Supporting Text

Derivation of the Algebraic Method of Spectral Deconstruction: Consider a population with an overall mutant frequency F, composed of $N = \Sigma N_i$ individuals with i mutations each ($i = 0, 1, 2 \ldots$), from which are sequenced $M_s = \Sigma M_i$ mutants with i mutations each ($i = 1, 2, 3 \ldots$). Assume that N contains two subfractions, S_1 and S_2, with respective mutation frequencies f_1 and f_2 ($f_1 < f_2$). Assume further that the mutations in each subpopulation are distributed at random so that $P(i) = f^i e^{-f}/i!$ and that subpopulation 1 does not contribute to the mutants carrying multiple mutations. Then ratio of triples to doubles among sequenced mutants will be $M_3 / M_2 = (NS_1 f_1^3 e^{-f_1}/ 6 + NS_2 f_2^3 e^{-f_2}/ 6) / (NS_1 f_1^2 e^{-f_1}/ 2 + NS_2 f_2^2 e^{-f_2}/ 2)$. Because subpopulation 1 makes a negligible contribution to the multiples, $NS_1 f_1^2 e^{-f_1} / 2 << NS_2 f_2^2 e^{-f_2} / 2$ and $NS_1 f_1^3 e^{-f_1} / 6 << NS_2 f_2^3 e^{-f_2} / 6$, so that $M_3 / M_2 \approx f_2 / 3$ or $f_2 \approx 3M_3 / M_2$. Because subpopulation 1 does not contribute, the total number of mutants with two mutations is $N_2 = S_2 N f_2^2 e^{-f_2} / 2$. The ratio of N_2 to all mutants is N_2 / FN and of M_2 to all sequenced mutants is M_2 / M_s. These two ratios are equal, so that $N_2 = FNM_2 / M_s$. Thus, $FNM_2 / M_s = S_2 N f_2^2 e^{-f_2} / 2$, which rearranges to $S_2 = 2FM_2 / M_s f_2^2 e^{-f_2}$. Because there are only two subpopulations, $S_1 = 1 - S_2$. The fraction of the population that contains no mutations is $1 - F = S_1 e^{-f_1} + S_2 e^{-f_2}$. Rearranging, $e^{-f_1} = (1 - F - S_2 e^{-f_2}) / S_1$, so that $f_1 = -\ln[(1 - F - S_2 e^{-f_2}) / S_1]$.

Distributions of Multiples in Genetic Space. The null hypothesis H_0 is that the two components of each double are distributed uniformly in genetic space, that is, they appear to be a random sample from the observed spectrum. Let p be the number of possible locations in the genetic space, and let n denote the number of doubles; in the present instance, $P = 60$ and $n = 21$. We then calculated the probability distribution of D_s under the null hypothesis. Let O_d and E_d denote the observed and the expected frequencies for a given d. E_d is obtained by using the probability distribution of D_s obtained under the null hypothesis. Then a test statistic to reject the null hypothesis is given by $T = \Sigma(O_d - E_d)^2 / E_d$ summed from $d = 1$ to $p - 1$. Because n is small relative to p, for several values of d, the observed frequency is zero and the expected frequency is very small. For this reason, we grouped the data from the reactions with accessory proteins into intervals. We computed the total observed frequency and the corresponding expected frequency within each group and then computed the test statistic T by using these frequencies. The value of the test statistic for the observed data are $T = 0.948$.

Because the total sample size is small, rather than using χ^2 tables to assess the statistical significance of this value of T, we obtained the critical values with bootstrap methodology. A total of 20,000 bootstrap samples were generated under the null hypothesis, and for each sample we computed the test statistic T. The bootstrap P value is then defined to be the proportion of bootstrap T values that exceed the observed value. This gave a P value of 0.97. Thus, we failed to reject the null hypothesis. In addition, the distance measure D_m is highly correlated with the distance measure D_s (the correlation coefficient between the two being $\approx$0.98) so that the two distance measures appear to be linearly related. Consequently, if the null hypothesis H_0 were tested by using D_m, we would expect the resulting P value also to be very high.

We did not perform a similar analysis for the data from the reaction without accessory proteins because n was only 12. However, after log-transforming the data to ensure approximate normality and homoscedasticity of variances, we compared the mean values of D_s between the populations with and without accessory proteins. Using a t test, we found no difference in the mean values ($P = 0.83$). Similarly, the difference between the mean values of D_m between the populations with and without accessory proteins was not significant ($P = 0.42$). The lower half of Table 4 summarizes the means and standard deviations for these two samples. In summary, the doubles observed in these two experiments appear to be composed of random samples of the underlying spectra.