www.molecularsystemsbiology.com

molecular
systems
biology

Quantifying epistatic interactions among the components constituting the protein translation system

Tomoaki Matsuura[1], Yasuaki Kazuta[1], Takuyo Aita[2], Jiro Adachi[1] and Tetsuya Yomo[1,3,4,*]

[1] Department of Bioinformatics Engineering, Graduate School of Information Science and Technology, Osaka University, Suita, Osaka, Japan, [2] Graduate School of Science and Engineering, Saitama University, Saitama, Japan, [3] Exploratory Research for Advanced Technology (ERATO), Japan Science and Technology Agency (JST), Suita, Osaka, Japan and [4] Graduate School of Frontier Biosciences, Osaka University, Suita, Osaka, Japan
* Corresponding author. Department of Bioinformatics Engineering, Graduate School of Information Science and Technology, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan. Tel.: +81 6 6879 4171; Fax: +81 6 6879 7428; E-mail: yomo@ist.osaka-u.ac.jp

In principle, the accumulation of knowledge regarding the molecular basis of biological systems should allow the development of large-scale kinetic models of their functions. However, the development of such models requires vast numbers of parameters, which are difficult to obtain in practice. Here, we used an *in vitro* translation system, consisting of 69 defined components, to quantify the epistatic interactions among changes in component concentrations through Bahadur expansion, thereby obtaining a coarse-grained model of protein synthesis activity. Analyses of the data measured using various combinations of component concentrations indicated that the contributions of larger than 2-body inter-component epistatic interactions are negligible, despite the presence of larger than 2-body physical interactions. These findings allowed the prediction of protein synthesis activity at various combinations of component concentrations from a small number of samples, the principle of which is applicable to analysis and optimization of other biological systems. Moreover, the average ratio of 2- to 1-body terms was estimated to be as small as 0.1, implying high adaptability and evolvability of the protein translation system.
Molecular Systems Biology **5**: 297; published online 18 August 2009; doi:10.1038/msb.2009.50
Subject Categories: synthetic biology
Keywords: Bahadur expansion; epistatic interaction; fitness landscape; *in vitro* translation system; protein synthesis

Introduction

The protein translation reaction, one of the most important regulators of cell behavior, involves the interactions of a large number of components, and has been studied extensively because of its importance in the cell (Nierhaus and Wilson, 2004). A reconstruction of an *Escherichia coli*-based *in vitro* translation system using protein components, highly purified on an individual basis, showed that 36 enzymes and ribosomes are sufficient to carry out protein translation (Shimizu *et al*, 2001). These minimal protein components include the ribosomal proteins; initiation, elongation, and release factors; aminoacyl-tRNA synthetases; and enzymes involved in energy regeneration. In addition, many studies have characterized the properties of such individual proteins in detail, for example, by kinetic analysis and three dimensional structural determination (e.g., Maier *et al*, 2005; Qin *et al*, 2006).

In principle, the accumulation of knowledge regarding the molecular basis of protein translation systems should allow the development of large-scale kinetic models of the entire reactions (Jamshidi and Palsson, 2008), which would provide insight into the complete relationship between the concentrations of the components, and the yield or rate of protein synthesis. Once these are obtained, we will have a complete understanding of the kinetic mechanism of the reaction that, for example, will allow prediction of the rates and/or yields under a given set of conditions. However, the development of a large-scale kinetic model requires a vast number of rate constants under a given set of conditions, which are difficult to obtain in practice. Thus, a coarse-grained model of the reaction is important (Covert *et al*, 2003; Price *et al*, 2004; Smallbone *et al*, 2007; Jamshidi and Palsson, 2008), which still provides insight into the kinetic mechanism as well as allows prediction.

One way of obtaining a coarse-grained model is to quantify the epistatic interactions (Boone *et al*, 2007; Poelwijk *et al*, 2007) among the components comprising the protein translation system. We use the term 'epistasis,' which is often used in the field of genetics (Boone *et al*, 2007; Poelwijk *et al*, 2007). Epistasis refers to the deviation from the expected phenotype when perturbations are combined. For example, negative epistasis means that although individual gene knockouts are dispensable, they become lethal when combined. The term epistasis is also used to refer to the interaction between the effects of mutations on the properties of proteins, which is also referred to as mutational nonadditivity. Here, we extend the usage of this term to express the interactions among the concentration changes of the components constituting biological systems.

Let us assume a system showing an activity *f* is composed of two components with concentrations $(c_i^0, c_j^0$; see Figure 1A). Furthermore, assume that the system alters the activity to $f + \Delta f$ by modulating the concentrations of the two components to (c_i^1, c_j^1). The difference in activity because of these concentration changes (Δf) is written as:

$$\Delta f = w_i + w_j + w_{ij} \qquad (1)$$

where w_i is the effect of altering the concentration of component *i* on the activity of the system, and w_{ij} is the interaction term (Figure 1A). When $w_{ij}=0$, the effects of altering the concentrations are additive and thus there is no epistatic interaction, whereas $w_{ij} \neq 0$ indicates that the two components show an epistatic interaction. The above example is a case with a system composed of two components, in which up to 2-body interactions may occur. However, a system composed of *n* components may show 2- to *n*-body interactions.

For interactions to be determined experimentally and quantitatively, the protein translation system should be composed of components the concentrations of which can be altered as required. Here, we used an *E. coli*-based *in vitro* translation system reconstituted from highly purified individual components, named the PURE system (Shimizu *et al*, 2001). As this system is prepared by mixing 69 defined components, the concentrations of which can be varied as desired, the protein synthesis activity of this system can be defined as a function of the concentrations of these 69 components. Using this system, we addressed the question: 'While it is possible to consider from 2- to 69-body interactions among the components, up to what body interaction terms make a significant contribution to protein synthesis activity of the system, and how large are the interaction terms?' Here, we report an analysis of the experimental results using Bahadur expansion (Solomon, 1961; Losee, 1994; Humphreys and Titterington, 1999), which gave quantitative values of the epistatic interactions among the components. This information provided insight into the kinetic mechanism of the reaction and also allowed us to predict the yield of the synthesized protein with various sets of component concentrations from small amounts of data. Our results are discussed with respect to adaptability and evolvability of the protein translation system.

Results

Defining three concentration vectors

The protein synthesis activity of the *in vitro* translation system used in this study (Shimizu *et al*, 2001) can be defined as a function of the concentrations of 69 components ($c_1, c_2, c_3, \ldots, c_{69}$). Note that molecules consisting of multiple elements, such as the ribosome, were counted as single components. We used the fluorescence intensity of GFP (green fluorescent protein) obtained after 3-h protein synthesis reaction at 37°C, with 300 nM mRNA of the *gfp* gene (Ito *et al*, 1999), as an indicator of the activity of this system, and defined activity (*f*) as the
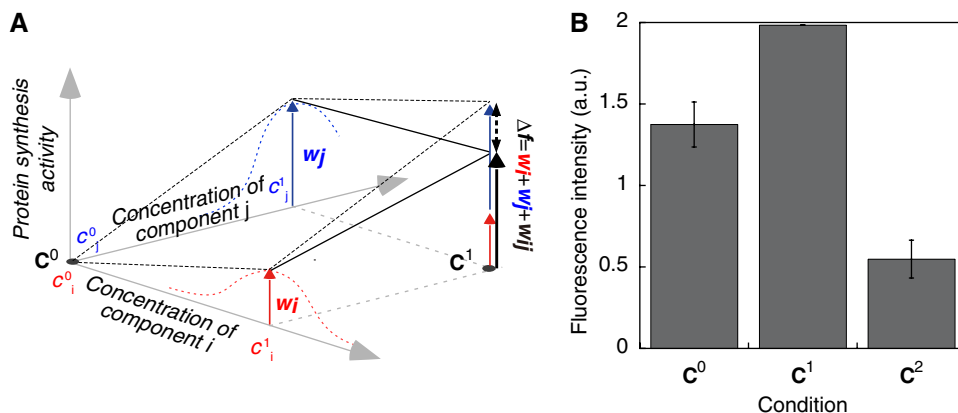

