

Functional Evolution within a Protein Superfamily

Research Article

Zhengping Yi,^{‡,#} Olga Vitek,[@] M. A. Qasim,[‡] Stephen M. Lu,^{‡,&}
Wuyuan Lu,^{‡,§} Michael Ranjbar,[‡] Jiangtian Li,[~] Michael C. Laskowski,[%]
Chris Bailey-Kellogg,^{^,*} and Michael Laskowski, Jr.[‡]

Departments of [‡]Chemistry, [@]Statistics, [~]Industrial Engineering, and [^]Computer Sciences,
Purdue University, West Lafayette, IN 47907-2038
[%] Department of Mathematics, University of Maryland

Current Addresses: [#]School of Life Sciences, Mail Code 4501, Arizona State University,
University Drive and Mill Avenue, Tempe, AZ 85287. zhengping.yi@asu.edu; [&]Ventria
Bioscience, 4110 N. Freeway Blvd., Sacramento, CA 95834; [§]Institute of Human Virology,
University of Maryland, Baltimore, MD 21201

* To whom correspondence should be addressed: Chris Bailey-Kellogg, 6211 Sudikoff
Laboratory, Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA.
Phone: 603-646-3385. Fax: 603-646-1672. Email: cbk@cs.dartmouth.edu.

Key words: functional evolution, protein superfamilies, functional conservation, sequence
hypervariability, specificity, protein-protein interactions

Running head: Functional Evolution within a Protein Superfamily

Abbreviations used:

| | |
|--------|--|
| CHYM | bovine chymotrypsin A |
| PPE | porcine pancreatic elastase |
| CARL | subtilisin Carlsberg |
| HLE | human leukocyte elastase |
| SGPA | <i>Streptomyces griseus</i> proteinase A |
| SGPB | <i>Streptomyces griseus</i> proteinase B |
| SRA | Sequence to Reactivity Algorithm |
| OM1 | avian ovomucoid first domain |
| OM3 | avian ovomucoid third domain |
| OMTKY3 | turkey ovomucoid third domain |
| BPTI | bovine pancreatic trypsin inhibitor |

ABSTRACT

The ability to predict and characterize distributions of reactivities over families and even superfamilies of proteins opens the door to an array of analyses regarding functional evolution. In this paper, insights into functional evolution in the Kazal inhibitor superfamily are gained by analyzing and comparing predicted association free energy distributions against six serine proteinases, over a number of groups of inhibitors: all possible Kazal inhibitors, natural avian ovomucoid first and third domains, and sets of Kazal inhibitors with statistically-weighted combinations of residues. The results indicate that, despite the great hypervariability of residues in the ten proteinase-binding positions, avian ovomucoid third domains evolved to inhibit enzymes similar to the six enzymes selected, while the orthologous first domains are not inhibitors of these enzymes on purpose. Hypervariability arises due to similarity in energetic contribution from multiple residue types; conservation is in terms of functionality, with “good” residues, which make positive or less deleterious contributions to the binding, selected more frequently and yielding overall the same distributional characteristics. Further analysis of the distributions indicates that while nature did optimize inhibitor strength, the objective may not have been the strongest possible inhibitor against one enzyme but rather an inhibitor that is relatively strong against a number of enzymes.

INTRODUCTION

The number of available protein sequences, now mostly translated from DNA sequences, is huge and growing very rapidly. Both their number and their rate of growth exceed greatly the available three-dimensional structures and the number of protein samples available for study. Therefore, algorithms for determination of protein reactivity (or function) from sequence alone are very much needed. These algorithms are likely to be most successful in dealing with groups of proteins, rather than the whole universe of proteins. The superfamilies of protein domains¹ appeared to us as logical groupings.

We have focused on the constituent domains of avian ovomucoids, and studied their reactivity with various serine proteinases.² Fig. 1 gives the complete sequence of an example ovomucoid, from turkey. Note that it has three domains. Most of our reactivity studies have been carried out on isolated third domains of avian ovomucoids and on recombinant variants of the third domains.³⁻⁶ Each of the ovomucoid domains functions as a standard mechanism,⁷ canonical⁸ protein inhibitor of serine proteinases of the Kazal superfamily of such inhibitors.⁹ The assignment to the Kazal superfamily suggests a function of inhibition of serine proteinases but both the specificities and the association constants of individual domains vary greatly against different serine proteinases. Some of them are strong inhibitors of trypsin, some of elastases, some of chymotrypsins and some are efficient inhibitors of several enzymes.³

Our laboratory has determined the sequences of third domains (OM3s) from 153 species.¹⁰ Due to polymorphism, the number of actual sequences is 159. Quite surprisingly, from the sequences for these 159 OM3s, it is clear that the subset of 10 positions in contact with the cognate enzyme fixes many more mutations than average for the entire molecule.¹⁰ Furthermore, this significant variability in the proteinase-inhibitor contact region is accompanied by huge changes in the inhibitor specificity and in the binding strength of the inhibitor to the enzyme. This creates a paradox since it is typically expected that structural and functional residues are strongly conserved.

In order to derive functional evolutionary explanations for the apparent paradox of contact residue hypervariability, this paper analyzes predicted reactivity distributions of Kazal superfamily inhibitors against various proteinases. The six selected enzymes, coming from a variety of sources, are bovine chymotrypsin A α (CHYM); porcine pancreatic elastases (PPE); subtilisin Carlsberg (CARL); *Streptomyces griseus* proteinases A and B (SGPA and SGPB, respectively); and human leukocyte elastases (HLE).⁴ All of them are strongly inhibited by the turkey ovomucoid third domain, OMTKY3 (Fig. 2),¹¹ which is the chosen wild type.⁴ To serve as the basis for comparing and contrasting the effects of functional evolutionary pressures on OM3s, we also study predicted reactivity distributions for a set of ovomucoid first domains (OM1s), as well as the set of all possible Kazal superfamily inhibitors. Our lab has determined ovomucoid first domains from 162 species of birds (Kato and Laskowski, unpublished data). Due to polymorphism, the number of actual sequences is 169. The set of $20^{10} \approx 10^{13}$ possible Kazal superfamily inhibitors is defined by all possible coded amino acid choices for the 10 contact residues. The sequence to reactivity algorithm (SRA) used to carry out the prediction is described in the Methods section below.

Analysis of the predicted reactivity distributions yields several significant insights into functional evolution within the Kazal superfamily:

1. Ovomuroid third domains have evolved to be excellent inhibitors of enzymes similar to the six enzymes selected. In contrast, OM1s did not evolve to be efficient inhibitors of the six enzymes selected.
2. While the OM3 family contains many very efficient inhibitors, its most efficient members are much weaker than the strongest possible members of the Kazal superfamily. While a number of suggested causes for this phenomenon are provided, we show that the reactivity gap between the strongest OM3 inhibitors and the strongest possible Kazal members narrows considerably for inhibitors which are simultaneously efficient against all six selected enzymes. The concept of “Meanzyme” was introduced to model this phenomenon. The ΔG° for an inhibitor against Meanzyme is defined as the average of the six ΔG° values for that inhibitor against all six enzymes studied. This ΔG° value serves as a measure of the overall strength of an inhibitor against a variety of enzymes.
3. Hypervariability of residues in the 10-position contact region does not imply lack of functional selection. In fact, evolutionary pressures on these residues are evident by analysis of distributions constructed from sets of sequences with statistically-weighted residues. Conservation is in terms of function: contribution to ΔG° . And naturally observed OM3 sequences are not random, but rather tend to be comprised of “good” residues, which make helpful or less deleterious contributions to the binding.

METHODS

A sequence to reactivity algorithm (SRA) was recently developed in our laboratory to predict the association standard free energies (ΔG°) of six serine proteinases with all possible members of the Kazal proteinase inhibitor superfamily subject to the restriction that either P2T or P1'E is present.^{3,4,6} The algorithm employs a data-driven first-order (or additive) model to predict ΔG° for an inhibitor. Experimental measurements of ΔG° were previously collected for each single substitution for each of the 10 variable contact residues. These measurements were used to determine $\Delta\Delta G^\circ$ s, which is the contribution of the single substitutions to the free energy of proteinase-inhibitor association (see equation 1). The total ΔG° under multiple substitutions is then predicted to be the sum of the $\Delta\Delta G^\circ$ values of the individual substitutions plus the ΔG° for the wild type, treating each substitution independently of the sequence context (equation 2).⁴

$$\Delta\Delta G^\circ(X_{wt} i X) = \Delta G^\circ(X_i) - \Delta G^\circ_{wt} \quad (1)$$

where i is the position where the replacement was made, X_{wt} is the wild type residue at this i th position and X the variant residue at this i th position.

$$\Delta G^\circ_{\text{predicted}} = \Delta G^\circ_{wt} + \sum_{i=1}^{10} \Delta\Delta G^\circ(X_{wt} i X) \quad (2)$$

Additivity in molecular recognition is not uncommon and has been studied by many research groups.¹²⁻¹⁸ For example, analysis of BPTI by alanine shaving clearly demonstrates additivity in BPTI-CHYM association.¹² Additivity was also employed to investigate the binding interactions between peptides and proteins of the class II major histocompatibility complex.¹⁵

Extensive tests have been performed to validate the algorithm in its application to Kazal Superfamily inhibitors. So far, there are 450 published cases where predicted and measured standard free energies were compared. Of these 289 (64%) were within experimental error of $2\sigma = 200$ cal/mol per substitution (a very tight assessment), 119 (26%) within $4\sigma = 400$ cal/mol per substitution, and only 42 (9%) fell outside these ranges.^{3,4,6}

Once an SRA has been developed and validated, as has been done by our laboratory for the Kazal superfamily inhibitors, entirely new approaches to study functionality are possible, as demonstrated in our previous paper published in *Proteins*.¹⁹ In the present work, we employ the published algorithms for efficient computation of distribution functions of reactivity of sets of Kazal inhibitors against the selected enzymes. We first characterize distribution functions for all possible Kazal superfamily inhibitors, and then turn to comparisons against the distribution functions of two orthologous families² within the Kazal superfamily, the avian ovomucoid third domains (OM3s) and first domains (OM1s). The computed distributions enable large-scale studies of similarities and differences in reactivity, and implications for evolution of function. Fig. 3 schematically depicts the distributions involved, and the characteristics of the distributions that support our functional evolutionary conclusions summarized in the **Introduction** section. It provides an intuitive summary of our approach for quick reference during the remainder of the paper. Figures and tables in the **Results** section detail the findings on the actual distributions, and the text describes the specific analysis steps.

