Published in final edited form as:

Hum Mutat. 2018 December ; 39(12): 1814–1826. doi:10.1002/humu.23616.

RheoScale: A Tool to Aggregate and Quantify Experimentally-Determined Substitution Outcomes for Multiple Variants at Individual Protein Positions

Abby M. Hodges¹, Aron W. Fenton², Larissa L. Dougherty², Andrew C. Overholt¹, and Liskin Swint-Kruse^{2,*}

¹Department of Natural, Health, and Mathematical Sciences, MidAmerica Nazarene University, Olathe, Kansas, USA;

²The Department of Biochemistry and Molecular Biology, The University of Kansas Medical Center, Kansas City, Kansas, USA

Abstract

Human mutations often cause amino acid changes (variants) that can alter protein function or stability. Some variants fall at protein positions that experimentally exhibit “rheostatic” mutation outcomes (different amino acid substitutions lead to a range of functional outcomes). In ongoing studies of rheostat positions, we encountered the need to aggregate experimental results from multiple variants, to describe the overall roles of individual positions. Here, we present “RheoScale” which generates quantitative scores to discriminate rheostat positions from those with “toggle” (most substitutions abolish function) or “neutral” (most substitutions have wild-type function) outcomes. RheoScale scores facilitate correlations of experimental data (such as binding affinity or stability) with structural and bioinformatic analyses. The RheoScale calculator is encoded into a Microsoft® Excel workbook and an R script. Example analyses are shown for three model protein systems, including one assessed *via* deep mutational scanning. The RheoScale calculator quickly and efficiently provided quantitative descriptions that were in good agreement with prior qualitative observations. As an example application, scores were compared to the example proteins’ structures; strong rheostat positions tended to occur in dynamic locations. In the future, RheoScale scores can be easily integrated into computational studies to facilitate improved algorithms for predicting outcomes of human variants.

Keywords

Deep mutational scanning; protein; personalized medicine; LacI/GalR; pyruvate kinase; TIM barrel

*To whom correspondence should be addressed. lswint-kruse@kumc.edu, Phone: 913-588-0399.

SUPPORTING INFORMATION

Microsoft® Excel and R versions of the RheoScale calculator are included as Supporting Material, along with Supporting Methods and five Supporting Figures. Citations in the Supporting Material include (Bell & Lewis, 2000; Fowler & Fields, 2014; Ha et al., 1989; Kalodimos et al., 2004; Larose, 2016; Markiewicz, Kleina, Cruz, Ehret, & Miller, 1994; Roscoe et al., 2013; Sturges, 1926; Swint-Kruse et al., 2002; Swint-Kruse et al., 1998; Taraban et al., 2008).

CONFLICT OF INTEREST

The authors have no conflicts of interest.

Results from genome sequencing show that any two unrelated people can have as many as 10,000 amino acid differences (variants) among their proteins (Lek et al., 2016; Ng et al., 2008). As such, effective personalized medicine requires the ability to reliably predict which variants give rise to medically-relevant functional change. Indeed, numerous computer algorithms have been developed to predict the impact of specific amino acid substitutions (*e.g.*, (Adzhubei et al., 2010; Bao, Zhou, & Cui, 2005; Bendl et al., 2014; Capriotti, Altman, & Bromberg, 2013; Capriotti, Calabrese, & Casadio, 2006; Capriotti, Fariselli, Calabrese, & Casadio, 2005; Choi, Sims, Murphy, Miller, & Chan, 2012; Hecht, Bromberg, & Rost, 2015; Mathe et al., 2006; Ng & Henikoff, 2001; Niroula, Urolagin, & Vihinen, 2015; Pejaver, Mooney, & Radivojac, 2017; Petukh, Dai, & Alexov, 2016; Petukh, Kucukkal, & Alexov, 2015; Ramensky, Bork, & Sunyaev, 2002; Reva, Antipin, & Sander, 2011; Schymkowitz et al., 2005; Stone & Sidow, 2005; Tang & Thomas, 2016)). However, several studies have noted the need to improve predictions (*e.g.* (Dong et al., 2015; Gray, Kukurba, & Kumar, 2012; Miller, Bromberg, & Swint-Kruse, 2017)), and this is a key goal of CAGI (the ongoing “Critical Assessment of Genome Interpretation”) (*e.g.* (Daneshjou et al., 2017)).

Instead of constructing a new algorithm, we have focused our attention on the *experimental* laboratory data that have provided the foundational definitions and assumptions underlying many algorithms. The focus of this work was to develop quantitative scores for the qualitative descriptions of experimental changes in function or stability that arise when multiple amino acids are substituted at given amino acid positions. By first understanding the general role of each position in a protein, we and others (Zhang, Norris, Schwartz, & Alexov, 2011) have hypothesized that the success of specific variant predictions will be enhanced.

One common expectation for “important” positions within a protein is that most substitutions abolish the protein’s function or greatly diminish the stability of the native conformation (“dead”). This outcome has been frequently observed for positions that require a particular amino acid chemistry and has provided rationale for alanine scanning experiments. The chemical constraint can often be observed in the evolutionary record as a conserved position in a protein sequence alignment. We previously made the analogy that this type of amino acid position is like a toggle switch with protein function either on (with the correct amino acid chemistry) or off (with a chemically-dissimilar variant) (Figure 1A, middle) (Meinhardt, Manley, Parente, & Swint-Kruse, 2013).

Conversely, other protein positions are expected to have little importance, and their amino acid substitutions are expected to have little effect on structure or function (Figure 1A, right). From an evolutionary perspective, many nonconserved positions are expected to be insensitive to amino acid substitution (“neutral”). From a structural perspective, many surface-exposed positions are also expected to be neutral.

Both of these expected scenarios tempt investigators to extrapolate the outcome from one variant (or a few) to the overall role of the position being substituted (neutral or toggle). However, this inference is not appropriate for a third class of protein positions: We observed this class in an experimental study of select nonconserved positions in LacI/GalR homologs

(Meinhardt et al., 2013; Meinhardt & Swint-Kruse, 2008; Tungtur, Egan, & Swint-Kruse, 2007; Tungtur, Meinhardt, & Swint-Kruse, 2010; Tungtur, Parente, & Swint-Kruse, 2011). Many substitutions did alter protein function which indicated that, despite their nonconservation, these positions were important (and thus, not neutral). Additionally, when multiple amino acid variants (~8–12) were substituted into the same position, a wide range of outcomes was observed (*e.g.* Figure 2A), from neutral to intermediate to dead (and thus, not a toggle position). Surprisingly, the outcomes did not correlate with either the common biochemical properties of amino acids (*e.g.* hydrophobicity, charge, *etc.*) or with the evolutionary frequencies observed for each type of amino acid (Meinhardt et al., 2013).