Figure 1 Schematic representation of the strategy for defining three concentration vectors. (**A**) Schematic with a system composed of two components *i* and *j*. Although the system is composed of 69 components, processes with two components are shown for simplicity. From the initial conditions, that is, $\mathbf{C}^0$ ($=c_i^0, c_j^0$), the concentration of component *i* (c_i^0) was varied to search for the concentration that maximizes the activity (c_i^1), whereas the concentrations of the other components remained fixed (red). The same was done with component *j* (blue). The activity of the system was then evaluated using the concentration vector $\mathbf{C}^1$ ($=c_i^1, c_j^1$). Identical optimization steps were carried out for another cycle to obtain $\mathbf{C}^2$. The height of the red arrow (w_i) plus the blue arrow (w_j) indicates the results expected when assuming additivity (no epistatic interaction, $w_{ij}=0$), the black bold arrow indicates the measured data, and dashed lines with arrows on both sides indicates the interaction term (w_{ij}). (**B**) Fluorescence intensity obtained with the GFP synthesis reaction using the concentration vectors $\mathbf{C}^0$, $\mathbf{C}^1$, and $\mathbf{C}^2$. The results of two independent trials are shown.

natural logarithm of fluorescence intensity (FI); $f=\ln(FI)$. Note that 3 h is the time duration in which the translation reaction is complete (Shimizu *et al*, 2001; Kazuta *et al*, 2008). Nevertheless, as the intensity value at 3 h is correlated with the initial reaction velocity (Supplementary Figure S1), f is considered to evaluate protein synthesis activity at the free energy level.

We first varied the concentrations of the components as described below and defined three different concentration vectors $\mathbf{C}^i=(c_1^i, c_2^i, c_3^i,\ldots, c_{69}^i)$ ($i=0,1,2$). Although the system is composed of 69 components, processes using two components are shown for simplicity in Figure 1A. The initial concentrations of 69 components $\mathbf{C}^0=(c_1^0, c_2^0, c_3^0,\ldots, c_{69}^0)$ were determined primarily based on the previous report by Shimizu *et al* (2001). The concentration of component i ($=1,2,\ldots,69$) was varied to search for the concentration that maximizes the GFP synthesis activity, whereas the concentrations of the other components remained fixed, and the concentration of component i for the largest activity c_i^1 was obtained (Supplementary Figure S2). The concentrations of components, the activity of those could not be improved by altering their concentration, were not altered from the initial value. In this way, we determined the concentration vector $\mathbf{C}^1=(c_1^1, c_2^1, c_3^1,\ldots, c_{69}^1)$. The identical optimization cycle was carried out from $\mathbf{C}^1$ to obtain $\mathbf{C}^2$ (values given in Supplementary Table S1). The entire dataset obtained when the concentrations of individual components were altered is shown in Supplementary Figure S2, and the text data are given in Supplementary Table S3.

The results of GFP synthesis reaction using $\mathbf{C}^0$, $\mathbf{C}^1$, and $\mathbf{C}^2$ are shown in Figure 1B. In case, there were no interactions among the concentration changes, the fluorescence intensity should increase monotonously, as the effects of optimizing the concentration of individual components would be accumulated. The observed intensity increased from $FI(\mathbf{C}^0)$ to $FI(\mathbf{C}^1)$, whereas it decreased from $FI(\mathbf{C}^1)$ to $FI(\mathbf{C}^2)$. These results indicated the presence of epistatic interactions among the components.

Grouping of 69 components into modules

This study was carried out to quantify the epistatic interactions among 69 components. Using our strategy (see below), if each component takes one of the two different states, exhaustive quantification of the interaction requires more than 10^{20} ($\approx 2^{69}$) measurements, which is obviously not feasible. To overcome this practical problem, we classified 69 'components' into three or four 'modules' and examined the extents of interactions among the modules (Box 1). As described below, we obtained similar results regardless of the modularization scheme used, and thus investigating the inter-module interactions led to elucidation of the inter-component interactions (see Box 2, and Supplementary information, Appendix I). The rationale behind the modularization experiments is illustrated in Box 2.

Box 1 shows a schematic representation of the modularization experiments. We prepared four modules from each of the concentration vectors $\mathbf{C}^0$ and $\mathbf{C}^1$, according to modularization scheme 1 (Figure 2A), yielding concentration vectors $(\mathbf{m}_1^t, \mathbf{m}_2^t, \mathbf{m}_3^t, \mathbf{m}_4^t)=\mathbf{C}^t$ ($t=0,1$), where $\mathbf{m}_k^t$ is the vector of the component's concentrations given by the modularization scheme. Then, the activity of the system was measured by recombining these modules (Box 1). Notations, such as '0000' and '1111' in

Box 1 Schematic representation of the modularization experiments
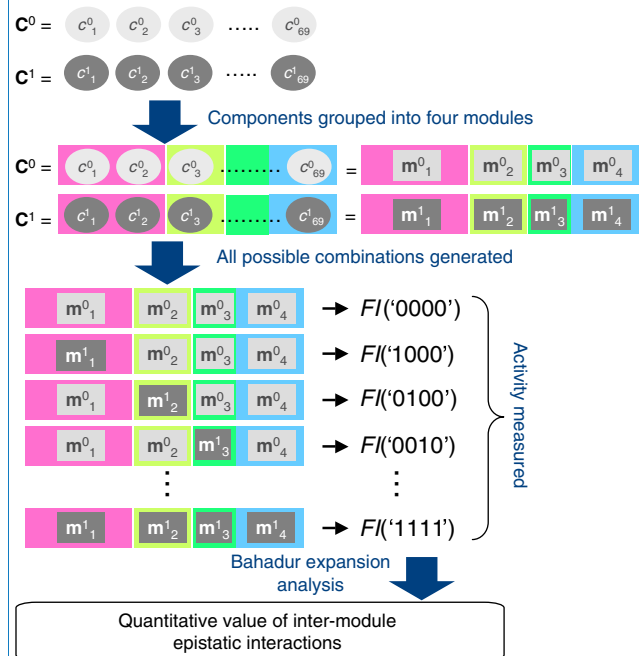
Box 1 The 69 components were grouped into four modules, yielding concentration vectors $(\mathbf{m}_1^t, \mathbf{m}_2^t, \mathbf{m}_3^t, \mathbf{m}_4^t)=\mathbf{C}^t$ ($t=0,1$), where $\mathbf{m}_k^t$ is the vector of the component's concentrations given by the modularization scheme. Then, the activity of the system was measured by recombining these modules. Notations, such as '0000' and '1111' indicate $(\mathbf{m}_1^0, \mathbf{m}_2^0, \mathbf{m}_3^0, \mathbf{m}_4^0)$ and $(\mathbf{m}_1^1, \mathbf{m}_2^1, \mathbf{m}_3^1, \mathbf{m}_4^1)$, respectively. As this 'sequence' (e.g., '0101'=$(\mathbf{m}_1^0, \mathbf{m}_2^1, \mathbf{m}_3^0, \mathbf{m}_4^1)$) gives a set of concentrations of all 69 components, fluorescence intensity (e.g., FI('0000')) is assigned for this sequence. Activity values of all possible sequences generated by recombining the modules '0000' and '1111' (denoted as '0000 × 1111') were measured. These data were subjected to Bahadur expansion analysis to obtain quantitative values of inter-module interactions. Note that investigation of the 'inter-module' interactions led to the elucidation of the 'inter-component' interactions (see Box 2 and Supplementary information, Appendix I).

