



# Triosephosphate Isomerase by Consensus Design: Dramatic Differences in Physical Properties and Activity of Related Variants

Brandon J. Sullivan<sup>1</sup>, Venuka Durani<sup>2</sup> and Thomas J. Magliery<sup>2,3\*</sup>

<sup>1</sup>Ohio State Biochemistry Program, The Ohio State University, Columbus, OH 43210, USA

<sup>2</sup>Department of Chemistry, The Ohio State University, Columbus, OH 43210, USA

<sup>3</sup>Department of Biochemistry, The Ohio State University, Columbus, OH 43210, USA

Received 18 May 2011;  
received in revised form  
23 July 2011;  
accepted 1 August 2011  
Available online  
4 August 2011

Edited by C. R. Matthews

## Keywords:

protein engineering;  
consensus design;  
sequence correlation;  
oligomeric state;  
molten globule

Consensus design, the selection of mutations based on the most common amino acid in each position of a multiple sequence alignment, has proven to be an efficient way to engineer stabilized mutants and even to design entire proteins. However, its application has been limited to small motifs or small families of highly related proteins. Also, we have little idea of how information that specifies a protein's properties is distributed between positional effects (consensus) and interactions between positions (correlated occurrences of amino acids). Here, we designed several consensus variants of triosephosphate isomerase (TIM), a large, diverse family of complex enzymes. The first variant was only weakly active, had molten globular characteristics, and was monomeric at 25 °C despite being based on nearly all dimeric enzymes. A closely related variant from curation of the sequence database resulted in a native-like dimeric TIM with near-diffusion-controlled kinetics. Both enzymes vary substantially (30–40%) from any natural TIM, but they differ from each other in only a relatively small number of unconserved positions. We demonstrate that consensus design is sufficient to engineer a sophisticated protein that requires precise substrate positioning and coordinated loop motion. The difference in oligomeric states and native-like properties for the two consensus variants is not a result of defects in the dimerization interface but rather disparate global properties of the proteins. These results have important implications for the role of correlated amino acids, the ability of TIM to function as a monomer, and the ability of molten globular proteins to carry out complex reactions.

© 2011 Elsevier Ltd. All rights reserved.

\*Corresponding author. Departments of Chemistry and Biochemistry, The Ohio State University, 100 West 18th Avenue, Columbus, OH 43210, USA. E-mail address: [magliery@chemistry.ohio-state.edu](mailto:magliery@chemistry.ohio-state.edu).

Abbreviations used: ANS, 1-anilinonaphthalene-8-sulfonic acid; cTIM, consensus TIM; ir-cTIM, interface reversion consensus TIM; ccTIM, curated consensus TIM; DHAP, dihydroxyacetone phosphate; GAP, glyceraldehyde-3-phosphate; MSA, multiple sequence alignment; *S.c.* TIM, *Saccharomyces cerevisiae* TIM; TIM, triosephosphate isomerase; TPR, tetratricopeptide repeat; TEV, tobacco etch virus; AUC, analytical ultracentrifugation; TCEP, tris(2-carboxyethyl) phosphine.

## Introduction

The sequence of amino acids in a protein encodes its physical and functional properties, but our ability to read that code is still very limited.<sup>1</sup> For example, there have been great successes in computational prediction and design of proteins in recent years,<sup>2,3</sup> but we are still far from a comprehensive, accurate model of the thermodynamic consequences of mutations.<sup>4,5</sup> In part, this is because natural proteins are typically only stabilized by 5–15 kcal mol<sup>-1</sup> over the unfolded state, and our knowledge of how to model the unfolded state is poor.<sup>6,7</sup> Remarkable functional designs of enzymes have also been achieved recently, but it remains exceedingly difficult to achieve catalytic efficiencies that compare to natural enzymes.<sup>8–10</sup> The effects of solvation, backbone motion, dynamics, and entropy are largely beyond our ability to predict or design.

One method of designing nonnatural sequences with native-like structures and functions is to look to statistical analysis of families of natural proteins. Genomic sequencing has given us vast databases of sequences of proteins that all have approximately the same structure and activity. This is basically a post-genomic formulation of the so-called “inverse folding problem”: what are all sequences in nature that adopt a particular fold?<sup>11</sup> In the limit, the conservation and variation of sequence features in a multiple sequence alignment (MSA) must contain all of the information necessary to design stable, active sequences. The question is: how do we read and apply that information? We were particularly interested in determining what information is encoded at the positional level (consensus/conservation) *versus* what is encoded by coupling between sites (correlation).

The idea of designing proteins, domains, or motifs from consensus is attractive because it makes intuitive sense that the most common amino acid in each position of an MSA is there for a reason (structural, functional, dynamic, etc.). Consensus sequences of motifs such as the tetratricopeptide repeat (TPR) and ankyrin repeat have been shown to be folded.<sup>12–14</sup> Enzymes, such as the fungal phytases, have been engineered using sequence consensus and have been shown to be active and stable. These consensus phytases were generated from 13 to 21 highly homologous sequences from near-neighbors in phylogeny.<sup>15–17</sup> Consensus-designed proteins generally have had higher thermal stabilities than the average proteins from which the consensus sequence was derived; however, some rational design considerations were applied to unconserved sites in many of these studies. Data from the phytases, antibodies, and thioredoxin suggest that about half the time, mutation of an amino acid to the most common amino acid in the MSA for that position is stabilizing.<sup>18–22</sup>

On the other hand, the most common amino acid in an unconserved site presumably has little informational value, and furthermore, unconserved sites may still be correlated to each other, which is lost in the consensus. For example, the consensus sequence of TPR motifs has a canonical charge of  $-7$  although individual TPRs have a  $0 \pm 2.5$  net charge, because the charged residues are largely poorly conserved surface residues that exhibit charge neutralization only when correlation is considered.<sup>23</sup> The distribution of information between consensus and correlation is not known, although design of WW domains using only consensus *versus* consensus plus correlation yielded a much larger fraction of folded proteins with incorporation of the correlation data.<sup>24,25</sup> When triosephosphate isomerase (TIM) was extensively mutated, virtually all structural positions could individually be mutated conservatively (e.g., Gln to Asn) with little effect on activity, but when all positions were simultaneously varied between the natural residue and a conservative replacement, only about 1 in  $10^{10}$  was active.<sup>26</sup> Therefore, interactions among sites appear to account for a great deal of the information in specifying a folded, active protein, but no experiments to date have elucidated the exact effects of these correlated mutations.

To start to answer this question, we proposed to engineer the pure consensus sequence of a complex protein architecture from a large, diverse enzyme family. Presumably, this pure consensus sequence would scramble or ablate many of the sequence correlations at poorly conserved sites and, as such, could act as “host” for interrogating the effects of “guest” correlation mutations. We selected the TIMs for this study, because they are a very well studied archetypal member of the  $(\beta/\alpha)_8$  proteins that make up 10% of all biological catalysts.<sup>27–29</sup> Because of their glycolytic function in the isomerization of dihydroxyacetone phosphate (DHAP) and glyceraldehyde-3-phosphate (GAP), virtually every organism has a TIM and therefore hundreds of sequences are available. TIM catalyzes a sophisticated reaction with nearly diffusion-limited kinetics and with coordinated motion in the catalytic cycle.<sup>30–35</sup> Furthermore, TIM barrel proteins have generally been difficult to engineer despite their ubiquity in nature.<sup>36</sup>

Here, we report the construction and characterization of closely related TIM proteins based purely on consensus, one from a “raw” sequence database and one from a later database curated of fragments and repeats. The raw consensus TIM (cTIM) is weakly active, poorly folded, and monomeric, in contrast to nearly all known natural TIMs, which are dimers. The curated consensus TIM (ccTIM) is dimeric, well folded, and fully active. We demonstrate that the oligomeric states are not a result of

**Table 1.** Apparent Michaelis–Menten parameters

	GAP			DHAP		
	$K_m$ (mM)	$k_{cat}$ ( $\text{min}^{-1}$ )	$k_{cat}/K_m$ ( $\text{M}^{-1} \text{min}^{-1}$ )	$K_m$ (mM)	$k_{cat}$ ( $\text{min}^{-1}$ )	$k_{cat}/K_m$ ( $\text{M}^{-1} \text{min}^{-1}$ )
<i>S.c.</i> TIM <sup>a</sup>	1.5	$5.22 \times 10^5$	$3.48 \times 10^8$	2.3	$4.5 \times 10^4$	$1.96 \times 10^7$
monoTIM <sup>b</sup>	4.1	$3.1 \times 10^2$	$7.56 \times 10^4$	12.2	12	$9.83 \times 10^2$
<i>S.c.</i> TIM <sup>c</sup>	$0.8 \pm 0.1$	$2.2 \pm 0.1 \times 10^5$	$3.0 \pm 0.5 \times 10^8$	$4 \pm 1$	$2.9 \pm 0.4 \times 10^4$	$8 \pm 2 \times 10^6$
Corr. <sup>d</sup>				$2 \pm 1$		$1.5 \pm 0.7 \times 10^7$
cTIM <sup>c,e</sup>	$1.0 \pm 0.1$	$8.5 \pm 0.4$	$8 \pm 1 \times 10^3$	$7 \pm 1$	$1.0 \pm 0.1$	$1.6 \pm 0.3 \times 10^2$
ir-cTIM <sup>c</sup>	$2.1 \pm 0.3$	$2.6 \pm 0.2$	$1.2 \pm 0.2 \times 10^3$	— <sup>f</sup>	— <sup>f</sup>	— <sup>f</sup>
ccTIM <sup>c</sup>	$0.34 \pm 0.04$	$7.1 \pm 0.2 \times 10^4$	$2.1 \pm 0.3 \times 10^8$	$5.3 \pm 0.4$	$1.48 \pm 0.05 \times 10^4$	$2.8 \pm 0.2 \times 10^6$
Corr. <sup>d</sup>				$2.4 \pm 0.6$		$6 \pm 2 \times 10^6$

<sup>a</sup> From Nickbarg and Knowles.<sup>28</sup>

<sup>b</sup> From Schliebs *et al.*<sup>56</sup>

<sup>c</sup> The parameters for DHAP turnover are apparent, as they have not been adjusted for arsenate inhibition.