We called these “rheostat” positions (Meinhardt et al., 2013) because, similar to a dimmer switch on a light, substitutions at such positions could be exploited by evolutionary processes to “dial” function up or down as organisms adapt to new niches. A review of other experimental studies suggests that rheostat positions occur in many proteins and that their effects could manifest in many parameters including binding affinities, rate constants, allosteric coupling, and/or protein stability (Swint-Kruse, 2016). We have reasoned that some variants at rheostat positions are likely to be medically-important: “Dead” variants at rheostat positions are catastrophic, and medically-important substitutions have been documented for various nonconserved positions (*e.g.* (de Beer et al., 2013; Pendergrass, Williams, Blair, & Fenton, 2006)). Furthermore, Alexov and colleagues have predicted that some positions with known disease-causing changes can also be substituted with harmless amino acids (Zhang et al., 2011); this complex behavior is indicative of a rheostat position.

Given the potential medical importance and complex substitution outcomes observed at known rheostat positions, we next considered whether discriminating rheostat positions from toggle positions could provide an avenue for improving substitution predictions (Miller et al., 2017). Results of this study showed that the functional outcomes from substituting rheostat positions were very poorly predicted by current algorithms, whereas those at toggle positions were well-predicted (Miller et al., 2017). Thus, the performance of current algorithms could be improved simply by filtering out variants at rheostat positions; for rheostat positions, new (and distinct) algorithms should be developed.

To that end, our long-term goals are (i) to reliably predict the *locations* of rheostat positions and (ii) to understand the biophysical changes that underlie their non-canonical substitution outcomes. For the latter, we are now carrying out relevant structural and functional studies in various model proteins. Among results from our ongoing experiments, we have identified positions that exhibit strong rheostat behavior with substitution outcomes that covered a wide functional range: from more active than wild-type, to near-wild-type, to functionally dead (*e.g.* (Meinhardt et al., 2013)). Results for other positions showed more nuanced and/or complex changes: some positions displayed weaker rheostatic behavior with modest but biologically significant variation from wild-type; other positions exhibited intermediate patterns (*e.g.* part rheostat and part neutral). These various outcomes are represented by simulated data shown in Figure 1A and B.

Together, these results show that assigning a position’s overall substitution behavior using a ternary rheostat/toggle/neutral classification (i) was insufficient to describe the full range of

experimental outcomes, (ii) required investigators to set arbitrary thresholds to discriminate the categories, (iii) returned qualitative descriptions of substitution outcomes, and (iv) was not easily used as input training sets for prediction algorithms or to correlate substitution outcomes with structural or evolutionary properties. These short-comings motivated us to develop the three quantitative scales (rheostat/toggle/neutral) reported herein.

As test cases, we applied these analyses to three experimental datasets for which multiple variants for each substituted position were available: nonconserved positions in the LacI/GalR paralogs (Meinhardt et al., 2013), an allosterically-regulated kinase for which multiple experimental functional parameters are available (pyruvate kinase) (Ishwar, Tang, & Fenton, 2015; Tang, Alontaga, Holyoak, & Fenton, 2017), and three TIM barrel orthologs of indole-3-glycerol phosphate synthase characterized *via* deep mutational scanning (Chan, Venev, Zeldovich, & Matthews, 2017). When used in combination, these rheostat, neutral, and toggle scales produced numerical scores that well-described the qualitative behaviors assigned to the experimentally-substituted positions and provided a facile means for analyzing large experimental datasets.

This report is accompanied by the “RheoScale” calculator tool (available as both a Microsoft® Excel workbook and as an R script) that can be used to quickly calculate position scores for a wide range of experimental data in a wide range of proteins. RheoScale scores can be easily compared with results from structural or sequence-based studies which we expect will aid in uncovering the biophysical parameters that lead to the varied substitution outcomes. To provide an example of this utility, we report initial structural analyses for each of the three protein test cases; results suggest rheostat positions occur in dynamic protein regions. In the future, RheoScale scores can also be used as a feature of protein positions in future prediction algorithms, which may lead to better determinations of variant outcomes.

Editorial Policies and Ethical Considerations

This work was carried out using data for recombinant proteins; review by an ethics board was not required.

RHEOSCALE CALCULATIONS

In developing scales to quantify substitution behavior, we considered the criteria used to subjectively define a position as rheostat, toggle, or neutral. These criteria were: (i) *the number of different functional (or stability) outcomes* that arose from varied amino acid substitutions at any one position; and (ii) *the range of change* observed for these outcomes, as compared to the total possible range determined using multiple positions. Both types of information can be captured in histograms. However, we did not necessarily expect (or observe) these histograms to follow a Gaussian distribution; standard measures of skewness and kurtosis were also not useful for analyzing these data.

Therefore, we developed novel methods to assign histogram bins and to analyze subsequent distribution patterns (see the attached Microsoft® Excel workbook or R script, Supporting Files). To illustrate and develop these calculations, simulated data were generated to

exemplify the three idealized substitution behaviors (rheostat, toggle, and neutral; Figure 1A). Other data were simulated to exemplify four non-ideal datasets (Figure 1B). Further details of the score calculations are provided within the Supplemental Information. An overview of the process is as follows:

For each position, the functional outcomes for each variant were sorted into histogram bins; results can be shown as bars (*e.g.* Figure 1C) for one data set or as dots (*e.g.* Figure 1D) to compare two or more data sets. Next, a score for neutral substitution behavior was calculated from the fraction of variants present in a bin centered on the wild-type value (Figure 1E). A score for toggle substitution behavior was calculated from the fraction of variants present in the bin corresponding to the non-functional value (hereafter referred to as “dead”; Figure 1E). Finally, a score for rheostat substitution behavior was calculated by determining the fraction of all possible bins that were occupied (Figure 1E). All three calculations were normalized so that scores range from 0.0 to 1.0 with higher scores indicating behaviors more consistent with the named behavior. (Note that rheostat scores never reached 0.0 because all positions always have a bin that contains wild-type values.)

In developing these score calculations, three factors were considered: (i) the required features for an experimental data set, (ii) the appropriate choice for the number of histogram bins, and (iii) the use of a weighting scale in calculating the rheostat score. The rationale and requisite details considered for each factor, as well as guidance for setting RheoScale parameters and analytic thresholds, can be found in the Supporting Methods. Directions for using the accompanying RheoScale Microsoft® Excel workbook calculator and the Rheoscale R script calculator are also presented in the Supporting Methods.

In brief, we recommend that experimental data sets should comprise at least ten variants per position. As with most analyses, data sets with smaller experimental errors required fewer variants than those with larger experimental error. Analysis was also facilitated by having data for multiple positions per protein; this aided in determining the full range of functional (or structural) values that can be accessed *via* single amino acid changes in the protein under study.

In choosing a bin number for histogram analyses, we recommend using the largest number that accounts for both the number of variants and the experimental error. The RheoScale calculator automatically aids determination of this value but, as in all histogram analyses, the investigator should systematically vary the bin number manually to empirically validate the parameter. A standardized bin number of 10 allowed good sampling for variant numbers that ranged from 8 to 20. A standardized bin number could also facilitate substitution comparisons across many different datasets and proteins. A weighted rheostat score can be used to provide more confidence that observed changes significantly differ from the wild-type value. Finally, when interpreting RheoScale scores, we empirically found a rheostat score 0.5 to correspond to subjectively assigned rheostat positions for datasets that were analyzed using ~10 bins.

