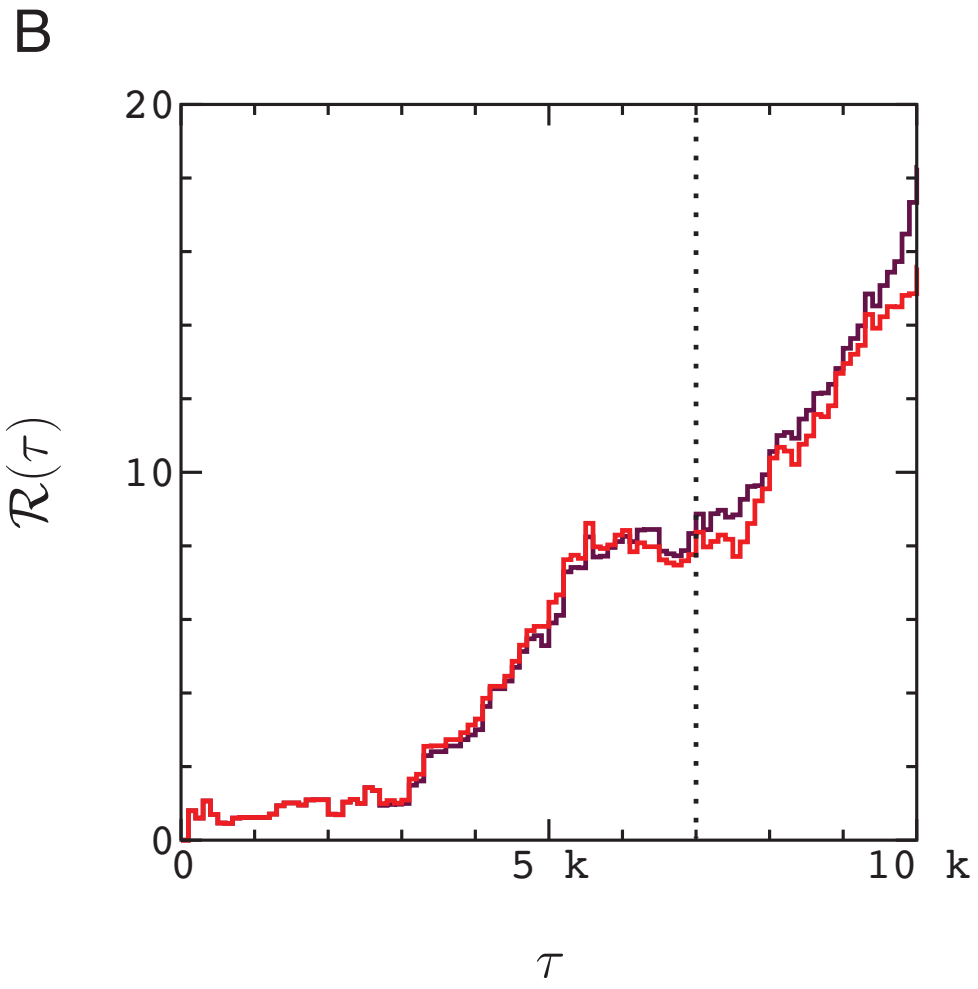

Figure 11B 25Jul2016