Box 1 indicate $(\mathbf{m}_1^0, \mathbf{m}_2^0, \mathbf{m}_3^0, \mathbf{m}_4^0)$ and $(\mathbf{m}_1^1, \mathbf{m}_2^1, \mathbf{m}_3^1, \mathbf{m}_4^1)$, respectively. As this 'sequence' (e.g., '0101'=$(\mathbf{m}_1^0, \mathbf{m}_2^1, \mathbf{m}_3^0, \mathbf{m}_4^1)$) gives a set of concentrations of all 69 components, fluorescence intensity is assigned for this sequence. Figure 2B shows the fluorescence intensities of all possible sequences generated by recombining the modules '0000' and '1111' (denoted as '0000 × 1111') (left), where 16 experimental data sets were obtained. Identical experiments were carried out by grouping $\mathbf{C}^1$ and $\mathbf{C}^2$ into four modules according to modularization scheme 1 (denoted as '1111 × 2222') (Figure 2B, right), or by grouping $\mathbf{C}^0$ and $\mathbf{C}^1$ into three modules according to modularization scheme 2 or 3 (Figure 2A and C) (denoted as '000 × 111'). Data shown in Figures 2B and C were subjected to Bahadur expansion analysis to quantify the inter-module interactions.

Inter-module interaction showed by Bahadur expansion

We defined the activity $f(\mathbf{x})$ of a sequence $\mathbf{x}$, where $\mathbf{x}=x_1x_2x_3x_4$ (e.g., $\mathbf{x}$='0110'), as the natural logarithm of the fluorescence

Box 2 Investigating the 'inter-module' interaction leads to elucidation of the 'inter-component' interactions

Grouped 'arbitrarily' into modules and quantified the 'inter-module' interactions

2-body inter-component interactions present

2-body inter-component interactions absent

2-body inter-module interactions detected

2-body inter-module interactions not detected

2-body inter-component interaction
2-body inter-module interaction
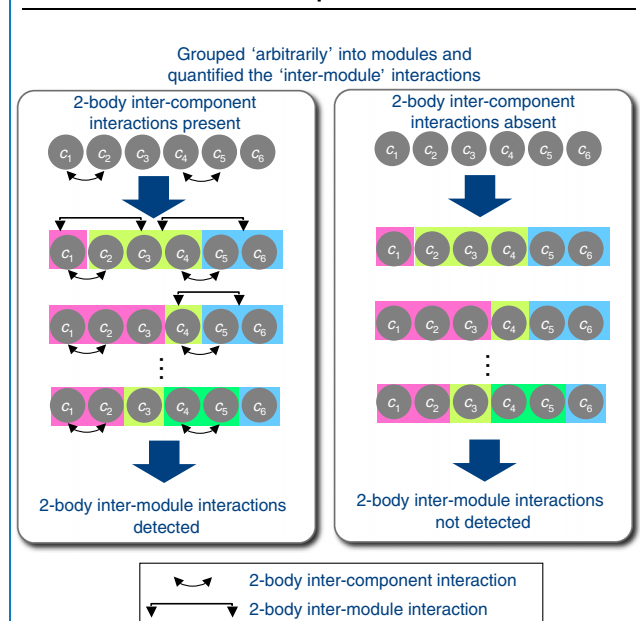
Box 2   Let us assume a system composed of six components. The six components were grouped arbitrarily into modules and the inter-module interactions were quantified using Bahadur expansion analysis. When 2-body interactions are present between the components, 2-body inter-module interactions are detected depending on the modularization scheme (left). However, when 2-body interactions are absent between the components, 2-body inter-module interactions are absent irrespective of the modularization scheme. Hence, when 'inter-module' interactions larger than 1-body interactions can be approximated to zero irrespective of how to define the modules, that is, irrespective of the modularization scheme (grouping of components) and concentrations of individual components in each module, the 'inter-component' interactions larger than 1-body interactions can be approximated to zero. Similarly, when larger than 2-body inter-module interactions are absent, larger than 2-body inter-component interactions are absent. In this way, investigating the 'inter-module' interaction leads to elucidation of the 'inter-component' interactions. For the mathematical description, see Supplementary information, Appendix I.

intensity $FI(\mathbf{x})$; $f(\mathbf{x}) = \ln(FI(\mathbf{x}))$. We carried out Bahadur expansion analysis (Solomon, 1961; Losee, 1994; Humphreys and Titterington, 1999), which is similar to Fourier expansion, to map a set of experimental activity values into an orthonormal system in which bases represent 1-body, 2-body, 3-body, *etc.*, interaction terms (for further details, see Materials and Methods). In the case of four-letter sequences, Bahadur expansion converts 2^4 activity values into 2^4 different interaction terms (f_0, w_i, w_{ij}, w_{ijk}, and w_{ijkl}, see below), which can be compared with each other. For example, using '0000 × 1111' and '1111 × 2222' in Figure 2B, a set of experimental activities for all 16 ($= 2^4$) sequences are mapped into the following orthonormal system consisting of 16 bases ($1, z_1, z_2, z_3, z_4, z_1z_2, z_1z_3, \ldots, z_1z_2z_3z_4$):

$$
\begin{aligned}
f(\mathbf{x}) = f_0 &+ \sum_{i=1}^{4} w_i z_i + \sum_{i=1}^{3} \sum_{j=i+1}^{4} w_{ij} z_i z_j \\
&+ \sum_{i=1}^{2} \sum_{j=i+1}^{3} \sum_{k=j+1}^{4} w_{ijk} z_i z_j z_k + w_{1234} z_1 z_2 z_3 z_4
\end{aligned} \tag{2}
$$

where z_i is determined by converting a letter x_i as follows:

$$
z_i = z_i(x_i) = \begin{cases} -1, & \text{if } x_i = 0 \\ +1, & \text{if } x_i = 1 \end{cases} \tag{3}
$$

and f_0, w_i, w_{ij}, w_{ijk}, and w_{ijkl} are the 0th, 1st, 2nd, 3rd, and 4th order Bahadur coefficients, respectively. The 0th order coefficient (f_0) is an average activity over all sequences, and the 1st order coefficient (w_i) is the 1-body contribution of a module i. The terms w_{ij}, w_{ijk}, and w_{ijkl} are 2-, 3-, and 4-body contributions, respectively, which represent the epistasis caused by inter-module interactions.

The calculated Bahadur coefficients are shown in Figure 3A. The absolute values of the coefficients became smaller as the order increased for both '0000 × 1111' and '1111 × 2222.' Note that if the activities are assigned as random numbers for all sequences, then all coefficients obtained using Bahadur expansion take an identical weight on average as with white noise. These results indicate that higher order terms make less of a contribution to the activity. Next, the coefficient of determination (R^2) was calculated for each Bahadur coefficient (Figure 3B). The R^2 value for each Bahadur coefficient is equivalent to the R^2 (square of the correlation coefficient R) of regression analysis between the calculated and experimental activities, in which the calculated value was obtained from equation (2) by setting all other coefficients to 0. We confirmed that higher order terms make smaller contributions to the activity. Furthermore, the activity for each sequence was calculated using the obtained coefficients but by truncating equation (2) at the 1st, 2nd, 3rd, and 4th order, respectively. The inset of Figure 3B shows R^2 values for the correlations between the calculated and experimental data. These R^2 values are equivalent to those obtained by cumulating the elemental R^2 values up to the 1st, 2nd, 3rd, and 4th order, respectively. The R^2 value reached more than 0.96 even with truncation at the 3rd and 4th order, indicating that truncation at the 2nd order is sufficient to explain the experimental results. That is, larger than 2-body interactions among the modules can be approximated to zero.