EXAMPLE APPLICATIONS

To develop the RheoScale calculator, we used both simulated data (Figure 1 and Supp. Figure S1) and experimental data previously published for the LacI/GalR paralogs (Meinhardt et al., 2013). We further tested the calculator on published data for pyruvate kinase, an enzyme for which multiple experimental functional parameters are available (Ishwar et al., 2015; Tang et al., 2017), and three TIM barrel orthologs that were characterized using deep mutational scanning (Chan et al., 2017). Finally, we considered the scores in light of known structural features to identify potential features of rheostat positions.

LacI/GalR homologs

Subjective rheostat assignments were integral to a previous study of LacI/GalR transcription repressor homologs (Meinhardt et al., 2013). Therefore, we used that same data set to assess the calculator performance and to demonstrate a range of data presentation styles.

In the previous study, a set of ~1100 LacI/GalR variants were created by random mutagenesis at 12 structurally-analogous, nonconserved positions. The region targeted by these 12 positions was a structurally flexible linker (Ha, Spolar, & Record, 1989; Kalodimos et al., 2004; Swint-Kruse, Larson, Pettitt, & Matthews, 2002; Swint-Kruse, Matthews, Smith, & Pettitt, 1998; Taraban et al., 2008) that, in the presence of DNA, forms multiple interfaces (Swint-Kruse et al., 2002): with the DNA binding domain, the DNA, the regulatory domain, and between two monomers (Supp. Figure S2). For each variant, the ability to repress transcription was compared to a “no repressor” condition which served as the “dead” value for current calculations. Control experiments showed that *in vivo* repression correlated with K_d for DNA binding (Tungtur, Skinner, Zhan, Swint-Kruse, & Beckett, 2011; Zhan, Taraban, Trehwella, & Swint-Kruse, 2008) and that all variants produced folded protein capable of binding DNA which was also expressed at high levels (Meinhardt et al., 2013). The 12 positions under study were subjectively designated as rheostat, toggle, or neutral positions for each homolog (Table 4 of (Meinhardt et al., 2013)). Rheostat behavior dominated the data set (Meinhardt et al., 2013); four other positions interspersed in this region functioned as toggle positions (Miller et al., 2017; Suckow et al., 1996).

In the current analyses, the number of available variants led to recommended bin numbers that ranged from 7–10 depending upon the homolog. We considered using the same bin number for all homologs, but the number of available variants differed enough that we chose to optimize score calculations for each homolog. The average experimental error was small enough that it did not influence the bin number. As will often occur for biochemical data, some positions had fewer variants than bin number (*e.g.*, nine bins were used to analyze data for the homolog “LacI”, but position 55 had only six variants) and therefore the calculated rheostat scores were lower limits. Two of the LacI/GalR homologs were excluded from the current analyses because (i) the poor repression by the wild-type paralog caused this bin to be adjacent to the “dead” bin (homolog “LGhP” in (Meinhardt et al., 2013)) and (ii) the dataset had too few variants (homolog “LLhS” (Meinhardt et al., 2013)).

The RheoScale calculator provided a much faster way to analyze the LacI/GalR experimental data than subjective assignments (minutes instead of hours). Calculation results are shown in Figure 2. Plots for one representative position (LacI position 58) are shown in Figure 2A (original data) and Figure 2B (histogram). A composite histogram for all variants at all positions in the LacI homolog are shown in Figure 2C to depict the total range of change. Calculated scores are shown for each position in Figure 2D (*e.g.*, the scores for position 58 would be determined from the histogram shown in Figure 2B). Scores for multiple homologs are compared *via* various presentations shown in Figure 2E–G. As noted before (Meinhardt et al., 2013), each linker position showed varied levels of rheostat behavior among the homologs. For nine of the twelve substituted positions, most rheostat scores were above 0.5 (Figure 2E–G). In contrast, most neutral and toggle scores were usually < 0.5 (Figure 2G). Therefore, the calculator allowed for a rapid, quantitative evaluation of the rheostat nature for individual protein positions that was consistent with subjective assignments of rheostat substitution behavior.

In addition, new information was gained from considering the full range of change observed for individual proteins. For example, Figure 2C shows the binned data for all 104 variants that were created across 12 positions in the LacI linker region. This histogram showed that a full range of outcomes was observed in the available dataset, and the rheostat score calculated from “All” data was 1.0. Likewise, the “All” rheostat scores were 1.0 for each LacI/GalR homolog. Thus, the LacI/GalR homolog functions were perfectly tunable by single substitutions in the linker region. In the future, it will be interesting to determine whether the linker is a “hotspot” for tuning LacI/GalR functions, or whether the tunability is widespread throughout the structure.

The various presentation styles shown in Figure 2E–G highlighted different features of the substitution behaviors of each position. For a limited number of proteins, line graphs (Figure 2E) were useful for highlighting quantitative differences in the rheostat scores. Heat maps (Figure 2G) were more efficient at comparing equivalent positions in large numbers of homologs. Dot plots (Figure 2G) better showed the full range of scores obtained for each analogous position across all the homologs; this presentation emphasized which positions were most commonly rheostatic, neutral, or toggle-like.

In summary, the calculated substitution behavior scores for each position recapitulated previous findings (Meinhardt et al., 2013) and provided additional insights into the tunability of the LacI/GalR transcription repressor function.

Pyruvate kinase

To test the RheoScale calculator on novel biochemical data, we used data available for human liver pyruvate kinase (hL-PYK) (Ishwar et al., 2015; Tang et al., 2017). These data were generated to study allosteric communication between protein binding sites rather than rheostatic substitution behavior of individual positions, but our preliminary subjective inspection suggested that several positions showed rheostat substitution behavior.

Pyruvate kinase is a glycolytic enzyme subject to extensive allosteric regulation. The affinity of hL-PYK for its substrate, phosphoenolpyruvate (PEP), is reduced upon binding alanine

and enhanced upon binding fructose-1,6-bisphosphate (“Fru-1,6-BP”) (Fenton & Alontaga, 2009; Fenton & Hutchinson, 2009). Thus, the multiple binding events of this protein also provided a model system to explore whether multiple functional parameters are simultaneously altered by substitutions at rheostat positions.

In the available datasets, eight positions near the Fru-1,6-BP site (Ishwar et al., 2015) and nine positions near the alanine binding site (Tang et al., 2017) were extensively substituted (Supp. Figure S3). Five different parameters are routinely reported for hL-PYK variants: K_{a-PEP} is a kinetically-derived apparent affinity for PEP that, due to the rapid equilibrium nature of the enzyme (Boyer, 1969), has been treated like a true dissociation constant. K_{ix-Ala} and $K_{ix-F-1,6-BP}$ are dissociation constants for the two allosteric effectors. Q_{Ala} and $Q_{F-1,6-BP}$ are allosteric coupling constants calculated from the fold-change in K_{a-PEP} caused by the respective allosteric ligand (Reinhart, 1983, 2004). Of these five parameters, three are available for any one hL-PYK variant (K_{a-PEP} and the respective K_{ix} and Q values associated with the allosteric site being substituted).