RESULTS

Employing the sequence to reactivity algorithm, predictions of the standard free energies of association were carried out for the six selected enzymes interacting with all $20^8 \times 40 \approx 10^{12}$ possible members of the Kazal superfamily serine proteinase inhibitors (as always, subject to the restriction of P2T or P1'E) as well as for sequenced OM1s (98 out of 169) and OM3s (147 out of 159) which satisfy the restriction. Basic statistics of the distribution functions are provided in Table 1. Fig. 4 summarizes all distributions with “box-and-whisker” plots, and Fig. 5 provides several example plots of individual distributions.

The “all possible” Kazal distribution

Fig. 5a shows the distribution function of predicted ΔG° values for all possible members of the Kazal superfamily interacting with one of the six selected enzymes, SGPA. Plots for the other five enzymes are similar (see Fig. 4). The horizontal axis is labeled in an unconventional direction to preserve the expectation that the strongest inhibitors are on the right, and the weakest on the left. Our laboratory can confidently measure values of ΔG° between -4 kcal/mol ($K_a \approx 1 \times 10^3 \text{ M}^{-1}$) and -17.5 kcal/mol ($K_a \approx 1 \times 10^{13} \text{ M}^{-1}$). This is a 10 order of magnitude range, but it is completely dwarfed by the 43 kcal/mol (32 orders of magnitude) range from $\Delta G^\circ_{\text{max}}$ to $\Delta G^\circ_{\text{min}}$. The ranges are comparable for other enzymes we study (Table 1a). The area underneath the curve for inhibitors stronger (more negative) than -4 kcal/mol is marked in green and that for inhibitors weaker than -4 kcal/mol is marked in black. While the experimental techniques of measurement are improving, it seems highly unlikely that the black area can be eliminated. The -17.50 kcal/mol upper limit of measurement ($K_a = 1 \times 10^{13} \text{ M}^{-1}$) causes a lot of technical problems but no deep intellectual ones. Clearly, methods to measure beyond this limit can be devised and are being devised.

The “measurable” division (green dotted line in Fig. 4; black/green division in Fig. 5) allows us to ask what fraction of all possible Kazal inhibitors measurably ($\Delta G^\circ = -4.0$ kcal/mol to -17.5 kcal/mol) inhibits SGPA. The answer (64%) is surprisingly high. The value for SGPA is the highest among the 6 enzymes we tested; the lowest is 22% for PPE (Table 1a).

The ΔG° value of an inhibitor belonging to the measurable fraction will not satisfy many of our experimental colleagues, who regard millimolar and occasionally micromolar inhibitors as inefficient. We therefore introduce another dividing line in Figs. 4 and 5 for “efficient” inhibitors, at $\Delta G^\circ = -11$ kcal/mol ($K_a \approx 1 \times 10^8 \text{ M}^{-1}$). The “efficient fraction” is then the proportion of sequences predicted to react more strongly than this value. We see that the fraction of efficient inhibitors is only 6% even for SGPA and dips to 1% for PPE (Table 1a).

Table 1a further shows that the standard deviation of ΔG° is 3.96 kcal/mol for SGPA, and it averages 4.39 kcal/mol over all six enzymes. As would be expected from the large sample size (10^{12}), all six curves are almost normal, with skewness positive but very small.

The following subsections further analyze this distribution and related ones, in order to address significant questions of functional evolution within the Kazal inhibitor superfamily.

Comparison of distribution functions

The distribution curve of predicted ΔG° for 147 OM3 sequences against the six serine proteinases was calculated and the basic statistics are given in Table 1b. As an example, the distribution curve of ΔG° against SGPA is shown in red in Fig. 5b. The shape of the ovomucoid

third domain curve differs greatly from the all possible Kazal curve (reproduced for comparison in Fig. 5b, in solid black/green). The OM3 curve is heavily skewed, while the all possible Kazal curve is basically a normal distribution. OM3 curves for all six enzymes have large positive skewness, averaging about 1.68 (see also Fig. 4). The predicted ΔG° values are within the upper limit of the measurement range (at least -17.5 kcal/mol) while a very small portion of them (< 8%) are outside the lower limit of the measurement range (exceeding -4.0 kcal/mol).

Refer again to Fig. 3, step 1, for illustrations of metrics used to compare the distributions. The most striking thing is that the third domains are generally very good inhibitors of the six selected enzymes. The efficient fraction (with $\Delta G^\circ \leq -11$ kcal/mol, $K_a \geq 1 \times 10^8$ M⁻¹) is larger than 60% for OM3s against any enzyme, whereas the fraction is less than 7% for the set of all possible Kazal inhibitors. Table 1b indicates similar efficiency against the 6 enzymes. The 5th quantile is another statistical measure of a distribution, indicating the value for which only 5% of the distribution is smaller (stronger). This statistic is not “pulled” by inefficient values as much as, say, the mean would be, and thus provides a clear indication of the overall strength of the good sequences in the distribution. Table 2 and Fig. 4 show that the 5th quantile for the OM3 distribution is much stronger than that for all possible Kazal inhibitors, against any enzyme.

The distribution curve for the 98 OM1 sequences was also calculated; statistics are in Table 1c, distributions are in Fig. 4, and an example curve against SGPA is plotted in blue in Fig. 5b. Table 1c indicates that the skewness is negative for CHYM and PPE and positive for the others. Although the shape of the OM1 curve for SGPA clearly differs from that of the all possible Kazal curve, in contrast to OM3s, the efficient fraction for OM1s ($\leq 2\%$) is comparable to that for the all possible Kazal distribution ($\leq 6\%$). The 5th quantile (Table 2) is in fact worse than that of the all possible distribution, for example, -10.71 kcal/mol for OM1s against SGPA, compared to -11.32 kcal/mol for all possible and -15.51 kcal/mol for OM3s.

While the contact residues in OMs are considered hypervariable, we note that the variation is in fact different for OM1s and OM3s, thus explaining the significant differences in their reactivity distributions. Table 3 lists the amino acid residues with the greatest frequency at each position for OM1 and OM3. These differ substantially, and consequently so do their ΔG° values. In fact, the ΔG° values of such sequences for OM3 are much stronger than those for OM1.

The 147 OM3 sequences and 98 OM1 sequences are considered to be samples and representatives of all natural OM3s and OM1s. Thus we need to characterize the uncertainty of our estimates (efficient fraction and 5th quantile), which are based on the samples at hand but seek to characterize the set of all natural OM3s and OM1s. Since the distribution of predicted ΔG° values for the sets of natural sequences is not normal and therefore doesn't satisfy the assumptions for standard statistical tests, we used the bootstrap test, a resampling-based statistical method²⁰ to acquire information about the uncertainty of our statistical estimators. The bootstrap test chooses random samples with replacement from the set of predicted ΔG° values for natural sequences, such that the number of elements in each resample equals the number of elements in the original data set. The procedure generally needs to be repeated many times (a thousand or more).

The predicted ΔG° values for 147 natural OM3s were sampled 10000 times. The efficient fraction for each resample was recorded, and the minimum and maximum efficient fractions were obtained. As can be seen from Table 4, the Bootstrap test of the efficient fraction

shows that for SGPA, the minimum efficient fraction in an OM3 resample is 0.68; that is, at least 68% of predicted ΔG° values are more negative (stronger) than -11 kcal/mol. On the other hand, the maximum efficient fraction in an OM1 resample is only 0.09. This is strong quantitative evidence that the efficient fraction for OM3s is much larger than that for all possible Kazal inhibitors, while OM1s have efficiency close to that of the set of all possible Kazal inhibitors.

We also applied the Bootstrap test to the 5th quantile of the OM3 and OM1 samples. The results, provided in Table 5, indicate a large gap between the 5th quantile of the resamples for OM3 and OM1 families: the maximum 5th quantile (weakest) in 10000 OM3 resamples is -15.43 kcal/mol while the minimum 5th quantile (strongest) in 10000 OM1 resamples is only -11.41 kcal/mol. This is strong quantitative evidence that the natural OM1s have basically the same strength as predicted for the entire set of Kazal inhibitors (5th quantile for SGPA is -11.32 kcal/mol): about 95% of the ΔG° s for natural OM1s are within the same range of 95% of the ΔG° s for all possible Kazal inhibitors. On the contrary, the OM3s are overall stronger.

Reactivity strength

From Table 1, we can see that there is a large difference between the ΔG°_{\min} of the best inhibitor among the 147 OM3s (average -15.80 kcal/mol) and the ΔG°_{\min} of all possible inhibitors (average -21.61 kcal/mol) for each of the six enzymes. Further, note that the skewness for the OM3 distribution curve (see Figs. 4 and 5 and Table 1b) suggests that values of ΔG° significantly more negative than the current ΔG°_{\min} are unlikely to be found even if the set were to be expanded to cover all OM3s from approximately 10000 extant bird species.²¹ Another feature is that the six ΔG°_{\min} values for the 147 OM3 inhibitors are quite similar to each other, with the largest absolute difference being only 1.20 kcal/mol between SGPA and SGPB (Table 1b). In contrast, the largest absolute value difference between the ΔG°_{\min} values for all possible Kazal inhibitors is 5.25 kcal/mol (Table 1a). This large variation is probably a reflection of the phenomenon that some enzymes are easier to inhibit than others. The relative absence of such variation in natural sequences suggests that nature may have put a limit on the strength of inhibition of various enzymes.