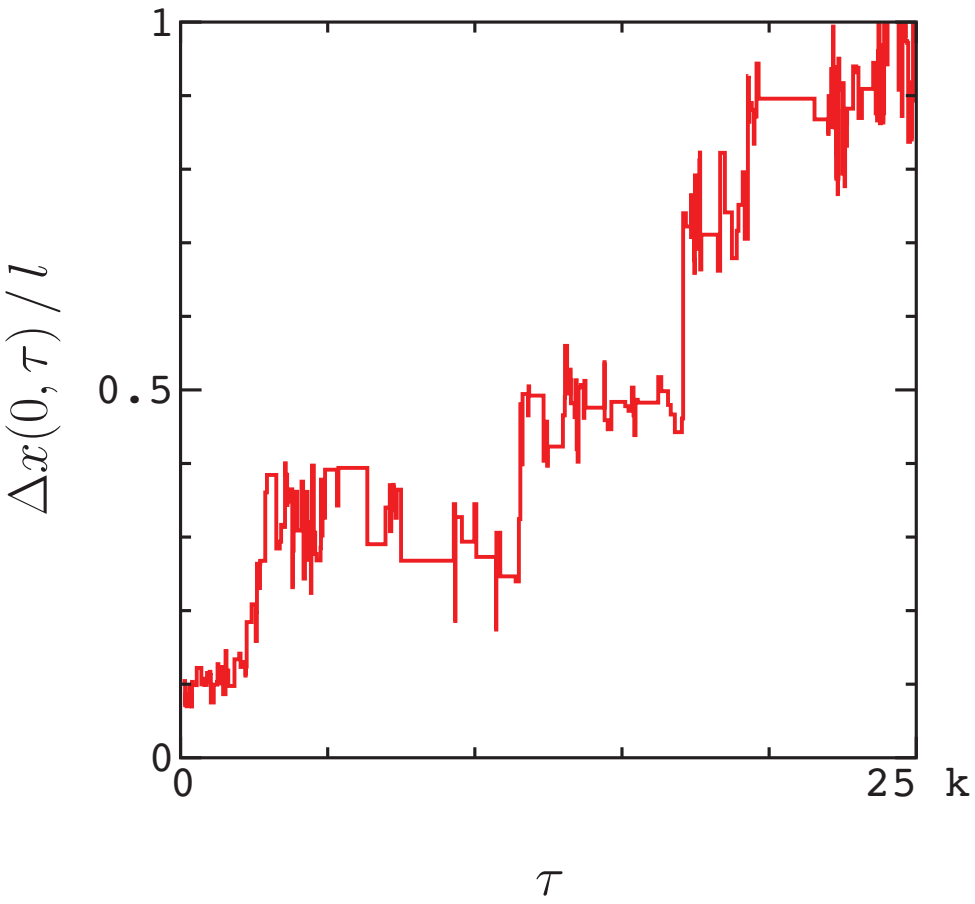

Figure 12 25Jul2016

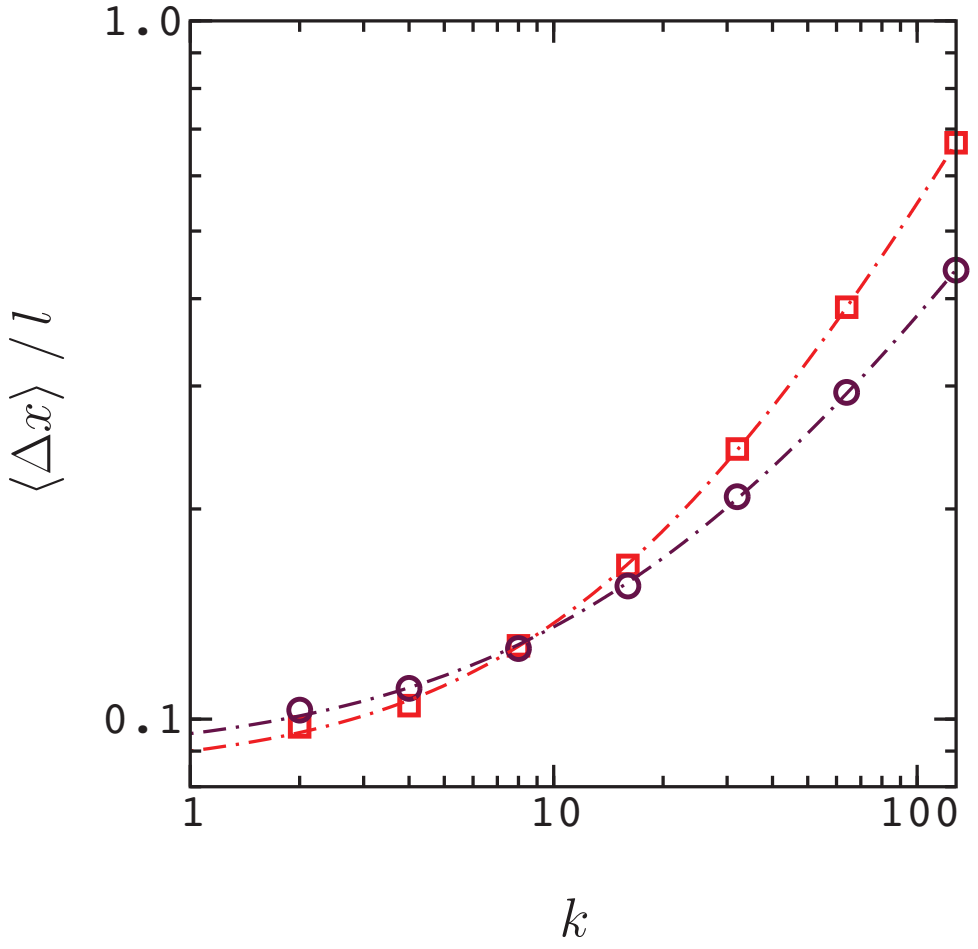