For RheoScale score calculations, the “dead” value for K_{a-PEP} was represented by variants lacking detectable catalytic activity over the substrate range assayed. In considering these variants, we encountered two challenges. The first was to assign a numerical value to “dead” K_{a-PEP} . For equilibrium dissociation constants, larger values represent diminished function; for “dead” hL-PYK variants, K_{a-PEP} must be greater than any of the substrate concentrations assayed. However, the use of very large values in the RheoScale calculator would have artificially expanded the total max/min range, thereby compressing the other experimental data into a few bins. Thus, we chose a value two orders of magnitude larger than the highest substrate concentration used in the experimental assay; this is the smallest value not detectable in the assay (Figure 3A, black dot). For variants with detectable activity but incomplete binding curves (*i.e.*, no V_{max} , which is required for data fitting), we assigned a value equivalent to the largest substrate concentration used in the experimental assay.

The second challenge was that abolished PEP activity could be due to a catastrophic loss in any of (i) PEP binding (as reflected by the apparent affinity K_{a-PEP}), (ii) binding of the other substrate, ATP, (iii) catalysis (k_{cat}), which is required to measure the apparent affinity, or (iv) protein expression/stability. One way to simplify interpretation is to omit “dead” variants from the score calculations, but this would make the toggle score meaningless. The magnitude of the rheostat score would also change, but the comparisons among rheostat scores would have similar results. Thus, the data shown in figures below include variants with “no PEP activity”, but the interpretation limitations must be kept in mind for “dead” K_{a-PEP} .

The other four parameters – K_{ix} and Q values – were only determined when K_{a-PEP} could be measured which gives confidence that the protein was expressed, folded, and catalytically active. (Although we do not rule out the possibility that any position was a rheostat for protein stability; such data were not available.) For K_{ix} “dead” values, we again chose a value two orders of magnitude greater than the largest ligand concentration used (Figure 3B and D, black dots); for these variants, Q values could not be determined and thus were not part of scores calculated for Q . Q “dead” values occurred when allosteric regulation was

abolished (K_{a-PEP} was the same in the absence and presence of allosteric ligand); this corresponded to a value of 1 for both regulators (Figure 3C and E, black dots). Note that for activator Fru-1,6-BP, $Q_{F-1,6-BP} = 1$ is the range minimum (Figure 3C, black dot), whereas for inhibitor alanine, $Q_{Ala} = 1$ is the maximum (Figure 3E, black dot).

Figure 3 shows the histograms constructed from all variants for the five parameters, and Figure 4 shows the calculated scores for the five parameters for individual positions and “All” values. Due to the small average experimental error, the number of bins was determined using the number of available variants (bin numbers 9–13; Figure 3). K_{a-PEP} data comprised 211 variants distributed among 17 positions in the two allosteric sites. One position (483) had a high toggle score for this parameter (Figure 4); most variants at this position were “dead” and one was severely impaired. All other variants at all other positions had K_{a-PEP} values that fell within 7-fold of the wild-type value (Figure 3A, white dot). The narrow range of change in the current study leads to modest rheostat values for K_{a-PEP} for the “All” data and individual positions (Figure 4). A large, essentially-empty gap (occupied only by the 483 near-dead variant) occurred between the measured and “dead” values.

The narrow range of change in K_{a-PEP} and presence of the gap is intriguing. Values within the gap are accessible by single substitutions as shown by the 483 “near-dead” variant and results from a whole-protein alanine scan (Tang & Fenton, 2017). Further analysis of the current dataset showed the gap and the 7-fold distribution persisted when the two binding sites were analyzed separately from each other (Supp. Figure S4). The 7-fold distribution falls within the activation/inhibition range of the wild-type protein that results from alanine or Fru-1,6-BP binding. It remains to be seen whether other variants near the allosteric sites would fill this gap or whether substitutions in these allosteric regions generally have limited effects on K_{a-PEP} .

In contrast to K_{a-PEP} the other four parameters were highly tunable by single substitutions near their respective allosteric sites (Figure 3 B–E): In other words, all bins between “wild-type” and “dead” were well-populated. For these parameters, the rheostat scores for “All” data approached 1, four individual positions had rheostat scores significantly greater than 0.5 (Figure 4A; 56, 82, 446, and 531), and another six positions had rheostat scores near 0.5. Five other positions had high toggle scores – and correspondingly low rheostat scores – for one of the parameters (Figure 4; 444, 482, 483, 494, and 501). Finally, although a few positions had high neutral scores for K_{a-PEP} , no position was neutral in all measured parameters. That is, all positions contributed to changing at least one of the measured functional parameters.

Preliminary structural analyses do not identify an obvious characteristic that discriminates rheostat and toggle positions. The structure of Fru-1,6-BP-bound hL-PYK (Holyoak et al., 2013) shows that both rheostat and toggle positions make direct contacts to the ligand; this parallels findings for the LacI/GalR homologs, for which both rheostat and toggle positions contacted DNA (Meinhardt et al., 2013). (Note that a structure has not been determined for alanine-bound hL-PYK.) Intriguingly, positions with the highest rheostat scores (Figure 4; 56, 82, 446, and 531) were located outside of regular secondary structure (Supp. Figure S3) (Holyoak et al., 2013), and hydrogen-deuterium exchange experiments with the rabbit

muscle M1 isoform suggest that the regions around positions 55–56, 481–483, and 444–449 have dynamics that change in response to allosteric ligands (Prasannan, Villar, Artigues, & Fenton, 2013). These observations are consistent with the hypothesis that rheostat positions fall in regions with functionally-important dynamics.

In summary, these analyses demonstrated that hL-PYK, with its multi-faceted function, can serve as an interesting experimental system for elucidating the properties of rheostat and toggle positions.

Deep mutational scanning of TIM barrel isozymes

In combination with biological competition assays, recent advances in gene sequencing have enabled “deep mutational scanning” of structure/function outcomes for large libraries of amino acid variants (Fowler & Fields, 2014; Gray, Hause, & Fowler, 2017; Roscoe, Thayer, Zeldovich, Fushman, & Bolon, 2013). Although the assay output is (i) a combination of all possible changes in stability and function, and (ii) sensitive to the threshold of the biological competition assay (Mavor et al., 2016), data for a given position usually report results for all possible amino acid substitutions and thus may provide information about rheostat positions. Here, we have used RheoScale to analyze results for three TIM barrel isozymes that were exhaustively substituted in the central β barrel and parts of the flanking loops (Figure 5A and B) (Chan et al., 2017). This region is known to be important for TIM barrel stability and contains the enzyme active site (Chan et al., 2017; Gangadhara, Laine, Kathuria, Massi, & Matthews, 2013). This data set allowed us to test the score calculations with assay output common to deep mutational scanning (*i.e.*, a biological fitness score) and to compare and contrast rheostat positions in three proteins with structural and functional similarities (albeit with low sequence identities in the range of 30–40%).