On the basis of the available data, we can deal with one possible reason for inhibitors not being as strong as possible. This is because many of the inhibitors are defensive and are aimed against digestive enzymes of many different parasites. It is very likely that such inhibitors must have broad specificity against many different, albeit closely related enzymes. The broad specificity is achieved by some loss of maximal strength of inhibition against any particular enzyme. In order to model this phenomenon, the “Meanzyme” (mean enzyme) concept was introduced. The average of the six ΔG° values for an inhibitor against all six enzymes studied was calculated and defined as the ΔG° for that inhibitor against Meanzyme. This averaged ΔG° value serves as a measure of the overall strength of an inhibitor against a variety of enzymes, and relative nonspecificity in attaining that strength. Distributions of reactivities against Meanzyme, and corresponding statistics, can be calculated similarly to those against any other enzyme.¹⁹ Table 1 includes rows characterizing these Meanzyme distributions, Fig. 4 plots them, and step 2 in Fig. 3 illustrates a key to how we analyze the distribution against Meanzyme.

As can be seen from Table 1 and Fig. 4, the strongest nonspecific Kazal inhibitor (ΔG°_{\min} of all possible Kazal inhibitors against Meanzyme, -17.75 kcal/mol) is much weaker than the average ΔG°_{\min} against the six enzymes (-21.61 kcal/mol). However, for the 147 OM3s (Table 1b), the strongest nonspecific inhibitor (ΔG°_{\min} of OM3s against Meanzyme, -14.50 kcal/mol) is

only about 1 kcal/mol weaker than the average ΔG°_{\min} for the six enzymes (-15.80 kcal/mol). This “reactivity gap” is also much smaller than that between the strongest OM3 inhibitor and the strongest possible inhibitor against any particular enzyme (5.25 kcal/mol, as discussed above). These differences indicate that simultaneous strength is more important than absolute strength against a single enzyme. Furthermore, the difference between the ΔG°_{\min} of the best Meanzyme inhibitor among the 147 OM3s (-14.50 kcal/mol) and that among all possible inhibitors (-17.75 kcal/mol) is much smaller than the corresponding difference for the average of the six enzymes (-15.80 kcal/mol for OM3s and -21.61 kcal/mol for all Kazal). The smaller difference suggests that the simultaneous strength required of a strong Meanzyme inhibitor is significant, and the apparent reduction in the gap is not an artifact of averaging, which allows different inhibitors to be strong against different enzymes.

Functional evolutionary pressure

As mentioned before, the residues at the ten contact positions of OM3s are hypervariable.¹⁰ For example, in the 147 natural OM3s, at the P1 position there are 9 different amino acid residues, while at P2' there are 10 (see Table 6). Many sequences can be obtained by randomly combining just the observed residues at the ten positions, rather than allowing all 20 residue types. We call these the “natural OM3 combination” sequences. Since either P2T or P1'E has to be present, the total number of such sequences is $5 \times 2 \times 7 \times (6+1) \times 9 \times 10 \times 6 \times 9 \times 7 = 1.67E+7$. A resulting question is then whether the 147 natural sequences we have are essentially just randomly selected from this collection, or whether there is something special about them.

In order to determine if the 147 natural OM3s are in fact selected for functionality, we compared their ΔG° distribution function against that of the $1.67E+7$ natural OM3 combination sequences. Table 7a characterizes the combination distribution functions, and Fig. 3, step 3 provides intuition for our analysis. For all six enzymes and Meanzyme, the efficient fraction of the natural combination distribution is small (Table 7a) and comparable to that for the all possible Kazal inhibitor distribution (Table 1a). This stands in sharp contrast to the case of natural OM3s, which have efficient fractions greater than 60% (Table 1b), suggesting that the natural sequences are in fact much more optimized for inhibition of these enzymes and not simply selected randomly from the set of combined sequences. Furthermore, consider the plot in Fig. 6a of the OM3 combination distribution curve against SGPA (green and black). It is basically a normal distribution and has the same shape as the all possible Kazal curve (Fig. 5a) with small skewness, totally different from the heavily skewed natural OM3 curve (reproduced in red in Fig. 6a). These results all suggest that the OM3 combination distribution is similar to the Kazal distribution and greatly different from the natural OM3 distribution.

The OM3 sequences are hypervariable but apparently aren't simply randomly selected from a set of useful residues. The question remains as to what functional pressures were optimized by nature. As can be seen from Table 6, different residues have different frequencies at different positions. We can calculate a weight for each residue at each position by dividing the frequency of the residue at the position (entry in Table 6) by the number of total sequences (147 for OM3s). We can then weight a particular sequence by the product of the weights for its residues across the set of contact positions. This then allows us to calculate a “weighted natural combination distribution” by weighting each predicted ΔG° according to the sequence weight.

The weighted natural OM3 combination distribution was computed for each of $1.67E+7$ OM3 combination sequences. Table 7b summarizes the distribution statistics, while Fig. 6b plots the distribution against SGPA (black and green, with the natural OM3 distribution in red as

before). Comparison of this weighted distribution (Table 7b and Fig. 6b) with the unweighted one (Table 7a and Fig. 6a) shows a dramatic difference. The efficient fraction for the weighted distribution is greater than 60% for all six enzymes and Meanzyme, and the skewness is largely positive with an average of 1.10. The means of the weighted distribution and the natural OM3 distribution are almost identical. On the whole, the weighted distribution is very similar to the natural OM3 distribution, as the superimposition in Fig. 6b illustrates for the case of SGPA. The weighted combination curve looks like a smoothed natural OM3 curve, which is not unexpected. The difference between these two curves is partially due to the correlation of residues at different contact positions for the natural OM3 inhibitors. Such correlations do not affect the mean but do affect the overall distribution curve.

In order to demonstrate more clearly the difference between distributions without and with weights, let us consider one specific example. The P₁ position has been widely accepted as the most important functional position in serine proteinase inhibitors. The wild type OMTKY3 has P₁ Leu and there are other 8 residues appearing at P₁ position in the 147 OM3s. If these 9 residues are considered to be equally important, that is, having the same weight, then the average contribution of these nine P₁ residues to the binding to SGPA is 3.35 kcal/mol, which is deleterious because Leu is the second best residue at P₁. However, among the 147 OM3s, 63 have Leu and 59 have Met (the third best residue) at P₁. If the respective weights are assigned to the nine residues, the weighted average contribution to the binding becomes 0.70 kcal/mol, which is much better than 3.35 kcal/mol.

With the weighted OM3 natural combination distribution at hand, the probability of finding an inhibitor belonging to the OM3 family with stronger binding than the best OM3 naturally found so far can readily be obtained. The results are shown in Table 7. Although the probability for the weighted OM3 natural combination distribution is larger than that for the Kazal distribution (see “P best” in Table 7b and Table 1, respectively), it is still quite small for all six enzymes and Meanzyme (< 0.09). In addition, the difference (2.36 kcal/mol) of ΔG°_{\min} between natural OM3s (average -15.80 kcal/mol for 6 enzymes) and OM3 combination inhibitors (average -18.16 kcal/mol) is much smaller than that between natural OM3s and Kazal inhibitors, which is 5.81 kcal/mol. Finally, for Meanzyme, the best OM3 combination inhibitor (-15.94 kcal/mol) is closer in strength to the best found OM3 inhibitor (-14.50 kcal/mol) than is the strongest possible inhibitor among all Kazal inhibitors (-17.75 kcal/mol). Therefore, it looks like nature did a very good job in selecting the sequences for OM3.

The same procedure was applied to the analysis of predicted ΔG° distributions for natural OM1 combinations without and with weights. The observed frequency of amino acid residues at the ten contact positions for 98 natural OM1s is listed in Table 8. The total number of OM1 combination sequences is 8.4E+5 (4x2x5x(2+6)x5x5x7x5x3, accounting for the P₂T-P₁'E restriction). Characteristics of the predicted distribution functions are listed in Table 9. The predicted ΔG° distribution of the unweighted OM1 combination sequences (Table 9a) is quite similar to that of the Kazal superfamily (Table 1a), with a small efficient fraction and small skewness, indicating a normal distribution. Meanwhile, among the skewnesses for the weighted OM1 combination distribution for the six enzymes, some are positive, some are negative, and one is zero, and the average is 0.16 (Table 9b). In addition, the efficient fractions for the combination OM1 distributions without and with weights are almost identical to those for the natural OM1 distributions. Finally, there are 5 residues appearing at the P₁ position in the 98 OM1s. If the same weight were applied to these 5 residues, the average contribution to the binding to SGPA would be 4.53 kcal/mol. If the respective weights were assigned to the 5

residues, the weighted average contribution to the binding would be 4.41 kcal/mol, which is about the same. This is contrary to the result for P₁ residues for OM3s, indicating that nature optimized the residues at P₁ position for OM3s but not for OM1s.