Figure 13 25Jul2016

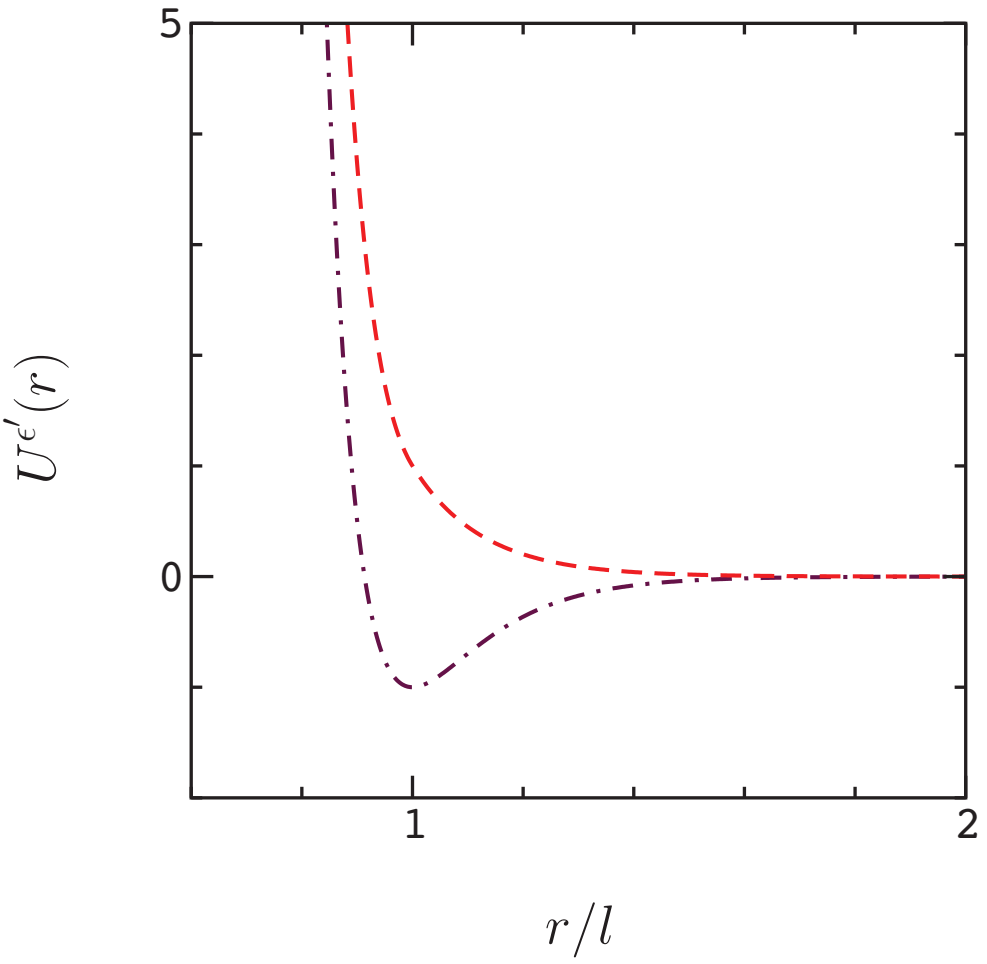

Figure 14 25Jul2016