For RheoScale score calculations, we used the guidelines established for the experimental study (Chan et al., 2017): The reported “fitness scores” for each variant were normalized by comparisons to the wild-type protein. The total range of change was defined as 1 to –1, with “dead” being –1 and “wild-type” being zero. Several measured values fell outside of this range, but various controls (including coding sequences with stop codons) led Chan *et al.* to designate this as the reliable range for interpretation. In both the previous experimental study and the current calculations, values outside of this range were reset to the appropriate max/min value. The data presented by Chan *et al.* in their Supplementary Figure 9 (Chan et al., 2017) were used to estimate average error for this data set of 0.125. Despite the large number of variants per position (nineteen), this relatively high error was the limiting factor when determining bin number (8 bins) for all three isozymes.

Analyses of RheoScale scores (Figure 5C) were consistent with many findings from the original study: No position had a strong neutral score, all the active site positions (Figure 5C, bottom row, colored positions) had very high toggle scores, and overall there was a high density of positions with toggle scores >0.5 especially when compared to the LacI/GalR or hL-PYK studies. This concentration of toggle positions in the substituted region is consistent with the role of the β barrel structure in stabilizing the overall protein structure (Gangadhara et al., 2013).

In addition, rheostat substitution behavior was observed for various positions in the three isozymes. All three isozymes had “All variant” rheostat scores of 1.0 (data not shown) indicating that the fitness score of each homolog was perfectly tunable by single substitutions in the region targeted for this study. Homolog comparisons showed that positions in the *T. thermophilus* isozyme (“Tt”) had fewer positions with high rheostat scores than the *S. solifatarictus* (“Ss”) or *T. maritima* (“Tm”) isozymes (Figure 6). The decrease in strong rheostat positions did not correlate with a corresponding increase in the number of strong toggle positions: the Tm isozyme had more (Figure 6). We also noted that several positions with high rheostat scores in one isozyme had very low rheostat scores in the other isozymes and correspondingly high toggle scores (Figure 5C).

We next looked for any structural patterns in the locations of the strongest rheostat positions. To that end, we emulated Chan *et al.* (Chan et al., 2017) by treating the substituted regions as repeats of the same structural unit (indicated with repeating arrows at the top of Figure 5C). This unit comprised ten amino acids starting with the last three positions of an N-terminal loop (“ α - β loop”) and the five positions that usually fell into a β strand followed by two positions from the beginning of a C-terminal loop (“ β - α loop”; Figure 5C and Figure 7).

When the scores for the 24 analogous structural units (8 units in each of three isozymes) were analyzed by position (Figure 7), several trends were evident. The first positions in the α - β loop and the β strand had the greatest propensity for high rheostat scores whereas the last three positions of the β strands were highly enriched for toggle positions (Figure 7). This results in a higher concentration of rheostat positions at the one end of the β barrel (Figure 5C), and a higher concentration of toggle positions at the other end, where the active site resides. Since the hL-PYK active site is also located on a TIM barrel domain (Supp. Figure S3), it will be interesting to determine whether the rheostat and toggle positions are similarly enriched on opposite sides of the barrel.

When the TIM structural unit scores were re-cast as heat maps (Supp. Figure S5), the toggle scores also showed a tendency to increase as the strands progressed around the barrel (top to bottom in the heat map). The latter is intriguing considering that a folding intermediate is known for a homologous TIM barrel. Strands 1–4 were folded in this intermediate, whereas strands 5–8 were unfolded (Rojasajakul, Wintrode, Vadrevu, Robert Matthews, & Smith, 2004). The increased toggle scores correlate with the units that fold last.

In summary, both rheostat and toggle positions were detected in data from deep mutational scanning studies of three TIM isozymes. When compared to structure, results suggest intriguing hypothesis about the locations of rheostat and toggle positions on TIM barrels.

CONCLUSION

When outcomes are predicted for amino acid substitutions, they are often assigned to binary categories such as “benign/pathological”. We and Alexov *et al.* (Zhang et al., 2011) have hypothesized that, by first understanding the overall role of an amino acid *position* in protein function or structure, the outcomes of individual substitutions might be better predicted. To

that end, Alexov *et al.* computed stability changes for 19 substitutions at three positions in spermine synthase and assigned two binary, positional categories: “tolerant/non-tolerant” (with respect to whether the majority of substitutions were predicted to alter stability), and “specific/non-specific” (with respect to the direction of the predicted change) (Zhang et al., 2011). In our experimental studies, results showed that the category of “non-tolerant” should be further divided into two classes of positions: rheostat and toggle (Meinhardt et al., 2013).

Here, we extend those concepts to accommodate the continuum between classes, from neutral- to rheostat- to toggle-like behaviors, that were observed in experimental data (Swint-Kruse, 2016). Analyses of data for model proteins show that the neutral, rheostat, and toggle RheoScale scores (i) recapitulated the findings of previously published manual analyses, and (ii) have proven useful for condensing outcomes from multiple amino acid substitutions into a simple quantitative descriptor for each substituted position. Use of the RheoScale calculator greatly sped analyses, and the use of multiple presentation styles emphasized different patterns in the scores and helped pinpoint individual positions and regions for more detailed study. In addition to analyzing data from biochemical studies, this calculator may be particularly useful for analyzing the large datasets that are arising from deep mutational scanning. Indeed, although the latter generally have high error for the individual data points, using the data from all 19 possible substitutions to calculate a general behavior for a position should amplify the signal to noise ratio.

For all three example protein datasets, RheoScale analyses showed that the full functional range was accessible to each protein (except $K_a\text{-}PEP$ for hL-PYK). Nevertheless, of the three experimental datasets analyzed, the LacI/GalR study had a greater fraction of positions for which rheostat character dominated than did the hL-PYK or TIM barrel proteins (Figure 2 versus Figure 3 and Figure 6). This was likely due to the fact that the LacI/GalR study was designed to explore the contributions of functionally-important, nonconserved positions. In contrast, the PYK study was designed to study allostery and the TIM barrel designed to study protein stability. In future studies designed to explore rheostat positions, it will be interesting to see the range of rheostat positions and the magnitudes of their scores that arise.

The trends illuminated by the calculated scores allowed us to note some intriguing structural similarities for the strong rheostat positions. In both LacI/GalR transcription repressors and in hL-PYK, the strong rheostat positions fell into regions known to have functionally important dynamics. The TIM barrel loops also showed a greater propensity for strong rheostat positions than did the β strands. As we continue to ferret out the biophysical roots of rheostat behavior, the calculated positional RheoScale scores could also provide useful input for machine learning algorithms and enable predictions for personalized medicine and protein engineering.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank James Leininger (MidAmerica Nazarene University) for assistance with histogram analyses. We thank Hanna Bradford (Blue Valley ISD Center for Advanced Professional Studies) for assistance analyzing data for the LacI/GalR homologs and Brittany Arce and Edina Kosa (KUMC) for assistance curating the hL-PYK data. We thank C. Robert Matthews (University of Massachusetts Medical School) for introducing us to the TIM isozyme data. This work was supported by grants from the W. M. Keck Foundation and the National Institutes of Health (R01 GM118589).