<sup>d</sup> Corrected kinetic parameters for *S.c.* TIM calculated with a literature<sup>28</sup> arsenate  $K_i$  value of 9.6 mM and for ccTIM calculated with an estimated arsenate  $K_i$  of  $5 \pm 2$  mM.

<sup>e</sup> For the GAP reaction catalyzed by cTIM, apparent values that do not account for high GAP concentration (>4 mM) points with lower velocity are shown, suggesting some type of substrate inhibition (Supplemental Fig. 3).

<sup>f</sup> The turnover of DHAP by ir-cTIM was not consistently observable above the background reaction.

defects at the interface but rather that global properties of the proteins differ dramatically. Those properties arise from sequence variations at unconserved sites, where correlated occurrences of amino acids may play a significant role.

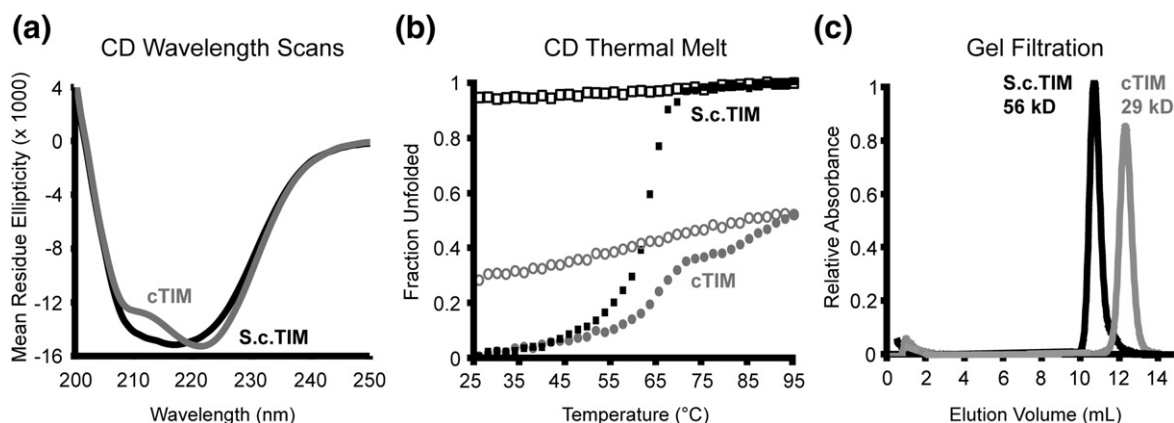
## Results

### Consensus TIM

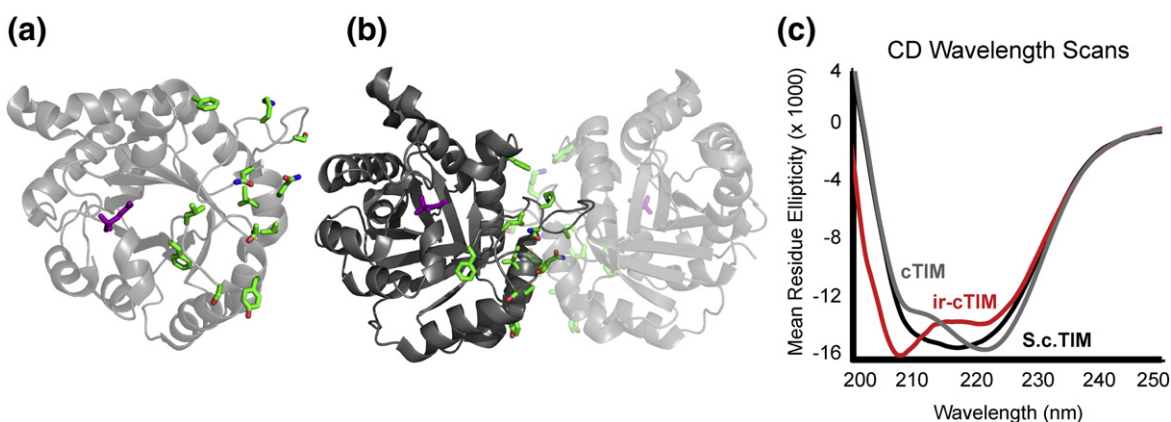
The consensus sequence of all TIMs was determined from the most common amino acid in each position of the Pfam alignment (version 18.04) of 639 sequences. Because hidden Markov model alignment is not well suited to deal with insertions relative to the seed alignment, the total number of

positions in the alignment (373) is much larger than the average length of a TIM sequence (235 aligned positions). Consequently, only positions with greater than 45% occupancy were selected, resulting in a sequence of 248 aa including four unaligned N- and C-terminal residues from *Saccharomyces cerevisiae* TIM (*S.c.* TIM). (*S.c.* TIM is also 248 aa long.) Because of the great evolutionary diversity of this ancient enzyme family, the consensus amino acid sequence is only 70% identical with that of *Tenebrio molitor* TIM, its closest known homolog.

The gene for the cTIM was assembled from synthetic oligonucleotides using a PCR scheme similar to the reassembly step in DNA shuffling.<sup>37</sup> The gene was cloned into two expression vectors, one under the control of the *tac* promoter and one under the control of the T7 promoter. The *tac* construct was transformed into DF502, an *Escherichia*



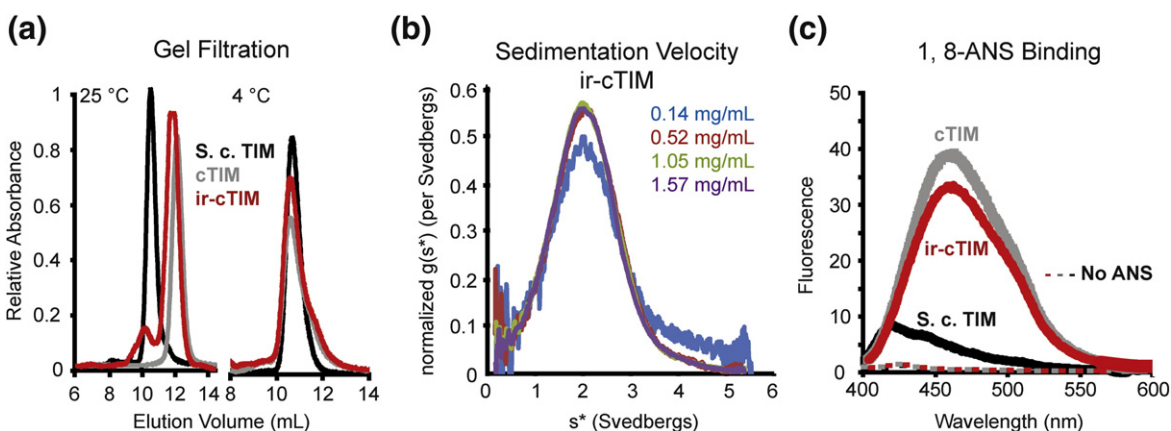
**Fig. 1.** cTIM structure and stability. (a) CD wavelength spectrum of cTIM and *S.c.* TIM. (b) Thermal melt and cooling of cTIM and *S.c.* TIM from the 222-nm CD data. Data collected at increasing temperatures are shown as closed points while data points collected during the reverse melt are shown open. (c) Gel-filtration chromatography shows that *S.c.* TIM elutes as a dimer, but cTIM elutes later with calculated molecular mass corresponding to monomeric TIM.



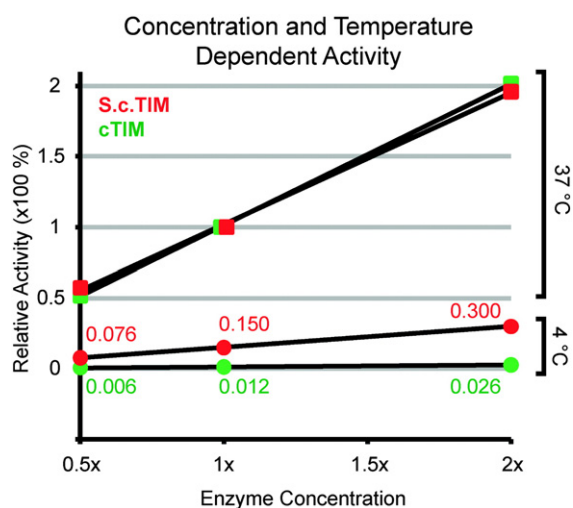
**Fig. 2.** *ir-cTIM* design. (a) The crystal structure of *S.c.* TIM (2YPI) is shown as an open monomer. The active-site bound inhibitor 2PG is shown in purple. The 12 mutations between *cTIM* and *ir-cTIM* are shown as sticks. These residues are all within 5 Å of the second chain, which reaches nearly into the active site. (b) The same rendering as in (a) but with the full dimer shown. (c) The CD wavelength spectrum of *ir-cTIM* shows similar ellipticity at 222 nm but significantly more signal at 205 nm.