To verify the statistical significance of these findings, we carried out a shuffling test. By shuffling the assignment of the observed activity values to sequences randomly, we generated 1000 sets of shuffled tables. Then, we carried out the same analysis as described above. In the case of shuffled data sets, the R^2 value for each Bahadur coefficient took an identical weight on average ($0.067 \approx 1/15$) as with white noise. Furthermore, the R^2 values calculated by truncating equation (2) at the 1st, 2nd, 3rd, and 4th order, respectively, were significantly smaller than the original data for the 1st and 2nd order truncation (inset of Figure 3B, black bar), indicating that the observation that larger than 2-body inter-module interaction can be approximated to zero is a physicochemical property of the *in vitro* translation system.

We then carried out the same analysis as described above with the data obtained by grouping the components into three modules (Figure 2C) and obtained the R^2 value for each Bahadur coefficient (Figure 3C). Consistent with the four module experiments, R^2 values decreased for higher order interaction terms. The inset of Figure 3C shows R^2 values for the correlations between the calculated and experimental data, in which the calculated values were obtained by 1st, 2nd, and 3rd order truncation, respectively.
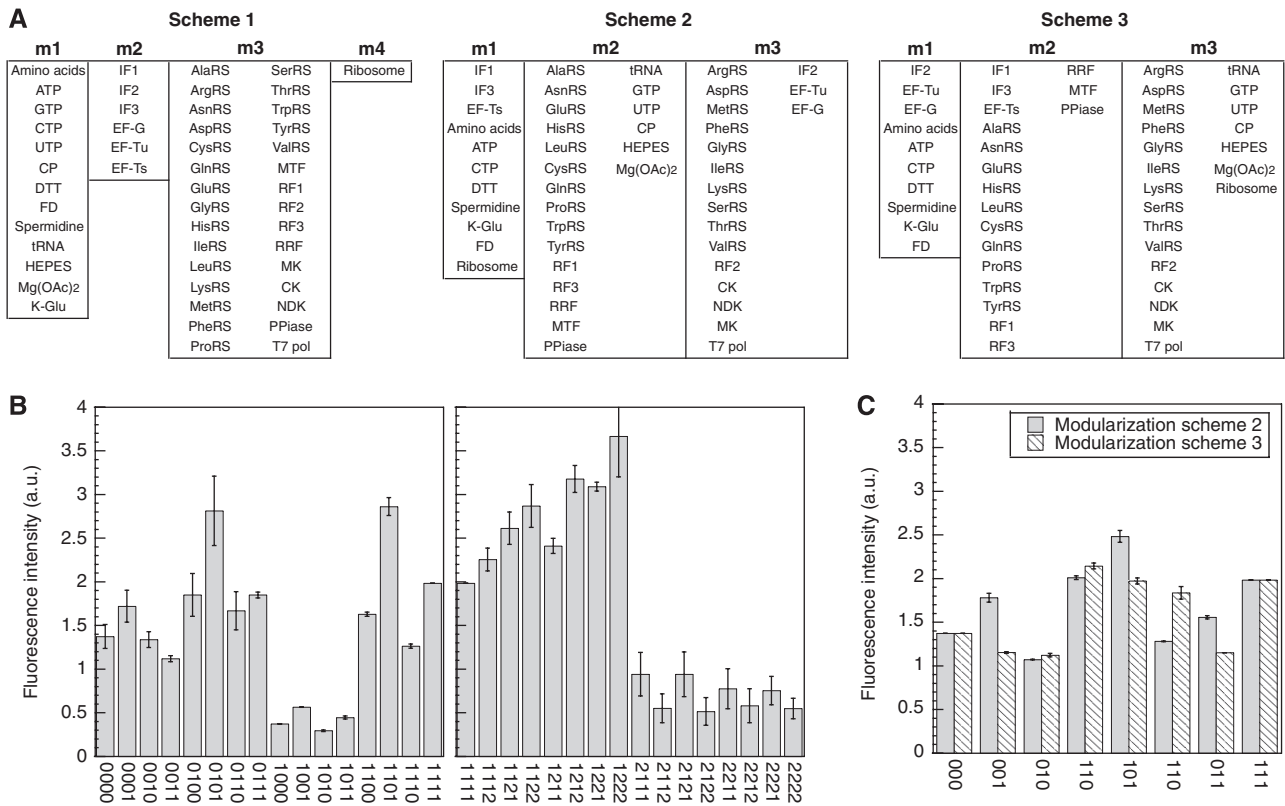
Figure 2 Grouping of the 69 components into modules to investigate the inter-module interactions. (**A**) Three modularization schemes used in this study. The 69 components were grouped into 4 (scheme 1) or 3 modules (schemes 2 and 3). See Supplementary Table S1 for abbreviations of the names of the components and their concentrations. (**B**) Combinatorial experiments of modules '0000 × 1111' (left) and '1111 × 2222' (right). Modularization was carried out according to scheme 1. Notations, such as '0101', indicate the concentration vector generated by combining the modules ($\mathbf{m}_1^0, \mathbf{m}_2^1, \mathbf{m}_3^0, \mathbf{m}_4^1$). Fluorescence intensities of synthesized GFP for each binary sequence are shown on the vertical axis. Results of two independent trials are shown. (**C**) Combinatorial experiments using modules '000 × 111.' Modularization was carried out according to scheme 2 or 3. Results of two independent trials are shown. Text data of (B) and (C) are given in Supplementary Table S4.
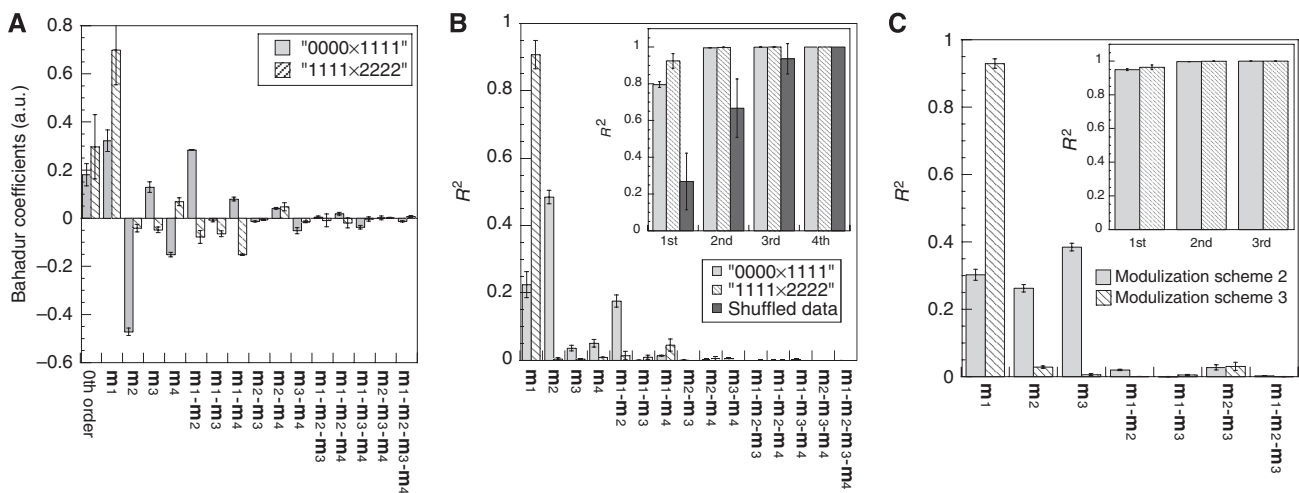

Figure 3 Results of Bahadur expansion analysis of the experimental data. (**A**) Bahadur coefficients determined from the data shown in Figure 2B. The heights of the bars labeled $\mathbf{m}_i$ and $\mathbf{m}_i$–$\mathbf{m}_j$ on the horizontal axis indicate the 1st order coefficient of module $\mathbf{m}_i$ and the 2nd order coefficient of the interaction between module $\mathbf{m}_i$ and $\mathbf{m}_j$, respectively. (**B, C**) R^2 values for each Bahadur coefficient from the results shown in Figure 2B (B) and Figure 2C (C). Insets show the R^2 values calculated by each order truncation of equation (m4). The results of two independent trials are shown.