Grant Numbers: This work was supported by grants from the W. M. Keck Foundation and the National Institutes of Health (R01 GM118589).

REFERENCES

- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, ... Sunyaev SR (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7, 248–249. doi:10.1038/nmeth0410-248 [PubMed: 20354512]
- Bao L, Zhou M, & Cui Y (2005). nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Research*, 33, W480–482. doi:10.1093/nar/gki372 [PubMed: 15980516]
- Bell CE, & Lewis M (2000). A closer view of the conformation of the Lac repressor bound to operator. *Nature Structural Biology*, 7, 209–214. doi:10.1038/73317 [PubMed: 10700279]
- Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, ... Damborsky J (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Computational Biology*, 10, e1003440. doi:10.1371/journal.pcbi.1003440 [PubMed: 24453961]
- Boyer PD (1969). The inhibition of pyruvate kinase by ATP: a Mg⁺⁺ buffer system for use in enzyme studies. *Biochemical and Biophysical Research Communications*, 34, 702–706. [PubMed: 5777786]
- Capriotti E, Altman RB, & Bromberg Y (2013). Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics*, 14 Suppl 3, S2. doi:10.1186/1471-2164-14-S3-S2
- Capriotti E, Calabrese R, & Casadio R (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22, 2729–2734. doi:10.1093/bioinformatics/btl423 [PubMed: 16895930]
- Capriotti E, Fariselli P, Calabrese R, & Casadio R (2005). Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*, 21 Suppl 2, ii54–58. doi:10.1093/bioinformatics/bti1109 [PubMed: 16204125]
- Chan YH, Venev SV, Zeldovich KB, & Matthews CR (2017). Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. *Nature Communications*, 8, 14614. doi:10.1038/ncomms14614
- Choi Y, Sims GE, Murphy S, Miller JR, & Chan AP (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, 7, e46688. doi:10.1371/journal.pone.0046688 [PubMed: 23056405]
- Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, ... Morgan AA (2017). Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*, 38, 1182–1192. doi:10.1002/humu.23280 [PubMed: 28634997]
- de Beer TA, Laskowski RA, Parks SL, Sipos B, Goldman N, & Thornton JM (2013). Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Computational Biology*, 9, e1003382. doi:10.1371/journal.pcbi.1003382 [PubMed: 24348229]
- Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, & Liu X (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, 24, 2125–2137. doi:10.1093/hmg/ddu733 [PubMed: 25552646]

- Fenton AW, & Alontaga AY (2009). The impact of ions on allosteric functions in human liver pyruvate kinase. *Methods in Enzymology*, 466, 83–107. doi:10.1016/s0076-6879(09)66005-5 [PubMed: 21609859]
- Fenton AW, & Hutchinson M (2009). The pH dependence of the allosteric response of human liver pyruvate kinase to fructose-1,6-bisphosphate, ATP, and alanine. *Archives of Biochemistry and Biophysics*, 484, 16–23. doi:10.1016/j.abb.2009.01.011 [PubMed: 19467627]
- Fowler DM, & Fields S (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, 11, 801–807. doi:10.1038/nmeth.3027 [PubMed: 25075907]
- Gangadhara BN, Laine JM, Kathuria SV, Massi F, & Matthews CR (2013). Clusters of branched aliphatic side chains serve as cores of stability in the native State of the HisF TIM barrel protein. *Journal of Molecular Biology*, 425, 1065–1081. doi:10.1016/j.jmb.2013.01.002 [PubMed: 23333740]
- Gray VE, Hause RJ, & Fowler DM (2017). Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. *Genetics*, 207, 53–61. doi:10.1534/genetics.117.300064 [PubMed: 28751422]
- Gray VE, Kukurba KR, & Kumar S (2012). Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. *Bioinformatics*, 28, 2093–2096. doi:10.1093/bioinformatics/bts336 [PubMed: 22685075]
- Ha JH, Spolar RS, & Record MT Jr. (1989). Role of the hydrophobic effect in stability of site-specific protein-DNA complexes. *Journal of Molecular Biology*, 209, 801–816. doi:0022-2836(89)90608-6 [pii] [PubMed: 2585510]
- Hecht M, Bromberg Y, & Rost B (2015). Better prediction of functional effects for sequence variants. *BMC Genomics*, 16 Suppl 8, S1. doi:10.1186/1471-2164-16-S8-S1
- Holyoak T, Zhang B, Deng J, Tang Q, Prasannan CB, & Fenton AW (2013). Energetic coupling between an oxidizable cysteine and the phosphorylatable N-terminus of human liver pyruvate kinase. *Biochemistry*, 52, 466–476. doi:10.1021/bi301341r [PubMed: 23270483]
- Ishwar A, Tang Q, & Fenton AW (2015). Distinguishing the interactions in the fructose 1,6-bisphosphate binding site of human liver pyruvate kinase that contribute to allostery. *Biochemistry*, 54, 1516–1524. doi:10.1021/bi501426w [PubMed: 25629396]
- Kalodimos CG, Biris N, Bonvin AM, Levandoski MM, Guennuegues M, Boelens R, & Kaptein R (2004). Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, 305, 386–389. doi:10.1126/science.1097064 [PubMed: 15256668]
- Larose DT (2016). *Discovering statistics* (Third edition ed.). New York: W.H. Freeman & Company, a MacmillanEducation imprint.
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, ... Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536, 285–291. doi:10.1038/nature19057 [PubMed: 27535533]
- Markiewicz P, Kleina LG, Cruz C, Ehret S, & Miller JH (1994). Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as “spacers” which do not require a specific sequence. *Journal of Molecular Biology*, 240, 421–433. doi:10.1006/jmbi.1994.1458 [PubMed: 8046748]
- Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, & Tavtigian SV (2006). Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Research*, 34, 1317–1325. doi:10.1093/nar/gkj518 [PubMed: 16522644]
- Mavor D, Barlow K, Thompson S, Barad BA, Bonny AR, Cario CL, ... Fraser JS (2016). Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *Elife*, 5, e15802. doi:10.7554/eLife.15802 [PubMed: 27111525]
- Meinhardt S, Manley MW, Becker NA, Hessman JA, Maher LJ, & Swint-Kruse L (2012). Novel insights from hybrid LacI/GalR proteins: family-wide functional attributes and biologically significant variation in transcription repression. *Nucleic Acids Research*, 40, 11139–11154. doi:10.1093/nar/gks806 [PubMed: 22965134]