DISCUSSION

The interaction of serine proteinases with their standard mechanism, canonical inhibitors is nearly lock and key rather than induced fit.⁴ The best possible coded residue at P₁ for SGPB is Leu which is present in a number of ovomucoid third domains including OMTKY3. The P₁ residue inserts in the S₁ pocket of SGPB in the inhibitor-enzyme complex.²² P₁Pro is the worst residue for SGPB. When P₁Leu is replaced by P₁Pro, the Pro residue also enters the S₁ cavity.²³ There is no frame shift that finds a better P₁ residue as happens in many protein ... protein associations, for example in protein substrates for serine proteinases. The lock and key behavior is not limited to the Kazal superfamily. In BPTI the P₁Arg variant is best and the P₁Gly variant worst for bovine β trypsin.²⁴ Both P₁Arg and P₁Gly react with the S₁ position of β trypsin.²⁵

All our distributions are based on the assumption that the lock and key association holds. The cases where ΔG° is positive involve several, not just one, deleterious substitutions. It is tempting to speculate that the lock and key behavior will no longer hold for multiple deleterious substitutions, but there is not, as yet, any clear experimental evidence for such behavior. Presumably, if lock and key did not hold, the very weak interactions in the black area would become stronger. This in turn would decrease the mean values in Table 1a for all possible Kazal inhibitors against all six enzymes as shown.

Since SGPA has a 64% measurable fraction (i.e. 64% of all possible Kazal inhibitors may have some inhibition against SGPA), it is a good candidate for testing inhibition from a new Kazal domain sequence. It is likely that any reasonable Kazal inhibitor will show some inhibition against SGPA, so an efficient testing strategy (experimentally or computationally) considers SGPA first.

Another interesting application of the superfamily wide distribution is to calibrate the strength of a new inhibitor. For example, if the new ΔG° is -17.50 kcal/mol, the probability of finding a Kazal inhibitor with a stronger ΔG° for SGPA is 6.9×10^{-5} (Table 10). In addition, the ΔG°_{\min} in Table 1a is the ΔG° value for the strongest possible inhibitor among the whole Kazal superfamily for an enzyme. Knowing the strength of the strongest possible inhibitor among a whole superfamily for an enzyme is obviously very helpful to the drug screening process.

Meanzyme is a new concept in our lab, proposed in order to gain intuition into what we have observed in a large number of inhibitor sequences as well as measured and predicted association equilibrium constants. The 159 sequences of ovomucoid third domains obtained in our lab contain 11 different residues at the P₁ position, the majority of which are Leu or Met but never an aromatic amino acid. An aromatic amino acid would make an inhibitor stronger for serine proteinases with large S₁ pockets but very bad for serine proteinases with small S₁ pockets. Nature has clearly chosen residues more suited to “Meanzyme” inhibition. In addition, in many cases it would seem advantageous for a system to optimize inhibitors that can inhibit a large number of closely related enzymes, rather than to be strictly specific and inhibit one or just a few enzymes. As an example, the role of plant inhibitors is mainly believed to be to inhibit the enzymes of the organisms that infest the plant. Since plants are susceptible to a variety of different types of infestations, an inhibitor that is directed towards a “Meanzyme” is desirable, as it would be effective against the various related serine proteinases the plant could encounter from these different types of infestations.

The Meanzyme concept was introduced to provide one very possible explanation for why nature didn't maximize the strength of natural inhibitors. There are some other explanations, which we now consider. (1) The maximal strength of natural inhibitors is very high already. The probability of finding a Kazal inhibitor in the all possible Kazal set with stronger binding than the best found OM3 inhibitor is quite small for all six enzymes ($< 2E-3$, Table 1b, column 10). It may be a waste of evolutionary energy to make stronger ones. (2) Excessively strong enzyme-inhibitor complexes may dissociate too slowly and may be too stable to undergo enzymatic degradation. (3) Inhibitors are often present along with enzyme zymogens. Very strong natural inhibitors already form complexes with zymogens. Presumably, stronger complexes would be undesirable and would rob the inhibitors of their control function. Some of these explanations were reviewed in a more general context.²⁶

The distribution functions of the predicted ΔG° values for OM3 and OM1 combination sequences were introduced to study the evolutionary pressures on the selection of residues at the contact positions. Comparison of the distributions for OM1 combination sequences without and with weights to those of all possible Kazal and 98 natural OM1s shows that the natural OM1s may just be a random sample from the unweighted natural OM1 distribution, and OM1s are not inhibitors on purpose for the six selected enzymes. It is very possible that they were optimized for some other enzymes. In contrast, comparison of the distributions for OM3 combination sequences without and with weights to those of all possible Kazal and 147 natural OM3s indicates that by selecting "good" residues (those which make helpful or less deleterious contributions to the binding) more frequently than "bad" residues, OM3s preserve their function as serine proteinase inhibitors against enzymes similar to the six proteinases selected. Because several residues can be considered "good," a position may appear to be quite variable. This answers the question of what is strongly conserved during the functional evolution, rather than sequence identity: the good functional residues.

Enormous effort has been made to experimentally measure the association constant K_a of isolated OM3s and OM1s against the six selected enzymes.⁴ More than 1200 K_a s have been measured in our lab. The results indicate that isolated OM3s generally inhibit the six selected proteinases efficiently. On the contrary, the results from a number of OM1s that were tested against the six enzymes indicated that OM1s usually are not efficient inhibitors against them. Only three OM1s we tested showed measurable inhibition against some of the six proteinases.⁴ However, OM1s may inhibit some other enzymes efficiently. For example, chicken ovomucoid first domain was found to be a good inhibitor of endoproteinase Lys-C with a K_a of about $1 \times 10^8 M^{-1}$.² It also inhibits trypsin-like enzyme from *Streptomyces erythraeus*.²⁷ Nonetheless, when chicken ovomucoid first domain was tested against the six proteinases, no inhibition was observed. The experimentally distinct inhibition for natural OM3s and OM1s against the six selected enzymes should not be surprising based on the predicted reactivity distributions for them. Experimental data agree with predictions very well here.

As discussed in the **Methods** section, the Sequence to Reactivity Algorithm (SRA) employs a data-driven first-order (or additive) model, which treats each substitution independently of the sequence context.⁴ Although cooperativity in molecular recognition is common, additivity in molecular recognition is not rare.¹²⁻¹⁸ Furthermore, we have found additivity to hold strikingly well in the Kazal superfamily.^{3,4,6} Thus we are confident in basing the results here on the predictions made by this algorithm, especially as our conclusions are drawn from analysis of large distributions displaying very significant differences.

In summary, ovomucoid third domains evolved to inhibit enzymes similar to the six enzymes we studied, in spite of the great hypervariability of residues in the ten proteinase-binding positions. Residues may not be strongly conserved during the course of evolution, but pressures nonetheless force combined selection from sets of “good” residues in order to preserve function. The function appears to balance absolute strength against any particular enzyme with simultaneous strength against a variety of enzymes.

ACKNOWLEDGMENTS

This research was supported at Purdue in part by funds from NIH grants GM10831 and GM63539, and a grant from the Showalter Trust. MCL was partially supported by NSF research grant DMS-0300080, and CBK was partially supported by NSF CAREER award IIS-0237654.

REFERENCES

1. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. *Science* 2003;300(5626):1701-1703.
2. Kato I, Schrode J, Kohr WJ, Laskowski M, Jr. Chicken ovomucoid: determination of its amino acid sequence, determination of the trypsin reactive site, and preparation of all three of its domains. *Biochemistry* 1987;26(1):193-201.
3. Laskowski M, Jr., Qasim MA, Yi Z. Additivity-based prediction of equilibrium constants for some protein-protein associations. *Curr Opin Struct Biol* 2003;13(1):130-139.
4. Lu SM, Lu W, Qasim MA, Anderson S, Apostol I, Ardelt W, Bigler T, Chiang YW, Cook J, James MN, Kato I, Kelly C, Kohr W, Komiyama T, Lin TY, Ogawa M, Otlewski J, Park SJ, Qasim S, Ranjbar M, Tashiro M, Warne N, Whatley H, Wieczorek A, Wieczorek M, Wilusz T, Wynn R, Zhang W, Laskowski M, Jr. Predicting the reactivity of proteins from their sequence alone: Kazal family of protein inhibitors of serine proteinases. *Proc Natl Acad Sci U S A* 2001;98(4):1410-1415.
5. Lu W, Apostol I, Qasim MA, Warne N, Wynn R, Zhang WL, Anderson S, Chiang YW, Ogin E, Rothberg I, Ryan K, Laskowski M, Jr. Binding of amino acid side-chains to S1 cavities of serine proteinases. *J Mol Biol* 1997;266(2):441-461.
6. Qasim MA, Lu W, Lu SM, Ranjbar M, Yi Z, Chiang YW, Ryan K, Anderson S, Zhang W, Qasim S, Laskowski M, Jr. Testing of the additivity-based protein sequence to reactivity algorithm. *Biochemistry* 2003;42(21):6460-6466.
7. Laskowski M, Jr., Kato I. Protein inhibitors of proteinases. *Annu Rev Biochem* 1980;49:593-626.
8. Bode W, Huber R. Natural protein proteinase inhibitors and their interaction with proteinases. *Eur J Biochem* 1992;204(2):433-451.
9. Laskowski M, Qasim MA. What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochim Biophys Acta* 2000;1477(1-2):324-337.
10. Apostol I, Giletto A, Komiyama T, Zhang W, Laskowski M, Jr. Amino acid sequences of ovomucoid third domains from 27 additional species of birds. *J Protein Chem* 1993;12(4):419-433.
11. Ardelt W, Laskowski M, Jr. Turkey ovomucoid third domain inhibits eight different serine proteinases of varied specificity on the same Leu18-Glu19 reactive site. *Biochemistry* 1985;24(20):5313-5320.
12. Buczek O, Koscielska-Kasprzak K, Krowarsch D, Dadlez M, Otlewski J. Analysis of serine proteinase-inhibitor interaction by alanine shaving. *Protein Sci* 2002;11(4):806-819.
13. Dill KA. Additivity principles in biochemistry. *J Biol Chem* 1997;272(2):701-704.
14. Giel-Pietraszuk M, Barciszewska MZ. Additivity of interactions of zinc finger motifs in specific recognition of RNA. *J Biochem (Tokyo)* 2002;131(4):571-578.
15. McFarland BJ, Beeson C. Binding interactions between peptides and proteins of the class II major histocompatibility complex. *Med Res Rev* 2002;22(2):168-203.
16. Mildvan AS. Inverse thinking about double mutants of enzymes. *Biochemistry* 2004;43(46):14517-14520.