The R^2 value reached more than 0.99 even without the 3rd order coefficients regardless of the modularization scheme, indicating that truncation at the 2nd order is sufficient to explain the experimental results. Thus, we concluded that larger than 2-body interactions among the modules could be approximated to zero, regardless of the modularization scheme used.

Inter-component interaction of six components showed by Bahadur expansion

We aimed to quantify the epistatic interactions among 69 components. For this purpose, we grouped the components into modules to investigate the inter-module interactions, which still provided information on the inter-component interactions. This was based on the following theorem (see Box 2 for schematic explanations, and Supplementary information, Appendix I for mathematical descriptions):

If 'inter-module' interactions larger than 2-body can be approximated to zero irrespective of how to define the modules, that is, irrespective of the modularization scheme (grouping of components) and concentrations of individual components in each module, the 'inter-component' interactions larger than 2-body interactions can be approximated to zero.

In the previous section, we showed that 1- and 2-body inter-module interactions are sufficient to explain the experimental results with three different modularization schemes (Figure 3B and C), and with two different pairs of concentration vectors (Figure 3B). By applying the above theorem to the four observations, we developed the following conjecture: inter-component interactions larger than 2-body can be approximated to zero for the components comprising the protein translation system used. The question is whether four different experiments (Figure 2B and C) are sufficient to fulfill the arbitrariness. Rather than testing more different modularization schemes, we decided to conduct the experiment to quantify the inter-component interaction directly, which further suggested that the above conjecture is true.

We thus further investigated whether the above conjecture is true by directly measuring the inter-component interactions. We chose six components (magnesium acetate (Mg(OAc)$_2$), transfer RNA (tRNA), spermidine, potassium glutamate (K-Glu), NTPs, and creatine phosphate (CP)), which affected protein synthesis activity when their concentrations were altered. The experiment was designed such that each of the six components took the concentration in either C^1 or C^2, whereas

the concentrations of the remaining 63 components were fixed to C^1 (values are given in Supplementary Table S2). Therefore, the experimental conditions here can be written as a binary sequence of length six: for example, '111111'=($c^1_{Mg(OAc)2}$, c^1_{tRNA}, $c^1_{spermidine}$, c^1_{K-Glu}, c^1_{NTP}, c^1_{CP}) and '222222'=($c^2_{Mg(OAc)2}$, c^2_{tRNA}, $c^2_{spermidine}$, c^2_{K-Glu}, c^2_{NTP}, c^2_{CP}). The results of '111111 × 222222' are shown in Figure 4A. R^2 values calculated using the 1st–6th order truncation are shown in Figure 4B. The R^2 value reached more than 0.99 even without coefficients higher than 2nd order, indicating that 2nd order truncation is sufficient to explain the experimental results. These results were consistent with the conjecture, further suggesting that the above conjecture is true.

Relative contribution of 2-body to 1-body interaction terms on protein synthesis activity

We found that the activity of the system can be expressed by using up to the 2-body interaction terms (e.g. $f = f_0 + z_i w_i + z_j w_j + z_i z_j w_{ij}$). Therefore, we investigated the relative contribution of 2-body ($z_i z_j w_{ij}$) to 1-body ($z_i w_i + z_j w_j$) interaction terms on protein synthesis activity. We investigated these by plotting the relationship between ($z_i w_i + z_j w_j$) and ($z_i z_j w_{ij}$), which represents the sum of the effects of two perturbations (alteration of the concentrations of two components or modules individually), and the effects of interaction between the two, respectively (Figure 5). Larger '$z_i w_i + z_j w_j$' values tended to show larger '$z_i z_j w_{ij}$' values, indicating that larger interaction occurs when combining larger perturbations. We also calculated γ_{NA} ($= |z_{ij} w_{ij}| / |z_i w_i + z_j w_j|$) from the data shown in Figure 5 and obtained a median value of 0.16. This observation indicated that when simultaneously altering the component concentrations, the activity of the system can be reduced or increased on average by a factor of 0.16 from the sum of the effects of individual changes. Thus, the inter-component interaction in the protein translation system showed a small degree of interaction on average.
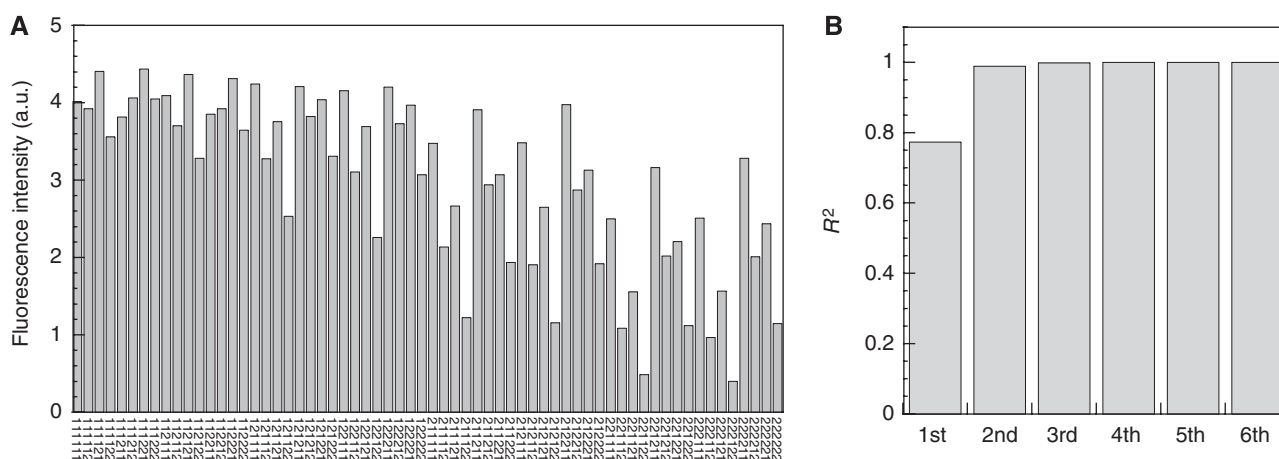