- Meinhardt S, Manley MW Jr., Parente DJ, & Swint-Kruse L (2013). Rheostats and toggle switches for modulating protein function. *PLoS One*, 8, e83502. doi:10.1371/journal.pone.0083502 [PubMed: 24386217]
- Meinhardt S, & Swint-Kruse L (2008). Experimental identification of specificity determinants in the domain linker of a LacI/GalR protein: bioinformatics-based predictions generate true positives and false negatives. *Proteins: Structure, Function, and Bioinformatics*, 73, 941–957. doi:10.1002/prot.22121
- Miller M, Bromberg Y, & Swint-Kruse L (2017). Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Scientific Reports*, 7, 41329. doi:10.1038/srep41329 [PubMed: 28134345]
- Ng PC, & Henikoff S (2001). Predicting deleterious amino acid substitutions. *Genome Research*, 11, 863–874. doi:10.1101/gr.176601 [PubMed: 11337480]
- Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, ... Venter JC (2008). Genetic variation in an individual human exome. *PLoS Genetics*, 4, e1000160. doi:10.1371/journal.pgen.1000160 [PubMed: 18704161]
- Niroula A, Urolagin S, & Vihinen M (2015). PON-P2: prediction method for fast and reliable identification of harmful variants. *PLoS One*, 10, e0117380. doi:10.1371/journal.pone.0117380 [PubMed: 25647319]
- Pejaver V, Mooney SD, & Radivojac P (2017). Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. *Human Mutation*, 38, 1092–1108. doi:10.1002/humu.23258 [PubMed: 28508593]
- Pendergrass DC, Williams R, Blair JB, & Fenton AW (2006). Mining for allosteric information: natural mutations and positional sequence conservation in pyruvate kinase. *IUBMB Life*, 58, 31–38. doi:10.1080/15216540500531705 [PubMed: 16540430]
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, & Ferrin TE (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25, 1605–1612. doi:10.1002/jcc.20084 [PubMed: 15264254]
- Petukh M, Dai L, & Alexov E (2016). SAAMBE: Webserver to predict the charge of binding free energy caused by amino acids mutations. *International Journal of Molecular Sciences*, 17, 547. doi:10.3390/ijms17040547 [PubMed: 27077847]
- Petukh M, Kucukkal TG, & Alexov E (2015). On human disease-causing amino acid variants: Statistical study of sequence and structural patterns. *Human Mutation*, 36, 524–534. doi:10.1002/humu.22770 [PubMed: 25689729]
- Prasannan CB, Villar MT, Artigues A, & Fenton AW (2013). Identification of regions of rabbit muscle pyruvate kinase important for allosteric regulation by phenylalanine, detected by H/D exchange mass spectrometry. *Biochemistry*, 52, 1998–2006. doi:10.1021/bi400117q [PubMed: 23418858]
- Ramensky V, Bork P, & Sunyaev S (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, 30, 3894–3900. [PubMed: 12202775]
- Reinhart GD (1983). The determination of thermodynamic allosteric parameters of an enzyme undergoing steady-state turnover. *Archives of Biochemistry and Biophysics*, 224, 389–401. [PubMed: 6870263]
- Reinhart GD (2004). Quantitative analysis and interpretation of allosteric behavior. *Methods in Enzymology*, 380, 187–203. doi:10.1016/S0076-6879(04)80009-0 [PubMed: 15051338]
- Reva B, Antipin Y, & Sander C (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*, 39, e118. doi:10.1093/nar/gkr407 [PubMed: 21727090]
- Rojsajjakul T, Wintrode P, Vadrevu R, Robert Matthews C, & Smith DL (2004). Multi-state unfolding of the alpha subunit of tryptophan synthase, a TIM barrel protein: Insights into the secondary structure of the stable equilibrium intermediates by hydrogen exchange mass spectrometry. *Journal of Molecular Biology*, 341, 241–253. doi:10.1016/j.jmb.2004.05.062 [PubMed: 15312776]
- Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, & Bolon DNA (2013). Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of Molecular Biology*, 425, 1363–1377. doi:10.1016/j.jmb.2013.01.032 [PubMed: 23376099]

- Schneider B, Knochel T, Darimont B, Hennig M, Dietrich S, Babinger K, ... Sterner R (2005). Role of the N-terminal extension of the (betaalpha)8-barrel enzyme indole-3-glycerol phosphate synthase for its fold, stability, and catalytic activity. *Biochemistry*, 44, 16405–16412. doi:10.1021/bi051640n [PubMed: 16342933]
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, & Serrano L (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, 33, W382–388. doi:10.1093/nar/gki387 [PubMed: 15980494]
- Stone EA, & Sidow A (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research*, 15, 978–986. doi: 10.1101/gr.3804205 [PubMed: 15965030]
- Sturges HA (1926). The choice of a class interval Case I Computations involving a single series. *Journal of the American Statistical Association*, 21, 65–66. doi:10.1080/01621459.1926.10502161
- Suckow J, Markiewicz P, Kleina LG, Miller J, Kisters-Woike B, & Müller-Hill B (1996). Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *Journal of Molecular Biology*, 261, 509–523. doi: 10.1006/jmbi.1996.0479 [PubMed: 8794873]
- Swint-Kruse L (2016). Using evolution to guide protein engineering: The devil IS in the details. *Biophysical Journal*, 111, 10–18. doi:10.1016/j.bpj.2016.05.030 [PubMed: 27410729]
- Swint-Kruse L, Larson C, Pettitt BM, & Matthews KS (2002). Fine-tuning function: correlation of hinge domain interactions with functional distinctions between LacI and PurR. *Protein Science*, 11, 778–794. [PubMed: 11910022]
- Swint-Kruse L, Matthews KS, Smith PE, & Pettitt BM (1998). Comparison of simulated and experimentally determined dynamics for a variant of the LacI DNA-binding domain, Nlac-P. *Biophysical Journal*, 74, 413–421. [PubMed: 9449341]
- Tang H, & Thomas PD (2016). PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*, 32, 2230–2232. doi:10.1093/bioinformatics/btw222 [PubMed: 27193693]
- Tang Q, Alontaga AY, Holyoak T, & Fenton AW (2017). Exploring the limits of the usefulness of mutagenesis in studies of allosteric mechanisms. *Human Mutation*, 38, 1144–1154. doi:10.1002/humu.23239 [PubMed: 28459139]
- Tang Q, & Fenton AW (2017). Whole-protein alanine-scanning mutagenesis of allostery: A large percentage of a protein can contribute to mechanism. *Human Mutation*, 38, 1132–1143. doi: 10.1002/humu.23231 [PubMed: 28407397]
- Taraban M, Zhan H, Whitten AE, Langley DB, Matthews KS, Swint-Kruse L, & Trewella J (2008). Ligand-induced conformational changes and conformational dynamics in the solution structure of the lactose repressor protein. *Journal of Molecular Biology*, 376, 466–481. doi:10.1016/j.jmb.2007.11.067 [PubMed: 18164724]
- Tungtur S, Egan SM, & Swint-Kruse L (2007). Functional consequences of exchanging domains between LacI and PurR are mediated by the intervening linker sequence. *Proteins: Structure, Function, and Bioinformatics*, 68, 375–388. doi:10.1002/prot.21412
- Tungtur S, Meinhardt S, & Swint-Kruse L (2010). Comparing the functional roles of nonconserved sequence positions in homologous transcription repressors: Implications for sequence/function analyses. *Journal of Molecular Biology*, 395, 785–802. doi:10.1016/j.jmb.2009.10.001 [PubMed: 19818797]
- Tungtur S, Parente DJ, & Swint-Kruse L (2011). Functionally important positions can comprise the majority of a protein's architecture. *Proteins: Structure, Function, and Bioinformatics*, 79, 1589–1608. doi:10.1002/prot.22985
- Tungtur S, Skinner H, Zhan H, Swint-Kruse L, & Beckett D (2011). *In vivo* tests of thermodynamic models of transcription repressor function. *Biophysical Chemistry*, 159, 142–151. doi:10.1016/j.bpc.2011.06.005 [PubMed: 21715082]
- Zhan H, Taraban M, Trewella J, & Swint-Kruse L (2008). Subdividing repressor function: DNA binding affinity, selectivity, and allostery can be altered by amino acid substitution of nonconserved residues in a LacI/GalR homologue. *Biochemistry*, 47, 8058–8069. doi:10.1021/bi800443k [PubMed: 18616293]