17. Mildvan AS, Weber DJ, Kuliopulos A. Quantitative interpretations of double mutations of enzymes. *Arch Biochem Biophys* 1992;294(2):327-340.
18. Wells JA. Additivity of mutational effects in proteins. *Biochemistry* 1990;29(37):8509-8517.
19. Li J, Yi Z, Laskowski MC, Laskowski M, Jr., Bailey-Kellogg C. Analysis of sequence-reactivity space for protein-protein interactions. *Proteins* 2005;58(3):661-671.
20. Davison AC, Hinkley DV. *Bootstrap Methods and Their Applications*. Cambridge, UK: Cambridge University Press; 1997.
21. Monroe BL, Jr, Sibley CG. *A World Checklist of Birds*. New Haven and London: Yale University Press; 1997.
22. Fujinaga M, Read RJ, Sielecki A, Ardelt W, Laskowski M, Jr., James MN. Refined crystal structure of the molecular complex of *Streptomyces griseus* protease B, a serine protease, with the third domain of the ovomucoid inhibitor from turkey. *Proc Natl Acad Sci U S A* 1982;79(16):4868-4872.
23. Bateman KS, Huang K, Anderson S, Lu W, Qasim MA, Laskowski M, Jr., James MN. Contribution of peptide bonds to inhibitor-protease binding: crystal structures of the turkey ovomucoid third domain backbone variants OMTKY3-Pro18I and OMTKY3-psi[COO]-Leu18I in complex with *Streptomyces griseus* proteinase B (SGPB) and the structure of the free inhibitor, OMTKY-3-psi[CH₂NH₂⁺]-Asp19I. *J Mol Biol* 2001;305(4):839-849.
24. Krowarsch D, Dadlez M, Buczek O, Krokoszynska I, Smalas AO, Otlewski J. Interscaffolding additivity: binding of P1 variants of bovine pancreatic trypsin inhibitor to four serine proteases. *J Mol Biol* 1999;289(1):175-186.
25. Helland R, Otlewski J, Sundheim O, Dadlez M, Smalas AO. The crystal structures of the complexes between bovine beta-trypsin and ten P1 variants of BPTI. *J Mol Biol* 1999;287(5):923-942.
26. Brooijmans N, Sharp KA, Kuntz ID. Stability of macromolecular complexes. *Proteins* 2002;48(4):645-653.
27. Nagata K, Yoshida N. Interaction between trypsin-like enzyme from *Streptomyces erythraeus* and chicken ovomucoid. *J Biochem (Tokyo)* 1984;96(4):1041-1049.
28. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* 1967;27(2):157-162.
29. Fujinaga M, Sielecki AR, Read RJ, Ardelt W, Laskowski M, Jr., James MN. Crystal and molecular structures of the complex of alpha-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 Å resolution. *J Mol Biol* 1987;195(2):397-418.
30. Bode W, Wei AZ, Huber R, Meyer E, Travis J, Neumann S. X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. *Embo J* 1986;5(10):2453-2458.

FIGURE LEGENDS

Figure 1. Complete sequence of turkey ovomucoid. Scissors indicate Spase V8 and CNBr cleavage sites, arrows indicate reactive site peptide bonds and bended lines indicate glycosylation positions. Nine disulfide bridges are indicated by bars. Schechter and Berger²⁸ notation starts at the reactive site peptide bond indicated by the arrow. The subscript indicates the distance from this bond. Residues preceding the reactive site are labeled $P_1, P_2 \dots P_n$; those following it $P_1', P_2' \dots P_n'$.

Figure 2. Covalent structure of turkey ovomucoid third domain, OMTKY3, a Kazal superfamily protein inhibitor. The three disulfide bridges are indicated by bars. Residues are numbered sequentially from Val⁶ to Cys⁵⁶. Schechter and Berger²⁸ notation starts at the reactive site peptide bond indicated by the arrow. The binding sites on the enzyme for these inhibitor residues are often referred to as S_1, S_2, \dots, S_n and S_1', S_2', \dots, S_n' respectively. X-ray structures revealed that in complexes of OMTKY3 with CHYM,²⁹ SGPB²² and HLE³⁰ there is a clear consensus of 12 contact residues. These are indicated in color. Of these 2 marked in green are clearly structural. The structural role of P_3 Cys¹⁶ is obvious, the P_{15}' Asn³³ side chain makes 3 hydrogen bonds to other residues. The remainder, called the variable contact residue set, is marked in blue. Among them are the most variable positions in the molecule¹⁰. Changing these residues generally results in large changes in binding.

Figure 3. Schematic overview of key steps in distribution analysis. (1) Comparison of the “all possible” distribution, the OM3 distribution, and the OM1 distribution indicates that, despite hypervariability in the contact residues, OM3 is indeed optimized for inhibition of these types of enzymes. We compare the measurable and efficient fractions (percentage of sequences stronger than -4 and -11 kcal/mol, respectively), as well as the 5th quantile (ΔG° value such that 5% of the distribution is stronger). (2) Comparison of the OM3 distribution against the real enzyme SGPA and the “averaged” enzyme, Meanzyme, shows that rather than seeking the strongest possible inhibitor, nature may have been optimizing for overall strength against an array of enzymes. The “reactivity gap” between the best observed and the best possible inhibitors is much smaller when considering simultaneous inhibition (i.e. against Meanzyme). (3) While there is significant variability in the contact residues, a form of functional conservation is still at work. Simply combining uniformly the observed residue types in OM3 yields a distribution much like the “all possible” one, while weighting the combination according to frequency of observation yields a distribution much like the natural OM3 one. What is strongly conserved during functional evolution, rather than sequence identity, is the use of good functional residues.

Figure 4. Distributions of predicted standard free energies of association against six enzymes and Meanzyme. (a) All possible Kazal inhibitors; (b) 147 OM3s; and (c) 98 OM1s. The plot for a particular enzyme includes a box from the first quartile to the third quartile, with a horizontal line at the median. Running vertically from the box are dotted-line whiskers capturing the extent of the non-outliers; here we show the absolute minimum and maximum of the distribution. An asterisk marks the 5th quantile of the distribution. The plots make apparent the

fact that OM3s tend to be much better inhibitors, with much more of their distributions stronger (more negative) than the -11 kcal/mol "efficient", shown with horizontal dotted lines. The plots also highlight the relative skew of the OM3 distributions toward strong binders, as compared to the other distributions (median is much closer to the minimum than to the maximum for all enzymes studied). This effect is particularly pronounced against Meanzyme, which represents simultaneous effectiveness against all six enzymes.

Figure 5. Example distributions of predicted standard free energies of association. (a) SGPA interacting with all possible members of the Kazal family; (b) SGPA interacting with all possible members of the Kazal family, OM3s (red), and OM1s (blue); (c) Meanzyme interacting with all possible members of the Kazal family, OM3s (red), and OM1s (blue). All predictions are subject to the P₂T-P₁'E restriction. Note the unconventional direction of the X axis, which was done to make strong binding be on the right and weak binding on the left. ΔG°_{\min} is the value for the strongest possible SGPA inhibitor in the Kazal family and ΔG°_{\max} is the weakest possible. The lower limit of measurement is -4 kcal/mol. The measurable area (green) comprises 64% of the total. The black area is where the interactions are too weak to measure. The red curve is for 147 OM3s isolated from different species of birds while the blue one is for 98 sequenced avian OM1s. These two distributions are also calculated, because the lowest values (in the black area) could not be measured. The curves are all normalized to unit area. The ΔG° for an inhibitor against Meanzyme is defined as the average of the inhibitor's ΔG° against the six enzymes studied.

Figure 6. Distributions of predicted standard free energies of association of SGPA for sets of inhibitors. (a) natural OM3 combinations (black and green) and 147 OM3s (red); (b) natural OM3 combinations with weights (black and green) and 147 OM3s (red). Natural OM3 combination sequences are obtained by considering all combinations of the residues appearing at each of the ten contact positions of 147 OM3s, subject to the P₂T-P₁'E restriction. The total number of combination sequences is 1.67E+7 (see text). The distribution of predicted ΔG° values for the OM3 combination sequences with weights is obtained by applying a weight for each sequence's ΔG° , according to the frequency of its residues (see Table 6). For direct comparison, the red curve for 147 OM3s is the same as that in Figure 5b.