Figure 4 Quantification of inter-component interactions. (**A**) Six components (Mg(OAc)$_2$, tRNA, spermidine, K-Glu, NTPs, and CP) were designed to take the concentration either in C^1 or C^2, whereas the concentrations of the other 63 components were fixed to C^1 (values are given in Supplementary Table S2). Therefore, the experimental conditions (concentration vector) here can be written as a binary sequence of length 6, for example, '111111'=($c^1_{Mg(OAc)2}$, c^1_{tRNA}, $c^1_{spermidine}$, c^1_{K-Glu}, c^1_{NTP}, c^1_{CP}) and '222222'=($c^2_{Mg(OAc)2}$, c^2_{tRNA}, $c^2_{spermidine}$, c^2_{K-Glu}, c^2_{NTP}, c^2_{CP}). The experimental results of '111111 × 222222' are shown. Text data are given in Supplementary Table S4. (**B**) R^2 values calculated by the 1st–6th order truncation of equation (m4) for the combinatorial experiments of six components.
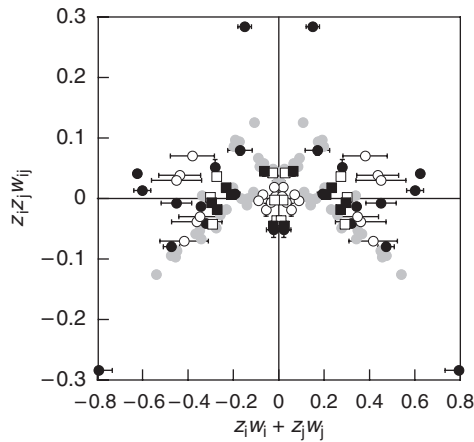
Figure 5 Relative contribution of 2-body ($z_iz_jw_{ij}$) to 1-body ($z_iw_i + z_jw_j$) interaction terms on protein synthesis activity. Combinatorial experiments with modules '0000 × 1111' (filled circles) and '1111 × 2222' (open circles), modularization of which was carried out according to scheme 1 (Figure 2B). Combinatorial experiments with modules '000 × 111' (Figure 2C) modularization of which was carried out according to modularization scheme 2 (filled boxes) or 3 (open boxes). Combinatorial experiments of six components '111111 × 222222' (gray circles; also see Figure 4A). The median of γ_{NA} was 0.16. As z_i takes $+1$ or -1 depending on the sequence, ($z_iw_i + z_jw_j$, $z_iz_jw_{ij}$) can take ($w_i + w_j$, w_{ij}), ($-w_i - w_j$, w_{ij}), ($w_i - w_j$, $-w_{ij}$), or ($-w_i + w_j$, $-w_{ij}$), and therefore the plots become symmetric.

Discussion

In the protein translation system used in this study, although 2- to 69-body inter-component interactions are conceivable, we have shown that larger than 2-body interactions can be approximated to zero. Note that this conclusion is valid with alteration of the concentrations of the components over the range tested in this study. The absence of larger than 2-body interactions (epistatic interactions) reported here does not indicate the absence of molecular complexes of more than two components. Obviously, the protein translation reaction proceeds by generating large complexes (Nierhaus and Wilson, 2004). Below, we discuss the interpretation of our results from the kinetic viewpoint, and also give an example of 2-body interaction from the molecular viewpoint.

Fluorescence intensity obtained experimentally (*FI*), which correlates with the initial reaction velocity (*v*) (Supplementary Figure S1A) can be factorized as follows:

$$\ln(v) \propto \ln(FI(c_0, c_1, \ldots, c_{69})) = \text{const} + \sum_{i=1}^{69} \ln(fnc_i(c_i))$$
$$+ \sum_{i<j} \ln(fnc_{ij}(c_i, c_j)) + \sum_{i<j<k} \ln(fnc_{ijk}(c_i, c_j, c_k)) + ..$$
$$+ \ln(fnc_{12...69}(c_0, c_1 \ldots, c_{69})) \quad (4)$$

where *fnc* is an arbitrary function and c_i is the concentration of component *i*. The presence of *t*-th term ($t=1, 2,\ldots, 69$) in the above equation is identical to the presence of the *t*-body interaction term in the Bahadur expansion (see Supplementary information, Appendix II for details). Thus, our results indicated that when factorizing the polynomial form of the

large-scale kinetic models, larger than 2nd order terms in the above equation can be approximated to zero. Although the absence of larger than 2-body interactions alone cannot show the detailed molecular mechanism, it is important to link the epistatic interaction and the physical interactions among the molecules. Therefore, we provide one example of a 2-body interaction below.

We considered GTP being utilized at various stages of the protein translation reaction. If two different enzymes (or reaction intermediates) compete for free GTP and the rate of the reaction catalyzed by the enzymes is limited by the GTP concentration, there will be a 2-body epistatic interaction between the enzymes (see Supplementary information, Appendix II for details). Similarly, if *n* enzymes compete for GTP, there will be *n*-body interactions. Thus, even in the absence of direct physical interactions among the enzymes, epistatic interactions occur through an indirect physical interaction through the GTP molecule. However, epistatic interactions disappear if the GTP concentration is sufficiently high such that the rates of the reactions catalyzed by the enzymes are no longer limited by the GTP concentration.

As biological systems consist of vast numbers of components, it would be useful to be able to predict the activity values under vast numbers of conditions with different combinations of component concentrations (Yin and Carter, 1996; Young *et al*, 1997; Arita *et al*, 2002; Benos *et al*, 2002; Chester *et al*, 2004; Wiedemann *et al*, 2004). The absence of larger than 2-body inter-component interactions means that activity values of the *in vitro* translation system can be predicted by estimating up to the 2nd order Bahadur coefficients. To estimate those for a binary sequence with a length of *n*, a set of activity of at least $_nC_0 + {_nC_1} + {_nC_2} = 0.5 \times (2 + n + n^2)$ sequences is needed. Once these coefficients are obtained, it is possible to predict the results of all other possible sequences ($2^n - 0.5 \times (2 + n + n^2)$). As an example, we tested the predictability using the data in which fluorescence intensity is defined by a binary sequence of length six (Figure 4A). In this case, at least 22 experimental data are needed to estimate the 2nd order Bahadur coefficients for prediction of the other 42 ($= 2^6 - 22$) results. A typical scheme for choosing the 22 data (and sequences) is as follows. First, pick a reference sequence (e.g., '111111'), and then all possible single-point mutants ('211111,' '121111,'…, '111112'), and the double-point mutants ('221111,' '212111,'…, '111122'). Note that although the selection strategy often follows the theory of the design of experiments (Fisher, 1966), our simple scheme was sufficient for accurate prediction as described below. Using the 22 sequence–activity relationships, up to the 2nd order Bahadur coefficients can be estimated using equation (m4) (Materials and methods), which then allow prediction of the remaining 44 samples. Figure 6A shows the correlation between the experimental and predicted data using '111111' as a reference sequence; the prediction showed good agreement with the experimental data. Figure 6B shows R^2 values calculated similarly using each of the 64 as a reference sequence. This rank order plot shows that the R^2 value was > 0.8 in 57 of 64 cases and thus high R^2 values could be obtained with 90% probability. Such high R^2 values were not obtained using the same prediction by the 1st order truncation,
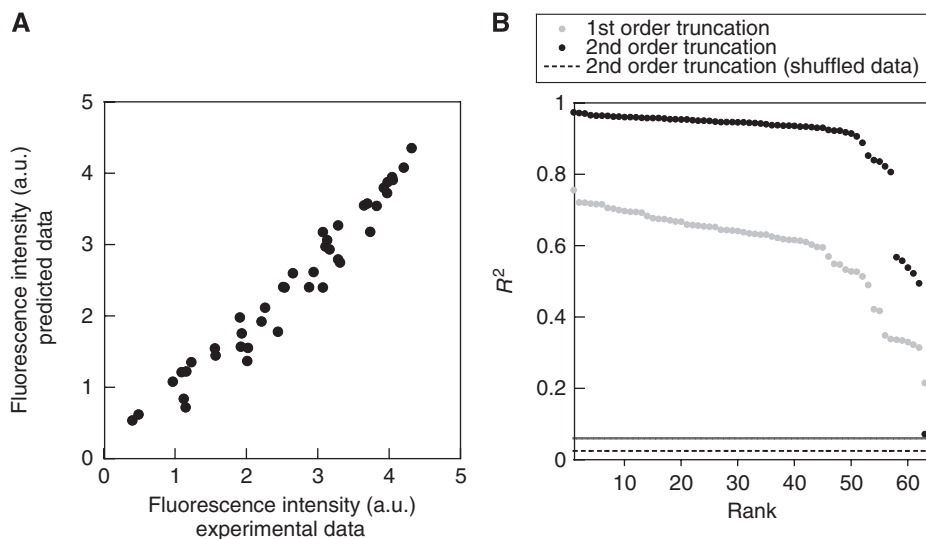
A