*coli* strain deficient in TIM and several other genes nearby in the chromosome.<sup>38</sup> Growth on lactate and glycerol minimal media was comparable to complementation with *S.c.* TIM using the same construct. However, DF502 growth was inconsistent in our hands, perhaps because of the very slow growth on minimal media due to the large number of metabolic genes knocked out in this strain. We turned to the recent Keio collection single-gene knockout of TIM,<sup>39</sup> which we lysogenized with DE3 phage to support transcription from the T7 promoter. At 5 μM IPTG, *cTIM* supported growth on lactate minimal media in 2–3 days and on glycerol minimal media in 4 days, while *S.c.* TIM resulted in growth in about 1 day on both media.

The *cTIM* protein could be overexpressed at very high levels in *E. coli* and was purified to near homogeneity using two-step IMAC purification with 6×His tag cleavage by tobacco etch virus (TEV) protease. To eliminate contamination by the endogenous *E. coli* TIM, the engineered TIMs were purified from the Keio TIM knockout DE3 strain. The Michaelis–Menten parameters were determined from steady-state kinetics for both directions of the isomerization reaction. The apparent  $K_m$  values for DHAP and GAP are comparable to those for *S.c.* TIM, but the apparent  $k_{cat}$  values are reduced by about  $10^4$ -fold (Table 1). Wild-type TIMs exhibit bimolecular kinetics close to the diffusion limit, but apparently weak growth can be supported with



**Fig. 3.** *ir-cTIM* characterization. (a) The elution volume from gel-filtration chromatography of *ir-cTIM* corresponds to a molecular mass close to monomer. At lower temperatures (4 °C), *cTIM* and *ir-cTIM* elute as dimers with a shoulder for monomeric species. (b) Sedimentation velocity shows that *ir-cTIM* is monomeric with no concentration-dependent oligomerization. (c) ANS binding of *S.c.* TIM exhibits a weak fluorescence peak at 420 nm. *cTIM* and *ir-cTIM* exhibit strong fluorescence with a red-shifted maxima of 460 nm, suggesting that they are both molten globular.



**Fig. 4.** The activity of cTIM and *S.c.* TIM studied under a series of temperatures and concentrations. The activity at 37 °C and 1x[E] was arbitrarily set at unity (100%) for both enzymes. The activity doubled and halved for the wild-type enzyme when the concentrations were increased and decreased twofold, respectively. This occurred at both temperatures. If cTIM were active as the dimer, one would expect doubling the concentration of enzyme to have a nonlinear effect on activity. Lowering the temperature from 37 to 4 °C led to a 13-fold reduction in reaction rate for the wild-type enzyme at each concentration. At 4 °C, we observe cTIM dimers by gel filtration. All other things being equal, if cTIM dimers were the active unit, we would expect less than a 13-fold decrease for cTIM. In fact, we see the opposite; the average activity decreases 80-fold between 37 and 4 °C at each enzyme concentration.

significant reductions in activity. Therefore, an active TIM was derived from consensus alone, albeit one with significantly reduced activity.

Far-UV circular dichroism (CD) spectra for cTIM and *S.c.* TIM are similar and consistent with similar  $(\beta/\alpha)_8$  architecture (Fig. 1a). Thermal denaturation was followed by CD spectroscopy at 222 nm (Fig. 1b). *S.c.* TIM unfolds in a single, irreversible step at about 60 °C. cTIM exhibits a similar pretransition baseline to *S.c.* TIM but does not unfold in a single step and is only ~50% unfolded at 95 °C. Unlike *S.c.* TIM, which precipitates at 95 °C, cTIM shows no signs of precipitation and exhibits some reversibility on cooling from 95 °C. This behavior is consistent with the thermal stabilization that has been observed for consensus mutations, although it is possible that more molten globule character is also exhibited by cTIM.<sup>12–17,40</sup>

With the exception of a few tetrameric TIMs from thermophiles, all known TIMs are homodimeric. The structure of TIM suggests that dimerization is necessary for full assembly of the active site by the interdigitation of loop 3 from the opposite monomer, and engineered monomeric TIMs exhibit

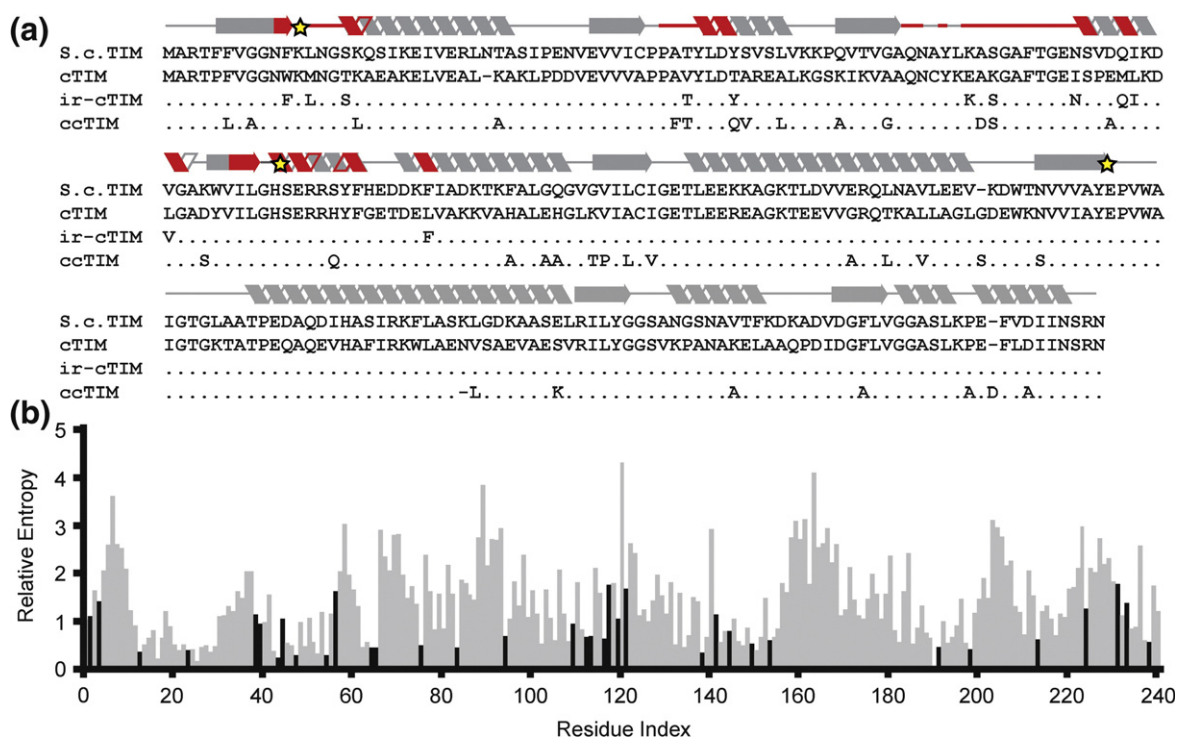
$k_{cat}/K_m$  values reduced by about  $10^4$ -fold.<sup>41–45</sup> The quaternary structure of cTIM was determined by gel-filtration chromatography (Fig. 1c). cTIM elutes significantly after *S.c.* TIM. Elution volumes were compared to a standard curve to determine apparent molecular masses; *S.c.* TIM eluted as the expected dimer (~56 kDa), but the consensus enzyme elutes as a monomer at room temperature with an apparent molecular mass of ~29 kDa. Surprisingly, the consensus sequence of over 600 dimeric proteins is a monomer.

### Engineering the interface of cTIM

Although the monomeric state of cTIM was a surprise, its activity is consistent with TIM variants intentionally engineered to be monomers.<sup>41–45</sup> These attempts to monomerize TIM involved deletions in the interfacial loop 3 and mutations that reversed charge pairing. We hypothesized that by choosing the most common amino acid at each position of cTIM, we had scrambled necessary amino acid interactions (i.e., correlations) at the dimer interface. To examine this hypothesis, we reverted the dimerization interface to the sequence observed in *S.c.* TIM, which is known to be dimeric. The 1YPI crystal structure reveals 40 residues within 5 Å of the opposite monomer. The 12 interface residues that differed between cTIM and *S.c.* TIM were mutated in cTIM to create an interface reversion cTIM (ir-cTIM; Fig. 2a and b).

The ir-cTIM was purified in similar yield to the original cTIM. CD spectra are similar, but ir-cTIM exhibits greater signal at 205 nm, suggesting more random coil (Fig. 2c). The thermal melts monitored at 222 nm were essentially identical. By gel-filtration chromatography, ir-cTIM elutes at a calculated molecular mass slightly larger than that of cTIM at room temperature (~42 kDa, Fig. 3a). Sedimentation velocity by analytical ultracentrifugation (AUC) confirmed that the protein is still monomeric at room temperature (Fig. 3b). Furthermore, ir-cTIM did not exhibit concentration-dependent oligomerization over a 10-fold range of concentrations (0.15–1.5 mg mL<sup>-1</sup>). The activity of ir-cTIM was decreased compared to cTIM and failed to complement the Keio TIM knockout on minimal media.