Table 1 Distributions of the predicted ΔG° for 1E+12 all possible Kazal inhibitors, 147 OM3s and 98 OM1s

| Enzyme | ΔG°_{\min} (kcal/mol) | ΔG°_{\max} (kcal/mol) | Measurable fraction (%) ^a | Efficient fraction (%) ^b | Mean (kcal/mol) | Standard deviation (kcal/mol) | Skewness | $\Delta(\Delta G^\circ_{\min})$ (kcal/mol) | P best ^c |
|---|---------------------------------------|---------------------------------------|--|---|--------------------|-------------------------------------|-------------|---|------------------------|
| (a) Kazal (1E+12) | | | | | | | | | |
| CHYM | -23.75 | 27.24 | 28 | 2 | -1.36 | 4.63 | 0.03 | | |
| PPE | -18.50 | 25.45 | 22 | 1 | 0.22 | 4.97 | 0.14 | | |
| CARL | -22.89 | 25.72 | 45 | 4 | -3.23 | 4.49 | 0.25 | | |
| SGPA | -21.49 | 21.36 | 64 | 6 | -5.16 | 3.96 | 0.34 | | |
| SGPB | -20.86 | 21.14 | 57 | 3 | -4.41 | 3.76 | 0.46 | | |
| HLE | -22.14 | 22.92 | 31 | 2 | -1.64 | 4.53 | 0.04 | | |
| Average | -21.61 | 23.97 | 41 | 3 | -2.60 | 4.39 | 0.21 | | |
| Meanzyme | -17.75 | 20.64 | 36 | 0 | -2.60 | 3.54 | 0.32 | | |
| (b) OM3s (147) OM3-Kazal | | | | | | | | | |
| CHYM | -16.31 | 1.18 | 93 | 64 | -11.40 | 3.92 | 1.21 | 7.44 | 2.61E-04 |
| PPE | -15.95 | 1.86 | 95 | 81 | -12.25 | 3.80 | 2.13 | 2.55 | 3.86E-06 |
| CARL | -15.29 | 0.82 | 95 | 73 | -11.73 | 3.61 | 1.52 | 7.60 | 1.64E-03 |
| SGPA | -16.43 | 0.34 | 98 | 84 | -13.00 | 3.26 | 1.79 | 5.06 | 4.18E-04 |
| SGPB | -15.23 | 1.51 | 98 | 65 | -11.82 | 3.02 | 1.79 | 5.63 | 3.31E-04 |
| HLE | -15.58 | -0.92 | 96 | 75 | -11.24 | 3.27 | 1.66 | 6.56 | 3.40E-04 |
| Average | -15.80 | 0.80 | 96 | 74 | -11.91 | 3.48 | 1.68 | 5.81 | 4.99E-04 |
| Meanzyme | -14.50 | -0.74 | 94 | 78 | -11.91 | 3.20 | 1.79 | 3.25 | 3.47E-05 |
| (c) OM1s (98) OM1-Kazal | | | | | | | | | |
| CHYM | -11.27 | 0.33 | 46 | 1 | -3.66 | 1.88 | -0.47 | 12.48 | 0.016 |
| PPE | -7.12 | 3.36 | 4 | 0 | -1.85 | 1.55 | -0.17 | 11.38 | 0.066 |
| CARL | -7.03 | 1.56 | 78 | 0 | -4.71 | 1.82 | 1.44 | 15.86 | 0.200 |
| SGPA | -11.41 | -0.98 | 94 | 2 | -8.68 | 2.30 | 1.51 | 10.08 | 0.048 |
| SGPB | -11.28 | -0.14 | 93 | 2 | -7.37 | 2.21 | 1.06 | 9.58 | 0.022 |
| HLE | -7.49 | 1.21 | 30 | 0 | -3.30 | 1.49 | 0.30 | 14.65 | 0.102 |
| Average | -9.27 | 0.89 | 58 | 1 | -4.93 | 1.88 | 0.61 | 12.34 | 0.076 |
| Meanzyme | -7.59 | -0.74 | 77 | 0 | -4.93 | 1.41 | 1.11 | 10.16 | 0.072 |

^a: $\Delta G^\circ \leq -4$ kcal/mol ($K_a \approx 1 \times 10^3$ M⁻¹)^b: $\Delta G^\circ \leq -11$ kcal/mol ($K_a \approx 1 \times 10^8$ M⁻¹)^c: The probability of finding a Kazal inhibitor stronger than the best found OM3/OM1, computed as the area under the curve of the "all possible" Kazal distributions to the right of ΔG°_{\min} for natural 147 OM3s and 98 OM1s

Table 2 The 5th quantile of distributions of predicted ΔG° values (kcal/mol) for 1E+12 all possible Kazal inhibitors, 147 OM3s and 98 OM1s

| Enzyme | CHYM | PPE | CARL | SGPA | SGPB | HLE |
|--------|--------|--------|--------|--------|--------|--------|
| Kazal | -9.06 | -7.68 | -10.34 | -11.32 | -10.12 | -9.08 |
| OM3 | -15.17 | -15.20 | -14.68 | -15.51 | -14.46 | -14.36 |
| OM1 | -5.78 | -3.60 | -6.17 | -10.71 | -9.86 | -5.58 |

* 5th quantile: 5% of the ΔG° values in the distribution are more negative (stronger) than the 5th quantile value.

Table 3 Amino acid residues with the greatest frequency at each of the 10 contact positions for 147 OM3s and 98 OM1s as well as predicted ΔG° values (kcal/mol) of the inhibitors comprised of these choices.

| | P6 | P5 | P4 | P2 | P1 | P1' | P2' | P3' | P14' | P18' | CHYM | PPE | CARL | SGPA | SGPB | HLE |
|-----|----|----|----|----|----|-----|-----|-----|------|------|--------|--------|--------|--------|--------|--------|
| OM3 | K | P | A | T | L | E | Y | M | G | N | -12.64 | -14.84 | -13.78 | -15.57 | -14.38 | -13.96 |
| OM1 | V | L | V | T | K | E | L | S | S | L | -4.86 | -2.84 | -5.29 | -10.73 | -9.05 | -4.05 |

Table 4 Bootstrap test of the efficient fraction of the natural OM3 and OM1 distributions of predicted ΔG° values (kcal/mol) for SGPA

| SGPA | Efficient fraction ^a | Efficient fraction range for resamples ^b | |
|-------|---------------------------------|---|------|
| | | min | max |
| Kazal | 0.06 | | |
| OM3 | 0.84 | 0.68 | 0.94 |
| OM1 | 0.02 | 0.00 | 0.09 |

^a: $\Delta G^\circ \leq -11$ kcal/mol ($K_a \cong 1 \times 10^8$ M⁻¹)

^b: The 147/98 predicted ΔG° s for the OM3/OM1 were sampled 10000 times respectively, the efficient fraction for each resample was recorded, and minimum and maximum efficient fractions were obtained

* For other enzymes, the results for the efficient fraction range are similar with a large gap between OM3 and OM1 families

Table 5 Bootstrap test of the 5th quantile of the natural OM3 and OM1 distribution of predicted ΔG° values (kcal/mol) for SGPA

| SGPA | 5th quantile ^a | 5th quantile for resamples ^b | |
|-------|---------------------------|---|--------|
| | | min | max |
| Kazal | -11.32 | | |
| OM3 | -15.51 | -15.81 | -15.43 |
| OM1 | -10.71 | -11.41 | -10.50 |

^a: 5% of the ΔG° values in the distribution are more negative (stronger) than the 5th quantile.

^b: The 147/98 predicted ΔG° s for the OM3/OM1 were sampled 10000 times respectively, the 5th quantile for each resample was recorded, and minimum and maximum 5th quantiles were obtained

* For other enzymes, the results for the range of the 5th quantile are similar with a large gap between OM3 and OM1 families.

Table 6 Observed frequency of amino acid residues at 10 contact positions for 147 OM3s

| Position | Residue and respective frequency | | | | | | | | | | Kinds of Residues | | | | | | | | | | |
|----------|----------------------------------|-----|---|----|---|----|---|---|---|---|-------------------|---|---|---|---|---|---|---|---|---|----|
| P6 | K | 137 | Q | 4 | R | 3 | T | 2 | M | 1 | | 5 | | | | | | | | | |
| P5 | P | 145 | H | 2 | | | | | | | | 2 | | | | | | | | | |
| P4 | A | 103 | V | 27 | D | 10 | G | 2 | T | 2 | E | 2 | S | 1 | 7 | | | | | | |
| P2 | T | 116 | S | 23 | L | 5 | A | 1 | M | 1 | R | 1 | | | 6 | | | | | | |
| P1 | L | 63 | M | 59 | A | 7 | V | 5 | Q | 4 | T | 3 | P | 3 | S | 2 | I | 1 | 9 | | |
| P1' | E | 143 | D | 4 | | | | | | | | | | | | | | | 2 | | |
| P2' | Y | 118 | D | 9 | Q | 5 | E | 4 | H | 4 | S | 2 | L | 2 | N | 1 | F | 1 | R | 1 | 10 |
| P3' | M | 67 | R | 62 | K | 7 | F | 6 | V | 5 | T | 1 | | | | | | | | | 6 |
| P14' | G | 81 | S | 40 | D | 13 | A | 6 | N | 3 | V | 1 | P | 1 | H | 1 | R | 1 | | | 9 |
| P18' | N | 112 | D | 18 | S | 8 | A | 4 | G | 2 | T | 2 | Y | 1 | | | | | | | 7 |

* Total sequences: the product of the number of residue kinds at each position, subject to the P2T-P1'E restriction.

* Weight: frequency of a residue at one position divided by 147, the total number of sequences.

* Weight for a ΔG° value: the product of the corresponding weights of the residues at each of the 10 contact positions.