B

Figure 6 Predicting the activity from a small number of samples. (**A**) Correlation between the predicted and experimental data (Figure 4A) on using '111111' as a reference sequence to obtain up to the 2nd order Bahadur coefficients. $R^2 = 0.933$ was obtained. (**B**) The rank order plot (or cumulative frequency distribution) of the 64 R^2 values, obtained from the correlation between the experimental and predicted data calculated using each of the 64 reference sequences. Predicted data were calculated by 2nd (black circles) or 1st (gray circles) order truncation of equation (m4). The gray dashed line and the bold line show the average and s.d. of the R^2 value obtained, respectively, when the prediction strategy of using up to the 2nd order Bahadur coefficients was applied to predict the respective 100 data sets in which the sequence–activity relationship was shuffled randomly.

indicating the necessity of 2nd order coefficients for accurate prediction. Furthermore, when the strategy of 2nd order truncation was applied to the prediction of the data sets in which the sequence–activity relationship was shuffled randomly, we obtained an average R^2 value of 0.025, indicating the necessity of considering up to 2-body interactions for accurate prediction. The methodology presented here is effective for prediction and optimization of other biological systems, particularly if their higher order epistatic interactions are estimated to be negligible as in the protein translation system.

Our results may be important to understand the evolvability and the adaptability of the protein translation system. Typically, the presence of epistatic interactions in a genetic interaction network indicates that the effects of 2 particular perturbations are mutually interdependent. For example, although individual mutations A and B are deleterious to the cell (decrease fitness), they become beneficial (increase fitness) when both mutations are combined. In such cases, accumulation of beneficial mutations in a population requires a longer time than in the absence of such interactions. This is because two mutations A and B have to be introduced simultaneously in the presence of interactions, whereas each beneficial mutation can be accumulated sequentially in the absence of such interactions. Using the genetic interaction network, analysis of the interactions is more qualitative than quantitative. A quantitative analysis of epistatic interactions among the mutations of proteins (mutational nonadditivity) has been carried out, and the extent of such nonadditivity has been shown to be small: the effects of two simultaneous mutations differ by an average of 10% from the sum of the effects of individual mutations (Wells, 1990; Dill, 1997; Matsuura *et al*, 1998; Man and Stormo, 2001; Aita *et al*, 2002; Bulyk *et al*,

2002) (see Supplementary information, Appendix III). This property has allowed their past evolutionary processes, as each beneficial mutation can be accumulated sequentially. Small values of nonadditivity can also explain why a number of directed evolution experiments succeeded in evolving protein function artificially (Arnold *et al*, 2001; Matsuura and Yomo, 2006).

We quantified the epistatic interactions using an *in vitro* translation system reconstituted only from components essential for the reaction. Therefore, unlike living cells that can tolerate single gene knockout of substantial fractions of the genes because of buffering by the presence of duplicate genes or alternative biological pathways (Kitano, 2004; Deutscher *et al*, 2006; Boone *et al*, 2007), a single knockout of any of the components of the present system is lethal (Shimizu *et al*, 2001). Using such a system, we estimated that the extent of epistatic interaction between the components constituting the system is $\gamma_{NA} = 0.16$ on average, and is thus small as mutational nonadditivity described above. This small epistatic interaction or nonadditivity suggests that the protein translation system has the potential to adjust the concentration of each of the components in a given environment without becoming trapped in local maxima, thus avoiding an exhaustive search in the concentration space. Similar to the protein evolution mentioned above, the system can accumulate beneficial mutations, for example, in the promoter regions thereby altering the component concentrations and enabling adaptation and evolution in a given environment or even in new environments. Although the extent of epistatic interaction estimated here is derived from the protein translation system, as all biological systems are the product of natural evolution, the small extent of epistatic interactions may be a general property of all living systems.

Materials and methods

In vitro translation system

All plasmids encoding the proteins included in the *in vitro* translation system used (PURE system) were kindly provided by Professor Ueda and Dr Shimizu (University of Tokyo). All proteins were purified according to protocols of Kazuta *et al* (2008) and Shimizu *et al* (2001), and ribosomes were purified according to the protocol of Ohashi *et al* (2007). For GFP synthesis, aliquots of 20 μl of the *in vitro* translation system containing four units of RNasin (Promega), 50 nM Alexa-Fluor647 (Invitrogen), and 300 nM GFPuv5 RNA were prepared and incubated at 37°C for 3 h in a real-time PCR system (Mx3005P; Stratagene). The concentrations of all other components (initiation, elongation, termination factors; aminoacyl-tRNA synthetases; energy regenerating enzymes; ribosomes; amino acids; and low molecular weight compounds) are listed in Supplementary Tables S1 and S2. Note that although we used RNA as a template for the reaction, T7 RNA polymerase was included in the system to retain the ability to also use a DNA template. Filter sets used for measuring the fluorescence intensities of GFP and AlexaFluor647 were 492/516 and 635/665 nm (excitation/emission wavelength), respectively. AlexaFluor647 was used as an internal control to normalize the differences in fluorescence intensity among the wells. The day-to-day variation of the data (typically <20%) was normalized using the internal controls. For example, assume that the control sample gave a value of FI_{C1} and FI_{C2} on day 1 and 2, respectively. The data obtained on day 2 were normalized by multiplying FI_{C1}/FI_{C2} to the obtained values.

RNA preparation

The GFP DNA fragment was amplified by PCR using PYRObest DNA polymerase (Takara) according to the manufacturer's instructions using pETG5tag (Sunami *et al*, 2006) as a template with the primers T7F (5′-TAATACGACTCACTATAGGG-3′) and G5tCys (5′-TTATTAACA ACATCCTGGACAACATTTGTAGAGCTCATCCAT-3′). The GFP used was GFPuv5, which was constructed previously by Ito *et al* (1999). The resulting PCR products were used directly for *in vitro* transcription by adding 150 μg of PCR fragments to 800-μl mixtures consisting of 40 mM Tris–HCl (pH 8.0), 8 mM MgCl$_2$, 5 mM DTT, 2 mM spermidine, 0.4 mM NTPs, and 20 μg T7 RNA polymerase, and incubated at 37°C for 5 h. RNA was purified using an RNeasy Midi Kit (QIAGEN) following the manufacturer's instructions.