When the gel-filtration chromatography was repeated at 4 °C (Fig. 3a), all three of the proteins (*S.c.* TIM, cTIM, and ir-cTIM) eluted as dimers. For cTIM, a shoulder on the dimer-weight peak suggests that both monomer and dimer are populated at 4 °C and 37 μM (1 mg mL<sup>-1</sup>), suggesting that this concentration is close to the  $K_d$  at this temperature. These results together suggest that the monomeric states of cTIM and ir-cTIM at room temperature may not be the result of inherent defects in the dimerization interface but rather nonnative global properties of the cTIM scaffold.



**Fig. 5.** Comparison of consensus TIM sequences. (a) Sequence alignment of *S.c.* TIM and consensus TIMs. Secondary structure shown for *S.c.* TIM with interface residues (within 5 Å of chain b) shown in red and active-site residues marked as stars. Periods denote the same amino acid as cTIM. (b) Plot showing the relative entropy (i.e., conservation) of each position in the TIM alignment. Residues that are mutated between cTIM and ccTIM are shown in black, while all other positions are shown in gray.

We also analyzed the binding of the three proteins to the hydrophobic dye 1-anilinonaphthalene-8-sulfonic acid (ANS). ANS is quenched in aqueous buffer but fluoresces strongly in lower dielectric environments such as organic solvent or when bound in the core of a protein. ANS binding is taken to be a sign of fluid tertiary structure exhibited by molten globules.<sup>46,47</sup> *S.c.* TIM shows a weak fluorescence emission peak at 418 nm, but both cTIM and ir-cTIM have strong red-shifted fluorescence with peaks at 460 nm (Fig. 3c). The 600-MHz <sup>1</sup>H, <sup>15</sup>N-heteronuclear single quantum coherence NMR spectrum of cTIM, however, displays a fair amount of amide peak dispersion for a protein of this size (Supporting Information, Fig. 7). Taken together, the biophysical data suggest that cTIM is monomeric and not as well folded as native TIMs at room temperature and above.

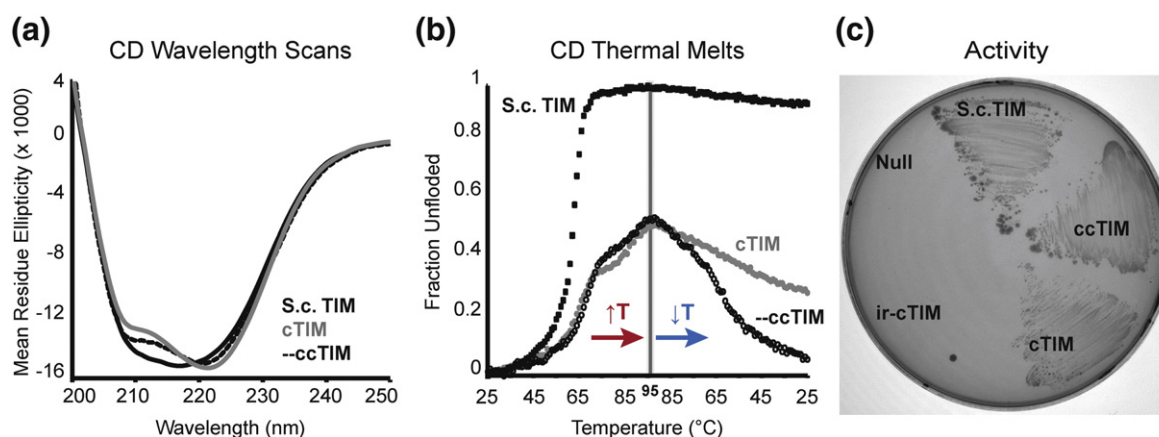
### Concentration and temperature studies

One could imagine that the weak activity of cTIM is due to weak activity in the monomer or to a small population of dimer. To examine further the weak activity of cTIM, we observed single-point kinetics over a range of enzyme concentrations at 4 and 37 °C (Fig. 4). *S.c.* TIM, which is dimeric at both temperatures across the whole range of concentra-

tions (16, 32, and 64 pM), increased in activity linearly with respect to concentration at both temperatures. Furthermore, there was a 13-fold decrease in activity at each concentration when the reaction was performed at 4 °C versus 37 °C. When cTIM was assayed under the same conditions (at 60–240 μM enzyme), we still observed a linear increase in activity with respect to concentration at both temperatures, but the activity was 80-fold lower at the lower temperature for all three concentrations. If activity required dimerization, we would have expected a nonlinear increase in activity at increasing concentration, as more of the dimeric state is populated, and we would have expected a smaller decrease in activity between 37 and 4 °C at all concentrations, since cTIM goes from mostly monomeric to mostly dimeric under these conditions. The composite data suggest that cTIM is active as a monomer with molten globular properties. It is worth noting that the dimeric species seen at 4 °C in cTIM and ir-cTIM may not be native-like dimers.

### Database curation

A third consensus TIM variant that we engineered shed light on the properties of the original cTIM. When we began the analysis for correlated occurrences of amino acids, we downloaded the then-



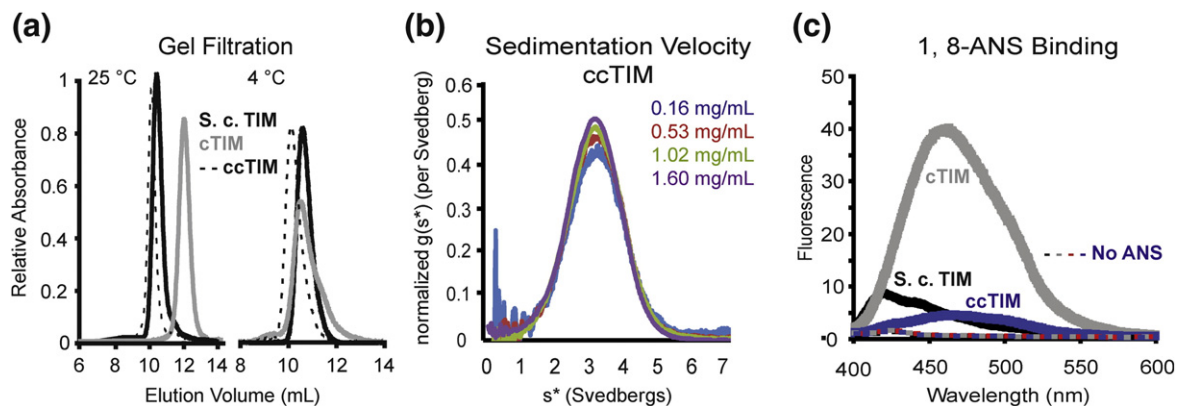
**Fig. 6.** ccTIM characterization. (a) CD wavelength spectra of consensus TIMs. (b) The consensus variants share similar unfolding patterns when ellipticity is monitored at 222 nm with temperatures ramping from 25 to 95 °C. If the melted samples are cooled back to room temperature, cTIM and ccTIM refold significantly as judged by an increase in ellipticity. ccTIM regains 95% of its initial ellipticity. (c) *In vivo* characterization of TIMs on lactate minimal media in the absence of IPTG. After 3 days of leaky expression at 37 °C, all but ir-cTIM complement the Keio(DE3) TIM knockout.

current version (22.0) of the Pfam database and curated it to remove repeated sequences and sequence fragments that did not represent full genes. More precisely, sequences with fewer than 205 aa (351 sequences) and exact sequence repeats (107 sequences) were removed from the 1239 sequence database to yield 781 nonredundant full-length sequences. A new ccTIM was created using a similar approach to occupancy as described for cTIM, resulting in a 248-aa sequence with 36 sequence differences from cTIM (34 substitutions, 1 insertion, and 1 deletion, Fig. 5). There was a single position in the alignment (which aligned with S.c. TIM residue 49) that was equally occupied by two residues: alanine and glutamine. The position was arbitrarily chosen to be Gln. The differences between cTIM and ccTIM arise from unconserved positions in which the most common amino acid differs, and consequently, we expected these changes to have

little impact. The amino acid bias of a position can be quantified by calculating the relative entropy between positional distribution and the distribution of amino acids in a neutral reference state, such as amino acid usage in all open reading frames in yeast. From this calculation, it is evident that only unbiased or weakly biased positions were affected (Fig. 5b). These positions tolerate virtually any amino acid in all TIMs, and therefore, only minor differences were anticipated between cTIM and ccTIM.