Table 7 Distributions of the predicted ΔG° for natural OM3 combinations without /with weight ^a

| Enzyme | ΔG°_{\min} (kcal/mol) | ΔG°_{\max} (kcal/mol) | Measurable fraction (%) ^b | Efficient fraction (%) ^c | Mean (kcal/mol) | Standard deviation (kcal/mol) | Skewness | $\Delta(\Delta G^\circ_{\min})$ (kcal/mol) | P best ^d |
|--|---------------------------------------|---------------------------------------|--|---|--------------------|-------------------------------------|-------------|---|------------------------|
| (1) Unweighted OM3 combinations, 1.67E+7 | | | | | | | | OM3 comb -Kazal | |
| CHYM | -18.44 | 13.57 | 53 | 5 | -4.20 | 4.09 | 0.09 | 5.31 | 5.76E-04 |
| PPE | -17.00 | 13.93 | 58 | 5 | -4.60 | 3.99 | 0.26 | 1.50 | 4.72E-05 |
| CARL | -16.56 | 12.92 | 53 | 3 | -3.97 | 4.00 | 0.33 | 6.33 | 7.23E-05 |
| SGPA | -19.02 | 8.43 | 76 | 8 | -6.29 | 3.55 | 0.31 | 2.47 | 4.47E-04 |
| SGPB | -17.59 | 7.86 | 71 | 3 | -5.51 | 3.15 | 0.26 | 3.27 | 2.28E-04 |
| HLE | -20.33 | 11.08 | 69 | 10 | -5.90 | 3.96 | 0.14 | 1.81 | 3.91E-03 |
| <i>Average</i> | -18.16 | 11.30 | 63 | 6 | -5.08 | 3.79 | 0.23 | 3.45 | 8.80E-04 |
| Meanzyme | -15.94 | 8.74 | 65 | 3 | -5.08 | 3.22 | 0.31 | 1.81 | 2.22E-04 |
| (2) weighted OM3 combinations | | | | | | | | Weighted comb -147 OM3s | |
| CHYM | -18.44 | 13.57 | 97 | 63 | -11.42 | 3.29 | 0.89 | -2.13 | 0.027 |
| PPE | -17.00 | 13.93 | 99 | 72 | -12.26 | 2.82 | 1.26 | -1.05 | 0.002 |
| CARL | -16.56 | 12.92 | 98 | 68 | -11.74 | 2.93 | 1.24 | -1.27 | 0.024 |
| SGPA | -19.02 | 8.43 | 99 | 80 | -13.02 | 2.56 | 1.23 | -2.59 | 0.023 |
| SGPB | -17.59 | 7.86 | 99 | 70 | -11.84 | 2.36 | 1.11 | -2.36 | 0.018 |
| HLE | -20.33 | 11.08 | 98 | 63 | -11.25 | 2.83 | 0.86 | -4.75 | 0.040 |
| <i>Average</i> | -18.16 | 11.30 | 98 | 69 | -11.92 | 2.80 | 1.10 | -2.36 | 0.022 |
| Meanzyme | -15.94 | 8.74 | 99 | 72 | -11.93 | 2.44 | 1.25 | -1.44 | 0.085 |

^a: Either P2T or P1'E has to be present in the Sequence

^b: $\Delta G^\circ \leq -4$ kcal/mol ($K_a \approx 1 \times 10^3$ M⁻¹)

^c: $\Delta G^\circ \leq -11$ kcal/mol ($K_a \approx 1 \times 10^8$ M⁻¹)

^d: The probability of finding a natural OM3 inhibitor stronger than the best found OM3, computed as the area under the curves of unweighted/weighted combination distributions to the right of the ΔG°_{\min} for 147 natural OM3s (see Table 1b)

Table 8 Observed frequency of amino acid residues at 10 contact positions for 98 OM1s

| Position | Residue and respective frequency | | | | | | | | | | | | Kinds of Residues | | |
|----------|----------------------------------|----|---|----|---|----|---|---|---|---|---|---|-------------------|---|---|
| P6 | V | 92 | A | 3 | E | 2 | G | 1 | | | | | | 4 | |
| P5 | L | 94 | V | 4 | | | | | | | | | | 2 | |
| P4 | V | 52 | L | 32 | A | 11 | D | 2 | I | 1 | | | | 5 | |
| P2 | T | 92 | P | 6 | | | | | | | | | | 2 | |
| P1 | K | 69 | E | 22 | Q | 3 | T | 2 | D | 2 | | | | 5 | |
| P1' | E | 31 | D | 30 | I | 25 | S | 6 | T | 3 | A | 2 | N | 1 | 7 |
| P2' | L | 86 | V | 7 | I | 2 | R | 2 | S | 1 | | | | 5 | |
| P3' | S | 66 | R | 16 | H | 9 | Q | 3 | I | 2 | A | 1 | L | 1 | 7 |
| P14' | S | 90 | T | 3 | G | 2 | N | 2 | M | 1 | | | | 5 | |
| P18' | L | 87 | S | 7 | M | 4 | | | | | | | | 3 | |

* Total sequences: the product of the number of residue kinds at each position, subject to the P2T-P1'E restriction.

* Weight: frequency of a residue at one position divided by 98, the total number of sequences.

* Weight for a ΔG° value: the product of the corresponding weights of the residues at each of the 10 contact positions.

Table 9 Distributions of the predicted ΔG° for natural OM1 combinations without /with weight ^a

| Enzyme | ΔG°_{\min} (kcal/mol) | ΔG°_{\max} (kcal/mol) | Measurable fraction (%) ^b | Efficient fraction (%) ^c | Mean (kcal/mol) | Standard deviation (kcal/mol) | Skewness | $\Delta(\Delta G^\circ_{\min})$ (kcal/mol) | P Best ^d |
|---|---------------------------------------|---------------------------------------|--|---|--------------------|-------------------------------------|--------------|---|------------------------|
| (1) Unweighted OM1 combinations, 8.4E+5 | | | | | | | | OM1 comb -Kazal | |
| CHYM | -13.57 | 7.59 | 23 | 0 | -1.94 | 2.81 | -0.14 | 10.18 | 4.79E-04 |
| PPE | -13.20 | 7.82 | 26 | 0 | -1.62 | 3.75 | -0.36 | 5.30 | 0.106 |
| CARL | -12.72 | 11.38 | 43 | 0 | -3.00 | 3.78 | 0.45 | 10.17 | 0.140 |
| SGPA | -14.70 | 2.33 | 74 | 2 | -5.76 | 2.59 | -0.05 | 6.79 | 0.014 |
| SGPB | -14.47 | 3.49 | 68 | 1 | -5.26 | 2.57 | -0.01 | 6.39 | 0.008 |
| HLE | -15.26 | 8.37 | 32 | 1 | -2.42 | 3.53 | -0.23 | 6.88 | 0.088 |
| <i>Average</i> | -13.99 | 6.83 | 45 | 1 | -3.33 | 3.17 | -0.06 | 7.62 | 0.059 |
| Meanzyme | -12.71 | 5.28 | 40 | 0 | -3.34 | 2.62 | -0.05 | 5.04 | 0.055 |
| (2) weighted OM1 combinations | | | | | | | | Weighted comb -98 OM1s | |
| CHYM | -13.57 | 7.59 | 46 | 0 | -3.86 | 2.26 | 0.00 | -2.30 | 2.22E-04 |
| PPE | -13.20 | 7.82 | 7 | 0 | -1.98 | 1.91 | -1.14 | -6.08 | 0.023 |
| CARL | -12.72 | 11.38 | 80 | 0 | -4.92 | 1.86 | 1.16 | -5.69 | 0.085 |
| SGPA | -14.70 | 2.33 | 99 | 4 | -8.84 | 1.69 | 0.77 | -3.29 | 0.017 |
| SGPB | -14.47 | 3.49 | 96 | 1 | -7.50 | 1.79 | 0.69 | -3.19 | 0.003 |
| HLE | -15.26 | 8.37 | 34 | 0 | -3.44 | 1.92 | -0.51 | -7.77 | 0.043 |
| <i>Average</i> | -13.99 | 6.83 | 60 | 1 | -5.09 | 1.91 | 0.16 | -4.72 | 0.029 |
| Meanzyme | -12.71 | 5.28 | 78 | 0 | -5.09 | 1.49 | 0.36 | -5.12 | 0.038 |

^a: Either P2T or P1'E has to be present in the Sequence^b: $\Delta G^\circ \leq -4$ kcal/mol ($K_a \approx 1 \times 10^3$ M⁻¹)^c: $\Delta G^\circ \leq -11$ kcal/mol ($K_a \approx 1 \times 10^8$ M⁻¹)^d: The probability of finding a natural OM1 inhibitor stronger than the best found OM1, computed as the area under the curves of unweighted/weighted combination distributions to the right of the ΔG°_{\min} for 98 natural OM1s (see Table 1c)

Table 10 The probability of finding a Kazal inhibitor stronger than $\Delta G^\circ = -17.50$ kcal/mol, computed as the area under the distribution curve of “all possible” sequences to the right of -17.50

| Enzyme | CHYM | PPE | CARL | SGPA | SGPB | HLE |
|-------------|---------|---------|---------|---------|---------|---------|
| Probability | 6.2E-05 | 1.2E-08 | 1.2E-04 | 6.9E-05 | 6.5E-06 | 2.0E-05 |