Bahadur expansion

Considering a set of all possible binary sequences with length n, we denote an arbitrary binary sequence by $\mathbf{x}=`x_1 x_2 \ldots x_n$,' where x_i typically takes 0 or 1 ($i=1,2,\ldots,n$), and we denote the set by $\mathbf{X}$. First, x_i is converted to z_i by:

$$z_i = z_i(x_i) = \begin{cases} -1, & \text{if } x_i = 0 \\ +1, & \text{if } x_i = 1 \end{cases} \qquad (m1)$$

Thus, we define the following function system:

$$\psi_0(\mathbf{x}) = 1$$
$$(\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots, \psi_n(\mathbf{x})) = (z_1, z_2, z_3, z_4, \ldots \ldots, z_n)$$
$$(\psi_{n+1}(\mathbf{x}), \psi_{n+2}(\mathbf{x}), \ldots, \psi_{n+_nC_2}(\mathbf{x})) = (z_1 z_2, z_1 z_3, \ldots \ldots, z_{n-1} z_n)$$
$$\vdots$$
$$\psi_{2^n-1}(\mathbf{x})) = z_1 z_2 z_3 z_4 \ldots \ldots z_n$$
$$(m2)$$

The set of functions $\{\psi_i(\mathbf{x})|i=0, 1, 2,\ldots, 2^n-1\}$ forms orthonormal bases of this vector space, that is, this function system satisfies the

following relationships:

$$\psi_i(\mathbf{x})^2 = 1$$
$$\frac{1}{2^n} \sum_{\mathbf{x} \in \mathbf{X}} \psi_i(\mathbf{x}) \psi_{i'}(\mathbf{x}) = \begin{cases} 1, & \text{if } i = i' \\ 0, & \text{if } i \neq i' \end{cases} \qquad (m3)$$
$$\frac{1}{2^n} \sum_{i=0}^{2^n-1} \psi_i(\mathbf{x}) \psi_i(\mathbf{x}') = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{x}' \\ 0, & \text{if } \mathbf{x} \neq \mathbf{x}' \end{cases}$$

Therefore, any function $f(\mathbf{x})$ is expanded as follows:

$$f(\mathbf{x}) = \sum_{i=0}^{2n-1} w_i \psi_i(\mathbf{x}) \qquad (m4)$$

where w_i is the Bahadur coefficient and is determined using:

$$w_i = \frac{1}{2^n} \sum_{\mathbf{x} \in \mathbf{X}} f(\mathbf{x}) \psi_i(\mathbf{x}) \qquad (m5)$$

An example for $n=4$ is shown in equation (2), which is shown as the sum of 1-, 2-, 3-, and 4-body interaction terms. Four-letter sequences, such as DNA, can be subjected to Bahadur expansion analysis (Arita *et al*, 2002). All calculations were carried out using Mathematica (Wolfram Research).

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Conflict of interest

The authors declare that they have no conflict of interest.

References

Aita T, Hamamatsu N, Nomiya Y, Uchiyama H, Shibanaka Y, Husimi Y (2002) Surveying a local fitness landscape of a protein with epistatic sites for the study of directed evolution. *Biopolymers* **64:** 95–105

Arita M, Tsuda K, Asai K (2002) Modeling splicing sites with pairwise correlations. *Bioinformatics* **18:** S27–S34

Arnold FH, Wintrode PL, Miyazaki K, Gershenson A (2001) How enzymes adapt: lessons from directed evolution. *Trends Biochem Sci* **26:** 100–106

Benos PV, Bulyk ML, Stormo GD (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res* **30:** 4442–4451

Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. *Nat Rev Genet* **8:** 437–449

Bulyk ML, Johnson PL, Church GM (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res* **30:** 1255–1261

Chester A, Weinreb V, Carter Jr CW, Navaratnam N (2004) Optimization of apolipoprotein B mRNA editing by APOBEC1 apoenzyme and the role of its auxiliary factor, ACF. *RNA* **10:** 1399–1411

Covert MW, Famili I, Palsson BO (2003) Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? *Biotechnol Bioeng* **84:** 763–772

Deutscher D, Meilijson I, Kupiec M, Ruppin E (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet* **38:** 993–998

Dill KA (1997) Additivity principles in biochemistry. *J Biol Chem* **272:** 701–704

Fisher RA (1966) *The Design of Experiments*, 8th edn. Edinburgh: London Oliver & Boyd

Humphreys K, Titterington DM (1999) The exploration of new methods for learning in binary Boltzmann machines. *Artif Intell Stat* **99:** 209–214

Ito Y, Suzuki M, Husimi Y (1999) A novel mutant of green fluorescent protein with enhanced sensitivity for microanalysis at 488 nm excitation. *Biochem Biophys Res Commun* **264:** 556–560

Jamshidi N, Palsson BO (2008) Formulating genome-scale kinetic models in the post-genome era. *Mol Syst Biol* **4:** 171

Kazuta Y, Adachi J, Matsuura T, Ono N, Mori H, Yomo T (2008) Comprehensive analysis of the effects of *Escherichia coli* ORFs on protein translation reaction. *Mol Cell Proteomics* **7:** 1530–1540

Kitano H (2004) Biological robustness. *Nat Rev Genet* **5:** 826–837

Losee Jr RM (1994) Term dependence: truncating the Bahadur–Lazarsfeld expansion. *Inf Process Manage* **30:** 293–303

Maier T, Ferbitz L, Deuerling E, Ban N (2005) A cradle for new proteins: trigger factor at the ribosome. *Curr Opin Struct Biol* **15:** 204–212

Man TK, Stormo GD (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res* **29:** 2471–2478

Matsuura T, Yomo T (2006) *In vitro* evolution of proteins. *J Biosci Bioeng* **101:** 449–456

Matsuura T, Yomo T, Trakulnaleamsai S, Ohashi Y, Yamamoto K, Urabe I (1998) Nonadditivity of mutational effects on the properties of catalase I and its application to efficient directed evolution. *Protein Eng* **11:** 789–795

Nierhaus KH, Wilson DN (2004) *Protein Synthesis and Ribosome Structure: Translating the Genome*. Weinheim: Wiley-VCH

Ohashi H, Shimizu Y, Ying BW, Ueda T (2007) Efficient protein selection based on ribosome display system with purified components. *Biochem Biophys Res Commun* **352:** 270–276

Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ (2007) Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445:** 383–386

Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2:** 886–897

Qin Y, Polacek N, Vesper O, Staub E, Einfeldt E, Wilson DN, Nierhaus KH (2006) The highly conserved LepA is a ribosomal elongation factor that back-translocates the ribosome. *Cell* **127:** 721–733

Shimizu Y, Inoue A, Tomari Y, Suzuki T, Yokogawa T, Nishikawa K, Ueda T (2001) Cell-free translation reconstituted with purified components. *Nat Biotechnol* **19:** 751–755

Smallbone K, Simeonidis E, Broomhead DS, Kell DB (2007) Something from nothing: bridging the gap between constraint-based and kinetic modelling. *FEBS J* **274:** 5576–5585

Solomon H (1961) *Studies in Item Analysis and Prediction*. Stanford, CA: Stanford University Press

Sunami T, Sato K, Matsuura T, Tsukada K, Urabe I, Yomo T (2006) Femtoliter compartment in liposomes for *in vitro* selection of proteins. *Anal Biochem* **357:** 128–136

Wells JA (1990) Additivity of mutational effects in proteins. *Biochemistry* **29:** 8509–8517

Wiedemann U, Boisguerin P, Leben R, Leitner D, Krause G, Moelling K, Volkmer-Engert R, Oschkinat H (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J Mol Biol* **343:** 703–718

Yin Y, Carter Jr CW (1996) Incomplete factorial and response surface methods in experimental design: yield optimization of tRNA(Trp) from *in vitro* T7 RNA polymerase transcription. *Nucleic Acids Res* **24:** 1279–1286

Young JS, Ramirez WF, Davis RH (1997) Modeling and optimization of a batch process for *in vitro* RNA production. *Biotechnol Bioeng* **56:** 210–220