### Curated consensus TIM

ccTIM expresses well in bacteria with yields approaching 50 mg L<sup>-1</sup>. CD wavelength spectra and thermal melt traces were essentially the same as those of cTIM (Fig. 6a and b). The ellipticities for the 222-nm minima corresponding to  $\alpha$ -helical structure are all within 7% when normalized for protein



**Fig. 7.** Structure of ccTIM. (a) ccTIM elutes near the calculated volume corresponding to dimer by gel-filtration chromatography. (b) Sedimentation velocity confirms that ccTIM is dimeric with no concentration dependence between 0.16 and 1.6 mg mL<sup>-1</sup>. (c) ANS binding of ccTIM yields a very weak fluorescence at 460 nm.

concentration, which was confirmed by SDS-PAGE and amino acid analysis. However, other biophysical properties turned out to be starkly different. When the thermal melt is reversed from 95 to 25 °C, ccTIM refolds almost quantitatively. There is a red shift in emission upon ANS binding, but the very low level of fluorescence suggests that ccTIM is much less molten than cTIM (Fig. 7c). The protein elutes from a gel-filtration column at room temperature with an apparent molecular mass of 66 kDa, slightly more than that of *S.c.* TIM or the calculated dimeric mass (Fig. 7a). AUC sedimentation velocity studies confirm that the protein is dimeric (50.5 kDa with 95% confidence) with less than 2% forming higher aggregates (Fig. 7b).

ccTIM is nearly as active as wild-type TIMs, with comparable DHAP and GAP  $K_m$  values and  $k_{cat}$  values of  $10^4$ – $10^5$   $\text{min}^{-1}$ . ccTIM complements growth in the Keio TIM knockout, leading to growth on minimal media similar to that of *S.c.* TIM and faster than that of cTIM (Fig. 6a). Surprisingly, although cTIM and ccTIM differ only in a relatively small number of unconserved positions and have similar structural and thermodynamic properties, cTIM is a molten globular monomer with weak activity and ccTIM is a native-like structured dimer with wild-type activity.

### Details of kinetic characterization

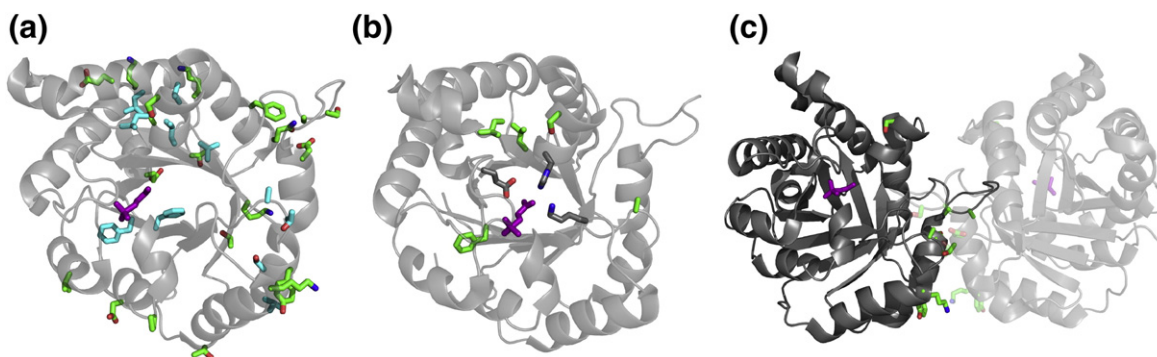
The catalyst of an isomerization reaction may not affect the thermodynamic equilibrium of its substrates. The Haldane relationship,  $(k_{cat}/K_m \text{ for GAP})/(k_{cat}/K_m \text{ for DHAP})$ , for TIM has been reported to be about 22.<sup>48</sup> The consensus-designed variants reported in Table 1 apparently have Haldane ratios of 50 for cTIM and 75 for ccTIM,

representing 2- to 3-fold combined error in the  $k_{cat}/K_m$  values. The majority of this error is manifested in the inflation of DHAP  $K_m$  values by competitive arsenate inhibition.<sup>49</sup> For cTIM, this is further complicated by the accurate determination of  $k_{cat}$  and  $K_m$  due to some type of substrate inhibition at high concentrations of DHAP (Supplemental Fig. 3). In the case of ccTIM, we estimated the  $K_i$  of arsenate by analyzing the DHAP reaction in the presence and absence of arsenate (Supplemental Fig. 6). The  $K_i$ ,  $5 \pm 2$  mM, yields an adjusted DHAP  $K_m$  of  $\sim 2.4$  mM, which translates to a Haldane relationship of  $35 \pm 13$ . This is within the range of previously reported data.

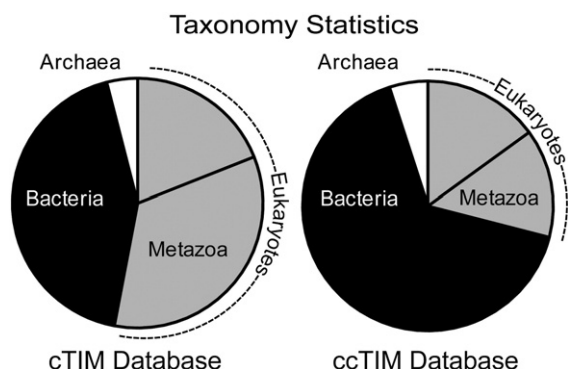
### A comparison between consensus TIMs

While cTIM is 70% identical with *T. molitor* TIM, ccTIM is only 61% identical with its nearest natural sequence neighbor, *Roseiflexus* sp. TIM, but cTIM and ccTIM are 85% identical with one another. Only one residue mutated between cTIM and ccTIM is within 5 Å of the active-site residues (K12, H95, and E165), the active-site lid (residues 166–176), or the 2PG inhibitor bound in crystal structure 2YPI (Fig. 8b). The only proximal mutated position (I127V) is close by virtue of a backbone–backbone interaction with E165. The mutations are spread throughout the protein secondary structures (17 in helices, 10 in sheets, and 8 in loops), and they are mainly solvent exposed (21 are more than 10% exposed in the dimer with an average exposure of 21%, Fig. 8a).<sup>50</sup> Except for F224A, the eight nonconservative mutations were on the protein surface.<sup>51</sup> Stated simply, there is no obvious reason for the dramatic differences between the properties of cTIM and ccTIM.

The 36 differences between the consensus TIMs are at largely unconserved positions (Figs. 5b and 9c).



**Fig. 8.** Sequence differences between cTIM and ccTIM. (a) The 36 mutations are shown as colored sticks with the active-site-bound inhibitor colored purple. Buried residues are cyan and surface-exposed residues are green. (b) The active site is shown by highlighting the catalytic residues (K12, H95, and E165) as gray sticks. The inhibitor 2PG is shown in purple. Only six residues have any atoms within 8 Å of the inhibitor, active-site residues, or active-site lid, loop 6, and none appear to be intimately involved with the active-site residues. (c) Eight mutations occur within 5 Å of the other monomer. The majority of these residues are surface exposed. A44F and V45T (A44 and T45 in *S.c.* TIM) are the only two positions that change solvent exposure between the monomer and the dimer. All figures were rendered in PyMOL using the 2PG-inhibitor-bound crystal structure, 2YPI.

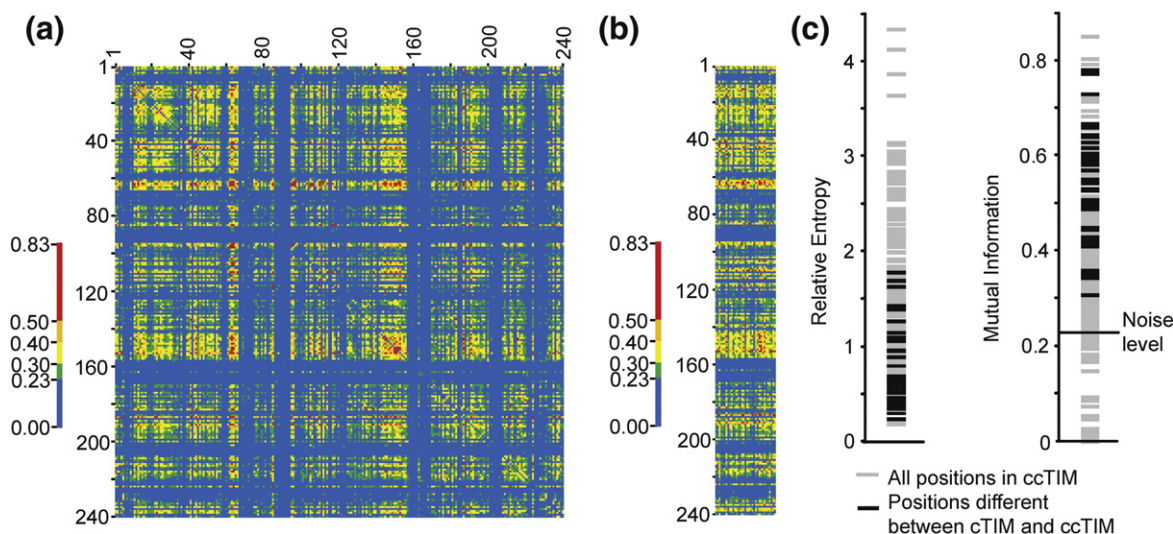


**Fig. 9.** Taxonomic distribution of sequences used to generate the cTIM and ccTIM sequences. The cTIM database had a much higher fraction of eukaryotic and metazoan sequences than the ccTIM database, which was dominated by bacterial sequences.