Figure 1

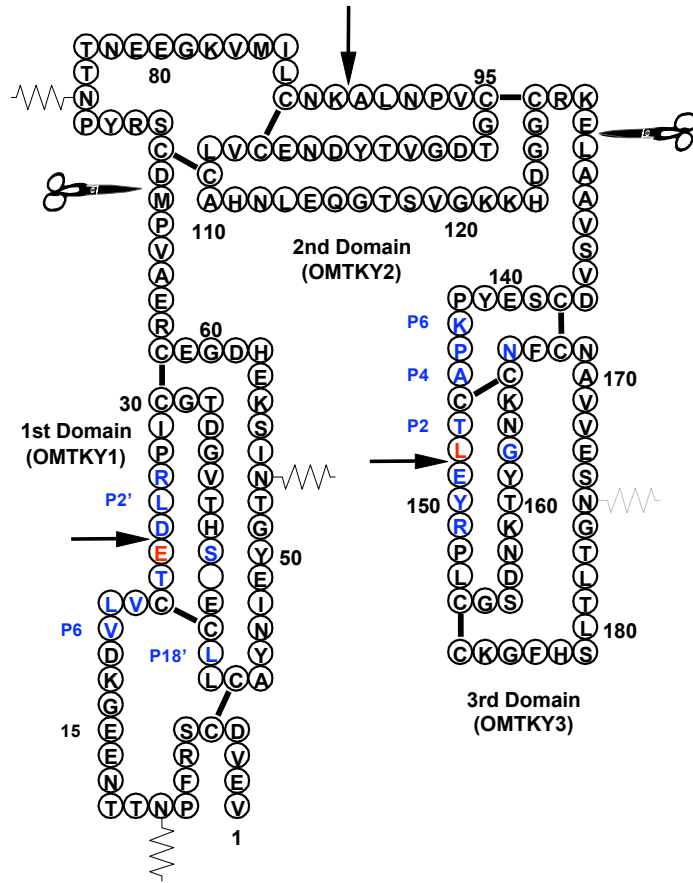


Figure 2

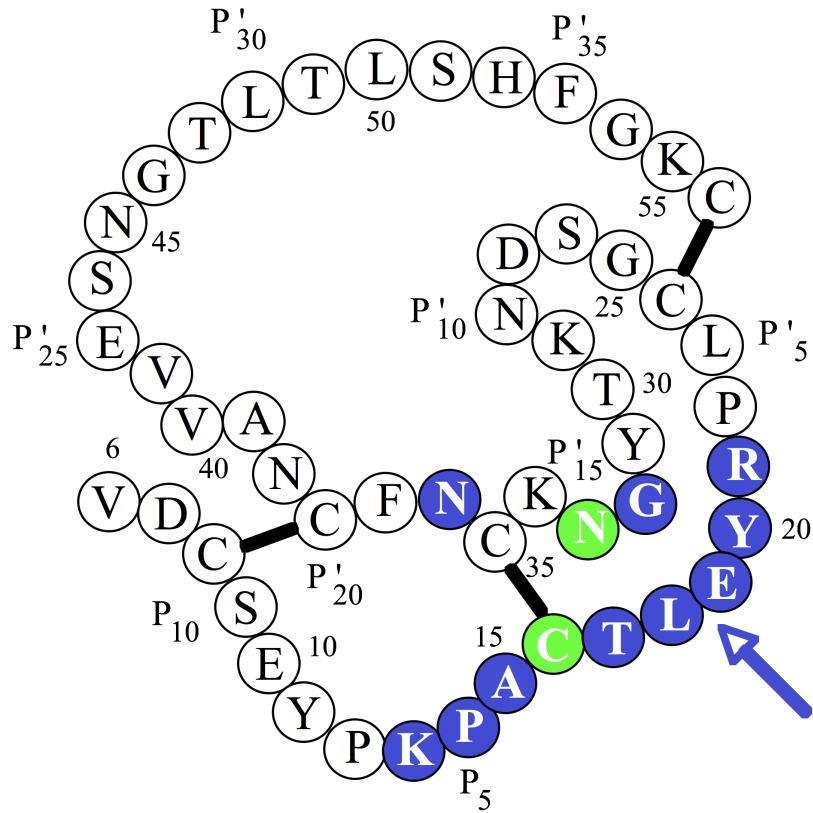


Figure 3

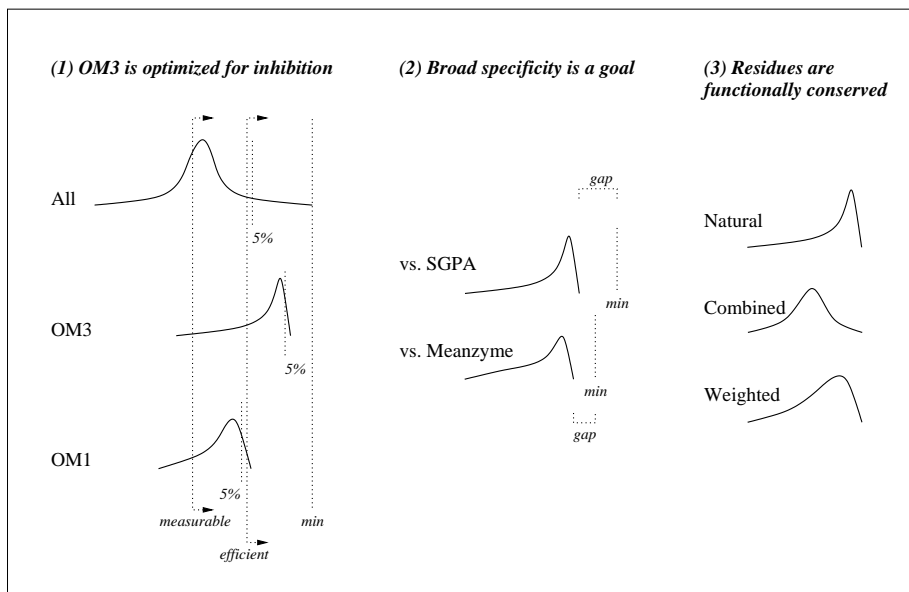


Figure 4

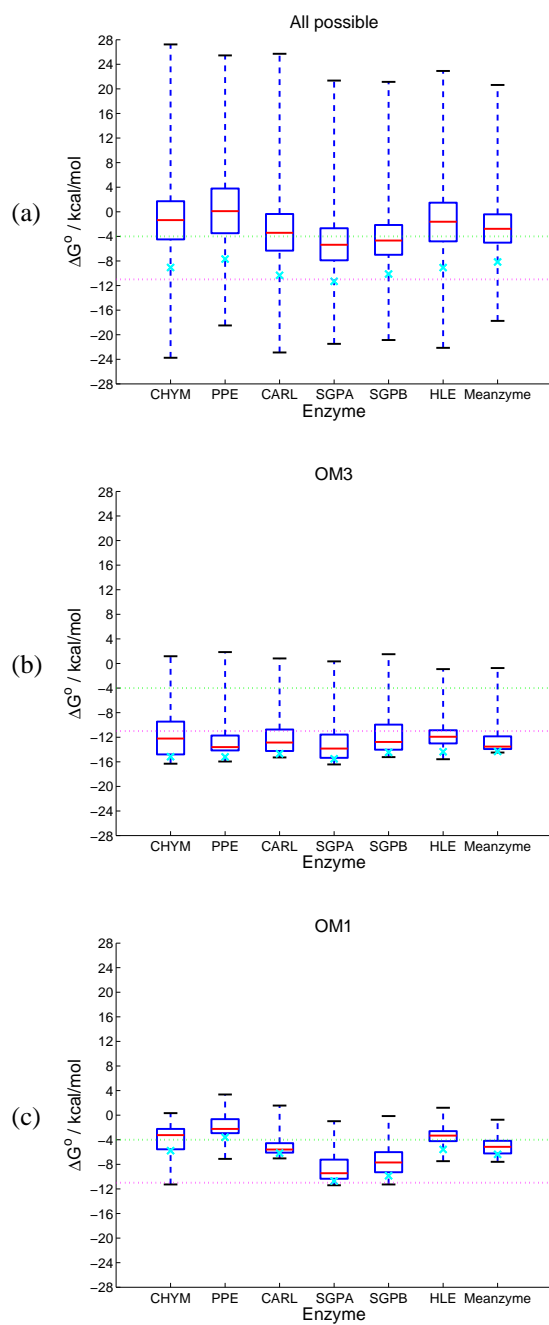


Figure 5

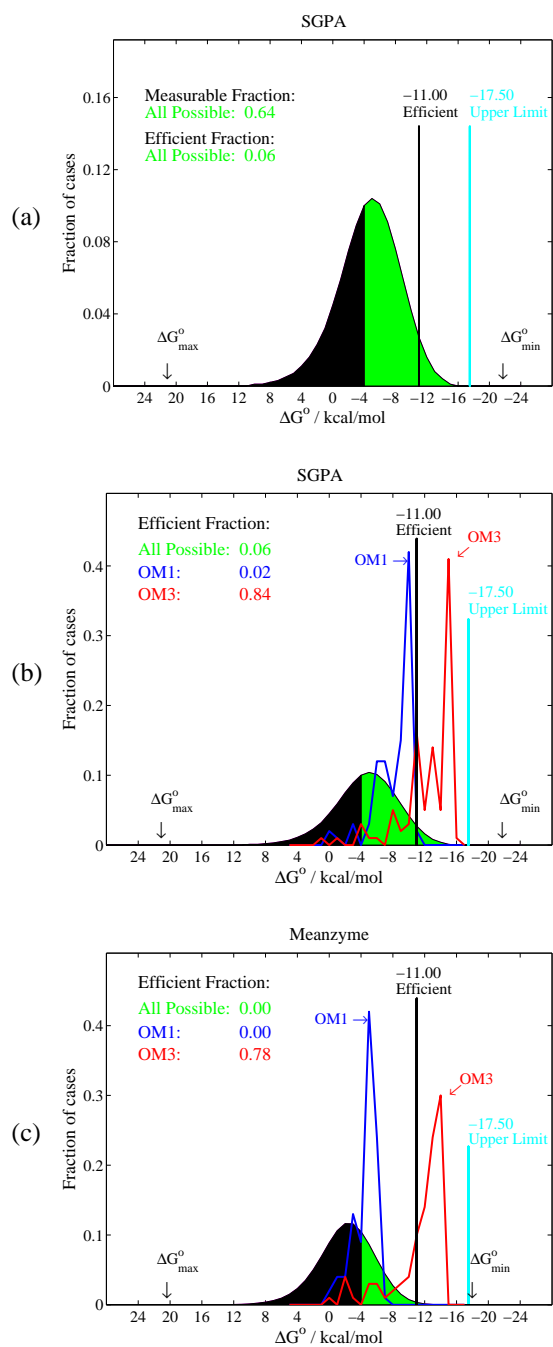


Figure 6