The average relative entropy compared to the neutral reference state is 1.42 for all positions *versus* 0.82 for the 36 varying positions. Most of the 12 mutated positions with relative entropies greater than 1.00 arise from distributions with a significant number of sequences occupied by 2 or 3 aa. For example, position 238 has a relative entropy of 1.38. The initial distribution was 169 Pro and 137 Ala out of 407 sequences occupied at this site. The curated distribution changed to 221 Pro and 325 Ala out of 720, switching the most common and next most common residues. A large fraction of the positions (11) were mutated to Ala. The mutations result in a

significant decrease in the charge of the protein ( $-11$  in the 240 aligned positions *versus*  $-5.5$  for cTIM,  $-3.5$  for *S.c.* TIM, and  $-5 \pm 5$  for TIMs overall). This phenomenon was seen before with the consensus sequence of the TPR motif, where it was shown to arise from scrambling of correlated surface charges.<sup>23</sup> We speculate that one major difference between cTIM and ccTIM may be in the extent of correlated occurrences of amino acids that are scrambled or broken.

To test this hypothesis, we analyzed statistical correlations between all positions in the MSA. These correlations, calculated here as the mutual information between the amino acid distributions at each pair of positions in the MSA, reveal that most sites in TIM do not exhibit strong correlations (Fig. 10a). However, the weakly conserved positions that change between cTIM and ccTIM are highly enriched in positions with strong sequence correlations (Fig. 10b and c). Another difference between cTIM and ccTIM is that the sequence databases used to construct them differ substantially in their phylogenetic distributions. Specifically, a greater fraction of sequences leading to ccTIM came from bacteria (Fig. 9). Our preliminary analysis of the correlation data suggests that networks of correlated residues differ in differing branches of phylogeny. A full analysis of TIM sequence correlations will be presented separately. Studies of further mutants of cTIM and ccTIM designed to assess the roles of individual mutations and correlated pairs or networks of mutations are underway.



**Fig. 10.** Sequence correlations in TIMs. (a) Heat map of the mutual information for all 57,600 pairwise positional correlations in the TIM alignment. Cool colors (blue and green) represent weak correlations, and warm colors (orange and red) represent strong correlations. (b) Heat map of the correlations observed for the 35 positions that differ between cTIM and ccTIM. (c) On the left, the relative entropy (sequence bias) of each position is plotted with those corresponding to the mutated positions in black. On the right, the maximum mutual information value for each position is plotted with those corresponding to the mutated positions in black. Most of the mutations between the consensus TIMs occur at nonconserved positions, but these sites are enriched in positions with strong sequence correlations.

## Discussion

One important lesson from this work is that, even for a large family of enzymes with significant evolutionary and sequence diversity that carry out a sophisticated and highly tuned reaction, native-like activity can be engineered from consensus alone. Natural TIMs exhibit nearly diffusion-controlled kinetics, which are believed to arise from a highly orchestrated cycle of loop motion and precise positioning of residues in the active site to stabilize the enediol intermediate and avoid the formation of a toxic methylglyoxal by-product. ccTIM is able to carry out this reaction at wild-type rates despite differing from the nearest natural TIM in 40% of its amino acids and never having been subject itself to evolution. This strongly argues that the vast majority of information for protein structure and function is encoded positionally, at the level of consensus, and not in higher-order correlations. It would be interesting in the future to examine methylglyoxal formation by the TIM variants engineered here.

However, the stark differences between ccTIM, cTIM, and ir-cTIM illustrate that there is more information in the sequence families than just the positional information. These proteins are all, in a sense, “consensus” variants. They differ in sites that are highly tolerant to mutation, and they arise from variations between the most common amino acids at those unconserved positions. There is no obvious reason that the particular set of amino acids at the 36 positions that differ between cTIM and ccTIM results in a weakly active monomer in the former case and a wild type-like dimer in the latter. However, the two proteins appear to differ in sites that are enriched in stronger statistical correlations, and the phylogenetic compositions of the databases leading to them differ substantially. We therefore speculate that differences in the extent to which sequence correlations were preserved or “broken” in the two consensus sequences may play a role in their properties. Experiments to test this idea are underway. It is worth noting that the large number of sequences used to construct these variants makes it possible to meaningfully assess sequence correlations.

While a native-like protein resulted from the curated database and a less-active molten globular protein resulted from the uncurated one, this does not necessarily suggest that curation is the key to successful consensus engineering. The sequence collections that are available significantly under-sample complete evolutionary history and are affected by researcher interest and organism availability. For consensus design, it is difficult to articulate a convincing reason that any one sequence (or even sequence fragment) should be included or omitted from a sequence library, since the process by which the library was created was inherently biased. Sequence fragments are a complication for correla-

tion analysis, but they simply add information to consensus analysis for the sites to which they correspond. Similarly, it is possible to imagine that many duplicates of a small number of unique sequences might bias the consensus sequence, but that was not the case here. Of the 1239 TIM sequences in Pfam 22, over 1000 of the sequences are unique and only 2 full-length sequences were found to be repeated more than 3 times (8 and 10 times).

We also do not think that the difference in size of the final cTIM and ccTIM databases (639 *versus* 781) had any significant effect in itself on the consensus sequence. We randomly removed 142 sequences from the ccTIM database in proportion to the taxonomic distribution in 20 separate trials. On average, the consensus sequence adopted  $3 \pm 2$  aa mutations relative to ccTIM with a range of 0–8. It is therefore possible to produce the same ccTIM sequence without additional sequences. A related factor that we completely neglect here is that sequence alignment quality is likely to have some effect, especially on weakly conserved positions. Weakly conserved stretches and regions with length heterogeneity (such as loops) are the most difficult to align with certitude. Larger numbers of sequences improve alignment quality, and further expansion of sequence databases will likely improve our understanding of weakly conserved positions and correlations among them.

The biophysical differences between cTIM and ccTIM are especially fascinating. Because of the way that the enzymes are designed, all of the conserved residues required for function (e.g., the Glu, His, and Lys in the active site) are present. The consensus enzymes exhibit similar CD spectra to yeast TIM, and even the weak activity of cTIM suggests that the proteins exhibit or at least sample highly similar structures to the natural TIMs. However, the oligomeric states and ANS binding data suggest that the primary difference between cTIM and ccTIM is in their global properties; that is, cTIM is more fluid and only dimerizes significantly at low temperature. It is still unclear how evolutionarily common mutations at 36 unconserved positions result in this difference. Structural and dynamic studies on cTIM and ccTIM are underway to understand better the nature of this change.

While it is difficult to prove beyond a shadow of a doubt, the preponderance of the evidence argues that cTIM is active as a monomer. The most convincing evidence is that cTIM activity increases in direct proportion to concentration (i.e., the specific activity is not concentration dependent, implying that any additional dimerization is not increasing activity) and that cTIM is reduced further in activity than *S.c.* TIM upon cooling to 4 °C, although *S.c.* TIM is a dimer at both concentrations and cTIM is significantly dimeric only at 4 °C. Further purification of cTIM by ion-exchange chromatography did not result in higher activity, and multiple preparations yielded similar

activities, suggesting that the problem is not simply that there is a large inactive population. Careful controls, including purification from a TIM-free strain, ensure that wild-type TIM contamination is not the cause of the activity.

Wierenga *et al.* have engineered several versions of trypanosomal TIM to be monomeric, which turned out to be a surprisingly difficult undertaking.<sup>41–45</sup> Even relatively radical mutations or deletions to the interfacial loop 3 resulted in significant amounts of dimer at higher concentrations. Similarly, Goraj *et al.* attempted to engineer monomeric TIMs from human TIM by interface mutations, but the results were monomer–dimer equilibria, as well as inactive proteins or concentration-dependent specific activities, implying that activity arises from the dimer. We believe that our concentration- and temperature-dependent kinetic studies provide some of the strongest evidence that TIM can function as a monomer. However, it is interesting that the trypanosomal monomeric mutants have similar  $k_{\text{cat}}$  values to cTIM and that the mechanism of monomerization is so different in cTIM/ir-cTIM (i.e., global scaffold changes *versus* interface mutations).

The unusual dynamic nature of cTIM calls to mind the loop motions present in the TIM catalytic cycle.<sup>30–35</sup> Movement of loop 6 occurs on the same time scale as catalysis. As it appears to form a lid on the active site, its motion is thought to be coordinated with catalysis. This loop motion has been observed directly by fluorescence and by solution-state and solid-state NMR. One possibility is that cTIM's low activity is due in part to dysregulation of the loop motions. We attempted to make single-Trp mutants of cTIM for <sup>19</sup>F-Trp incorporation and NMR studies analogous to those of McDermott *et al.*, but the single-Trp168 mutant (W11F W157F W191F) of cTIM is inactive. Further experiments to probe this issue in cTIM and ccTIM are underway.

Finally, it is a surprise that cTIM is even weakly active given its fluid nature, because the TIM reaction is thought to result from highly precise positioning of catalytic residues. The result is reminiscent of the recent discovery of Hilvert *et al.* that an engineered monomeric chorismate mutase from *Methanococcus jannaschii* (mMjCM) has similar catalytic efficiency to its native-like dimeric counterpart.<sup>52,53</sup> The balance of enthalpy and entropy changes upon substrate binding was dramatically altered for mMjCM but with little net effect on the overall free energy. It will be interesting to calorimetrically analyze the binding of cTIM to inhibitors.

## Materials and Methods

### Materials

The TIM single-gene knockout from the Keio collection and DF502 were acquired from the *E. coli* Genetic Stock

Center at Yale University. The Keio strain was lysogenized with Novagen's  $\lambda$ DE3 kit (69734). TEV protease was received from David S. Waugh via the American Type Culture Collection. ANS (A-47) was obtained from Invitrogen. The following reagents were used to determine enzymatic activity. Substrates DHAP (D7137), DL-GAP (G5251), and sodium arsenate (S9663) were purchased from Sigma Aldrich. NAD<sup>+</sup> (10 837 067 001) and NADH (10 107 735 001) were obtained from Roche. Coupling enzymes rabbit GAP dehydrogenase (G2267) and rabbit  $\alpha$ -glycerol-3-phosphate dehydrogenase (G6751) were purchased from Sigma Aldrich.

### Relative entropy calculations

The relative entropies for the ccTIM data set were calculated using the equation:

$$\text{R.E.} = \sum_x p_x \ln \frac{p_x}{f_x}$$

Here,  $p_x$  is the frequency of amino acid  $x$  usage at each position and  $f_x$  is the frequency of amino acid  $x$  usage in a neutral reference state.<sup>54,55</sup> Our reference state is amino acid codon frequencies in the *S. cerevisiae* genome. Amino acid usage in yeast is a close approximation to eukaryotic usage, but advantageously will not change as more sequences become available. *S. cerevisiae* amino acid codon usage is as follows (in percentage): A, 6; C, 1; D, 6; E, 7; F, 4; G, 5; H, 2; I, 7; K, 7; L, 10; M, 2; N, 6; P, 4; Q, 4; R, 4; S, 9; T, 6; V, 6; W, 1; Y, 3. By this calculation, a position with the same distribution as the reference state will have an RE of 0, while absolutely conserved positions will show REs greater than 3.

### Mutual information calculations

The magnitude of correlation between positions in the MSA is calculated as the mutual information:

$$\text{M.I.}(i, j) = \sum_i \sum_j p_{x,y} \ln \frac{p_{x,y}}{p_x p_y}$$

Here,  $p_x$  is the frequency of amino acid residue  $x$  at position  $i$ ,  $p_y$  is the frequency of amino acid residue  $y$  at position  $j$ , and  $p_{x,y}$  is the frequency of co-occurrence of amino acid residue  $x$  at position  $i$  and amino acid residue  $y$  at position  $j$ .<sup>55</sup> For example, if alanine occurs at position  $x$  in 50% of sequences and valine occurs at position  $y$  in 25% of sequences, at random, 12.5% of the data set should contain the Ala–Val pair. If the observed percentage is significantly higher or lower, the positions are correlated or anticorrelated, respectively. The mutual information is logarithmically related to the multinomial probability of observing the given joint distribution of all residues at the two sites given the individual distributions at those sites. The background noise level of mutual information for the data set is calculated by randomizing all columns of the data set and calculating the mutual information for the randomized data set. Increasing mutual information scores correspond to stronger correlations.

### Cloning

The genes for cTIM, ir-cTIM, and ccTIM were constructed from synthetic DNA using the method described

by Stemmer. The *S. cerevisiae* gene was PCR amplified from yeast strain YPH499. Standard PCR reactions consisted of 10 to 25 cycles of [95 °C for 30 s, annealing temperatures ranged from 50 °C to 72 °C for 30 s, with 72 °C extension for 45 s]. cTIM subcloning into pET11a was performed by restriction enzyme digestion and ligation at flanking NdeI and BamHI sites. ir-cTIM and ccTIM were cloned by ligation-independent techniques into pHLIC (V.D., B.J.S. and T.J.M., submitted). Both vectors, pET11a and pHLIC, yield TIMs with traceless hexahistidine affinity tags produced under T7 expression in *E. coli* strains carrying the  $\lambda$ DE3 lysogen. The vectors were transformed into electrocompetent DH10B and plated on LB agar supplemented with ampicillin. The identity of each gene was confirmed by analytical restriction digest and DNA sequencing at Genewiz, Inc.

### Expression and purification

The sequence-confirmed constructs were transformed into electrocompetent BL21(DE3) or Keio(DE3) for over-expression. Shake-flask cultures in 2 × YT media were grown to OD<sub>600</sub> ~ 0.7 and induced with 0.1 mM IPTG. Induced cultures continued to grow at 37 °C for 4 h or were moved to 30 °C for 8 h. Cell pellets from 1-L cultures were resuspended to 30 mL with lysis buffer [50 mM Tris-HCl, pH 8, 300 mM NaCl, 10 mM imidazole, 2 mM β-mercaptoethanol, and 1 mM tris(2-carboxyethyl)phosphine (TCEP)]. Five millimolar MgCl<sub>2</sub>, 0.5 mM CaCl<sub>2</sub>, 2 U mL<sup>-1</sup> DNase (Pierce), 200 ng mL<sup>-1</sup> RNase (Fisher), 0.25 mg mL<sup>-1</sup> lysozyme (Fisher), and 0.1% Triton X-100 were added to the suspensions for 30 min at 4 °C. After incubation, the cell pastes were either sonicated or disrupted by glass beads. Centrifugation was performed to separate the soluble and insoluble fractions. The soluble fraction was bound to 1.5 mL of 50% Ni-NTA agarose (Qiagen). After 1 h, the sample was poured into a pre-fritted column and washed (50 mM Tris-HCl, pH 8, 300 mM NaCl, 20 mM imidazole, 2 mM β-mercaptoethanol, and 1 mM TCEP) before elution (50 mM Tris-HCl, pH 8, 300 mM NaCl, 250 mM imidazole, 2 mM β-mercaptoethanol, and 1 mM TCEP). The 6 × His-TEV-TIM fusion was digested overnight by 6 × His-TEV protease with 5 mM DTT. After quantitative cleavage, the 6 × His tag and 6 × His-TEV protease are removed by a second Ni-NTA step. The purified sample was concentrated and stored at 4 °C in 50 mM phosphate buffer, 300 mM NaCl, 2 mM DTT, and 2 mM TCEP, pH 8. Protein concentration and purity were judged by SDS-PAGE and amino acid analysis.

### CD spectroscopy

Data were collected on a Jasco J-815 CD spectrometer. TIM samples were diluted to 12 μM in 50 mM potassium phosphate and 300 mM NaCl, pH 8. Far-UV spectra were recorded in triplicate from 195 to 250 nm. Wavelength scans recorded ellipticity every nanometer with 2-s integration time and 100 nm min<sup>-1</sup> scanning speed. Data points reporting HT voltages greater than 600 V were discarded. For thermal denaturation, ellipticity was monitored at 222 nm with temperatures increasing from 25 to 95 °C. Data were collected in 1 °C steps with 6-s temperature equilibration, 1 °C min<sup>-1</sup> ramping and 2-s integration. Data were exported and analyzed in Microsoft Excel 2003/2007.

### Gel filtration

Size-exclusion chromatography was performed on a Pharmacia-LKB FPLC. Protein samples were diluted to 37 μM in 50 mM Tris-HCl and 100 mM NaCl, pH 8, and eluted from a Superdex 75 10/300 column (GE Amersham) with 50 mM Tris-HCl and 100 mM NaCl, pH 8, at 0.4 mL min<sup>-1</sup>. Protein peaks were collected by observing the absorbance at 280 nm. Molecular masses for the TIM variants were calculated based on fits from known standards: alcohol dehydrogenase (9.6 mL), 150 kDa; albumin (11.0 mL), 66 kDa; carbonic anhydrase (11.8 mL), 29.0 kDa; cytochrome c (13.6 mL), 12.4 kDa; and aprotinin (15.6 mL), 6.5 kDa.

### Analytical ultracentrifugation

Sedimentation velocity studies were performed by the University of Connecticut AUC Facility. TIM samples were tested at concentrations ranging from 0.1 to 1.5 mg mL<sup>-1</sup> in 50 mM potassium phosphate, 300 mM NaCl, 1 mM TCEP, and 2 mM DTT, pH 8. Runs were performed at 55,000 rpm at 20 °C in a Beckman-Coulter XL-I analytical ultracentrifuge. Data were collected at 60-s intervals for 4 to 6 h. Data were analyzed with DcDt+, version 2.1.0; Sedfit, version 11.3; and Sedphat, version 5.01.

### ANS binding

Fluorescence spectra were recorded using a Perkin-Elmer LS50B spectrometer. The stock concentration of ANS in methanol was determined by absorbance at 372 nm using an extinction coefficient of 8 × 10<sup>3</sup> M<sup>-1</sup> cm<sup>-1</sup>. Protein at 25 μM was incubated with 5 μM dye at room temperature for 5 to 10 min. The samples were excited at 372 nm, and the emission spectra were obtained from 400 to 600 nm.

### Activity

$k_{\text{cat}}$  and  $K_{\text{m}}$  were calculated using the assays described by Plaut and Knowles at 37 °C and pH 7.4. Our data were collected in 96-well plates using a Molecular Devices Spectramax M5 plate reader. Detailed protocols for the activity assay are provided in the [Supplemental Information](#). The Keio(DE3) strain was grown on minimal media agar plates lacking six-carbon sugars. These plates were supplemented with M63 salts, 0.2% w/v glycerol or lactate, 1 mg L<sup>-1</sup> thiamine, 80 mg L<sup>-1</sup> histidine, and 50 mg L<sup>-1</sup> uracil. Plates contained ampicillin for plasmid selection and kanamycin for strain selection. Cells were grown for 1–4 days at 37 °C.

### Acknowledgements

B.J.S. was a National Institutes of Health Chemistry-Biology Interface Program Fellow and Ohio State Presidential Fellow. We are grateful to Deepti Mathur for technical assistance with some of the

enzyme preparation and kinetics. We thank Christopher Jaroniec and Jeffrey Lary for their expertise in NMR and AUC, respectively. This work was supported by The Ohio State University.

## Supplementary Data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.jmb.2011.08.001](https://doi.org/10.1016/j.jmb.2011.08.001)

## References

1. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
2. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
3. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science*, **278**, 82–87.
4. Cordes, M. H., Davidson, A. R. & Sauer, R. T. (1996). Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* **6**, 3–10.
5. Richards, F. M. (1997). Protein stability: still and unsolved problem. *Cell. Mol. Life Sci.* **53**, 790–802.
6. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.
7. Rose, G. D. & Wolfenden, R. (1993). Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **22**, 381–415.
8. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A. *et al.* (2008). De novo computational design of retro-aldol enzymes. *Science*, **319**, 1387–1391.
9. Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J. *et al.* (2008). Kemp elimination catalysts by computational enzyme design. *Nature*, **453**, 190–195.
10. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L. *et al.* (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels–Alder reaction. *Science*, **329**, 309–313.
11. Pabo, C. (1983). Molecular technology. Designing proteins and peptides. *Nature*, **301**, 200.
12. Main, E. R., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. (2003). Design of stable alpha-helical arrays from an idealized TPR motif. *Structure*, **11**, 497–508.
13. Mosavi, L. K., Minor, D. L., Jr. & Peng, Z. Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl Acad. Sci. USA*, **99**, 16029–16034.
14. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Pluckthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332**, 489–503.
15. Lehmann, M., Kostrewa, D., Wyss, M., Brugger, R., D'Arcy, A., Pasamontes, L. & van Loon, A. P. (2000). From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng.* **13**, 49–57.
16. Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L. *et al.* (2002). The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.* **15**, 403–411.
17. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. (2000). The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta*, **1543**, 408–415.
18. Steipe, B., Schiller, B., Pluckthun, A. & Steinbacher, S. (1994). Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.* **240**, 188–192.
19. Ohage, E. & Steipe, B. (1999). Intrabody construction and expression. I. The critical role of VL domain stability. *J. Mol. Biol.* **291**, 1119–1128.
20. Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellenhofer, G. *et al.* (2000). Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol.* **296**, 57–86.
21. Godoy-Ruiz, R., Perez-Jimenez, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. (2005). A stability pattern of protein hydrophobic mutations that reflects evolutionary structural optimization. *Biophys. J.* **89**, 3320–3331.
22. Pey, A. L., Rodriguez-Larrea, D., Bomke, S., Dammers, S., Godoy-Ruiz, R., Garcia-Mira, M. M. & Sanchez-Ruiz, J. M. (2008). Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins*, **71**, 165–174.
23. Magliery, T. J. & Regan, L. (2004). Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. *J. Mol. Biol.* **343**, 731–745.
24. Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B. & Ranganathan, R. (2005). Natural-like function in artificial WW domains. *Nature*, **437**, 579–583.
25. Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H. & Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. *Nature*, **437**, 512–518.
26. Silverman, J. A., Balakrishnan, R. & Harbury, P. B. (2001). Reverse engineering the (beta/alpha)<sub>8</sub> barrel fold. *Proc. Natl Acad. Sci. USA*, **98**, 3092–3097.
27. Alber, T., Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D., Rivers, P. S. & Wilson, I. A. (1981). On the three-dimensional structure and catalytic mechanism of triose phosphate isomerase. *Philos. Trans. R. Soc. London, Ser. B*, **293**, 159–171.
28. Nickbarg, E. B. & Knowles, J. R. (1988). Triosephosphate isomerase: energetics of the reaction catalyzed by the yeast enzyme expressed in *Escherichia coli*. *Biochemistry*, **27**, 5939–5947.
29. Nagano, N., Orengo, C. A. & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765.
30. Desamero, R., Rozovsky, S., Zhadin, N., McDermott, A. & Callender, R. (2003). Active site loop motion in triosephosphate isomerase: T-jump relaxation

- spectroscopy of thermal activation. *Biochemistry*, **42**, 2941–2951.
31. Rozovsky, S., Jogl, G., Tong, L. & McDermott, A. E. (2001). Solution-state NMR investigations of triosephosphate isomerase active site loop motion: ligand release in relation to active site loop dynamics. *J. Mol. Biol.* **310**, 271–280.
  32. Rozovsky, S. & McDermott, A. E. (2001). The time scale of the catalytic loop motion in triosephosphate isomerase. *J. Mol. Biol.* **310**, 259–270.
  33. Williams, J. C. & McDermott, A. E. (1995). Dynamics of the flexible loop of triosephosphate isomerase: the loop motion is not ligand gated. *Biochemistry*, **34**, 8309–8319.
  34. Kempf, J. G., Jung, J. Y., Ragain, C., Sampson, N. S. & Loria, J. P. (2007). Dynamic requirements for a functional protein hinge. *J. Mol. Biol.* **368**, 131–149.
  35. Wang, Y., Berlow, R. B. & Loria, J. P. (2009). Role of loop–loop interactions in coordinating motions and enzymatic function in triosephosphate isomerase. *Biochemistry*, **48**, 4548–4556.
  36. Gerlt, J. A. & Raushel, F. M. (2003). Evolution of function in (beta/alpha)<sub>8</sub>-barrel enzymes. *Curr. Opin. Chem. Biol.* **7**, 252–264.
  37. Stemmer, W. P., Cramer, A., Ha, K. D., Brennan, T. M. & Heyneker, H. L. (1995). Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, **164**, 49–53.
  38. Babul, J. (1978). Phosphofructokinases from *Escherichia coli*. Purification and characterization of the nonallosteric isozyme. *J. Biol. Chem.* **253**, 4350–4355.
  39. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M. *et al.* (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**; 2006.0008.
  40. Magliery, T. J., Lavinder, J. J. & Sullivan, B. J. (2011). Protein stability by number: high-throughput and statistical approaches to one of protein science's most difficult problems. *Curr. Opin. Chem. Biol.* **15**, 443–451.
  41. Schliebs, W., Thanki, N., Jaenicke, R. & Wierenga, R. K. (1997). A double mutation at the tip of the dimer interface loop of triosephosphate isomerase generates active monomers with reduced stability. *Biochemistry*, **36**, 9655–9662.
  42. Borchert, T. V., Abagyan, R., Jaenicke, R. & Wierenga, R. K. (1994). Design, creation, and characterization of a stable, monomeric triosephosphate isomerase. *Proc. Natl Acad. Sci. USA*, **91**, 1515–1518.
  43. Borchert, T. V., Pratt, K., Zeelen, J. P., Callens, M., Noble, M. E., Oppendoerfer, F. R. *et al.* (1993). Over-expression of trypanosomal triosephosphate isomerase in *Escherichia coli* and characterisation of a dimer-interface mutant. *Eur. J. Biochem.* **211**, 703–710.
  44. Borchert, T. V., Zeelen, J. P., Schliebs, W., Callens, M., Minke, W., Jaenicke, R. & Wierenga, R. K. (1995). An interface point-mutation variant of triosephosphate isomerase is compactly folded and monomeric at low protein concentrations. *FEBS Lett.* **367**, 315–318.
  45. Mainfroid, V., Terpstra, P., Beaugard, M., Frere, J. M., Mande, S. C., Hol, W. G. *et al.* (1996). Three hTIM mutants that provide new insights on why TIM is a dimer. *J. Mol. Biol.* **257**, 441–456.
  46. Christensen, H. & Pain, R. H. (1991). Molten globule intermediates and protein folding. *Eur. Biophys. J.* **19**, 221–229.
  47. Ptitsyn, O. B., Pain, R. H., Semisotnov, G. V., Zerovnik, E. & Razgulyaev, O. I. (1990). Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Lett.* **262**, 20–24.
  48. Veech, R. L., Raijman, L., Dalziel, K. & Krebs, H. A. (1969). Disequilibrium in the triose phosphate isomerase system in rat liver. *Biochem. J.* **115**, 837–842.
  49. Putman, S. J., Coulson, A. F., Farley, I. R., Riddleston, B. & Knowles, J. R. (1972). Specificity and kinetics of triose phosphate isomerase from chicken muscle. *Biochem. J.* **129**, 301–310.
  50. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics* **14**, 51–55, 29–32.
  51. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
  52. Pervushin, K., Vamvaca, K., Vogeli, B. & Hilvert, D. (2007). Structure and dynamics of a molten globular enzyme. *Nat. Struct. Mol. Biol.* **14**, 1202–1206.
  53. Vamvaca, K., Vogeli, B., Kast, P., Pervushin, K. & Hilvert, D. (2004). An enzymatic molten globule: efficient coupling of folding and catalysis. *Proc. Natl Acad. Sci. USA*, **101**, 12860–12864.
  54. Magliery, T. J. & Regan, L. (2005). Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics*, **6**, 240.
  55. Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory*, (2nd ed.), John Wiley & Sons, Inc., Hoboken, NJ.
  56. Schliebs, W., Thanki, N., Eritja, R. & Wierenga, R. (1996). Active site properties of monomeric triosephosphate isomerase (monoTIM) as deduced from mutational and structural studies. *Protein Sci.* **5**, 229–239.