Zhang Z, Norris J, Schwartz C, & Alexov E (2011). *In silico* and *in vitro* investigations of the mutability of disease-causing missense mutation sites in spermine synthase. PLoS One, 6, e20373. doi:10.1371/journal.pone.0020373 [PubMed: 21647366]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Data and calculations for simulated examples with 14 amino acid variants at each position. (A) Simulated functional (or structural) outcomes are shown for ideal rheostat, toggle, and neutral positions. (B) Simulated outcomes are shown for positions with complex substitution outcomes. For both panels A and B, “wild-type” variants are shown with stippled bars. Data were simulated with a 10% error. The dashed horizontal lines show the maximum and minimum simulated values. The “dead” variant corresponds to the maximum value, as would be the case for any equilibrium dissociation constant or Michaelis constant (K_d or

K_m). (C) Histograms created from the simulated data for the ideal rheostat, toggle, and neutral positions shown in panel A. Values corresponding to wild-type and dead protein are indicated with large circles for both panels C and D. (D) Histograms created from the four non-ideal simulated experimental data shown in panel B. In order to compare two different positions, symbols are used rather than bars. In both C and D, the bin number recommended by the calculator – 14 – was used. (E) Neutral (white), unweighted rheostat (gray), weighted rheostat (hatched), and toggle (black) scores calculated from the histograms shown in panels C and D. The dashed line at 0.5 is to aid visual inspection of the scores. Note that all of the non-ideal cases have rheostat scores close to 0.5. In the three complex examples, this value mirrors the fraction of variants that have rheostat character; the non-rheostat variants raise the toggle or neutral scores. For the “modest rheostat” example, the rheostat score of 0.5 arises because the variant outcomes span about half of the total possible functional range (dashed lines); the toggle and neutral scores remain very low.

Figure 2.

Example score calculations using data from a study of variants in LacI/GalR homologs (Meinhardt et al., 2013). (A) Representative experimental data are shown for 9 amino acid variants at position 58 in the dimeric version of LacI (gray bars). The dashed line shows the minimum value (strongest repression) observed in the overall study for a LacI variant; this value was obtained for a variant at a different position. The black bar indicates the value for “no repression” and corresponds to the value of a “dead” LacI variant. The experimental values in (A) were used to generate the histogram shown in (B); the bin number

recommended by the calculator – 9 – was used. The bins corresponding to the “wild-type” and “dead” values are shown with green and magenta dots, respectively. (C) Summary histogram for 104 single amino acid variants created by substituting 12 different positions in dimeric LacI. Note that all possible bins are occupied. (D) Calculated neutral (magenta), rheostat (green), and toggle (black) scores for 12 positions that were substituted in dimeric LacI, and rheostat scores for “All” data. Note that several LacI positions had neutral or toggle scores of 0.0. The dashed line at 0.5 is to aid visual inspection of the scores. (E-G) Comparisons of scores calculated for analogous positions in different LacI/GalR homologs. “LacI” was the dimeric version of the naturally occurring *E. coli* lactose repressor protein. The other LacI/GalR homologs listed in the figure (*e.g.* LLhF, LLhR) were chimeras comprising the LacI DNA binding domain and the linkers/regulatory domains from various naturally occurring homologs; for more details, see (Meinhardt et al., 2012). (E) If only a few homologs are to be compared, RheoScale scores for each position can be plotted on the same graph with different symbols for detailed appraisal. (F) For comparison of multiple homologs, these plots show the full range of RheoScale scores obtained at each position (magenta dots). The average and standard deviation for each position are shown with black bars. The dashed line at 0.5 is to aid visual inspection of the data. Most of the positions in this region of the LacI/GalR homologs have strong rheostat scores and low neutral and toggle scores. (G) A heat map is another useful way to compare scores across many homologs. In this example, the color legend indicates the strength of weighted rheostat scores for the homologs on the Y axis and the positions listed on the X axis; gray boxes with black “X’s” denote positions with an insufficient number of variants to reliably calculate a score. Colored boxes with gray X’s indicate positions with fewer variants than the bin number; this results in the rheostat score being a lower limit.

Figure 3.

Composite histograms for functional parameters of hL-PYK variants. Histograms were created using experimentally determined parameters ((A) K_{a-PEP} , (B) K_{ix-Ala} , (C) Q_{Ala} , (D) $K_{ix-F-1,6-BP}$ and (E) $Q_{F-1,6-BP}$) for all available variants at all positions. Experimental data were taken from (Ishwar et al., 2015; Tang et al., 2017). Values on the x-axes correspond to the upper value of the bin, in log scale. The number of variants used for the composite calculations (“n”) are listed at the top of each panel. As described in the text, variants that lacked catalytic activity for PEP did not have measurements for K_{ix} or Q_{ax} ; variants that

lacked K_{ix} binding activity did not have measurements for Q_{ax} . Bins containing the “wild-type” (white) and “dead” (black) values are shown with dots.

Figure 4.

Calculated scores for substituted positions in and near the allosteric sites of hL-PYK.

Experimental data were taken from (Ishwar et al., 2015; Tang et al., 2017). Calculated scores (rheostat-top, neutral-bottom left, toggle-bottom right) are shown for three functional parameters at each of the noted positions. Dashed vertical lines separate the scores for individual positions. The dashed horizontal line at $Y=0.5$ is to aid visual inspection of the data. Rheostat scores calculated from the composite of all variants at all positions (“All”) are shown at the left of the top panel.

Figure 5.

Structures and RheoScale scores for TIM barrel isozymes. (A) Side and (B) top views of the ribbon diagram for a TIM barrel structure. The substituted regions (Chan et al., 2017) are colored with alternating green and magenta. Each of these regions comprise a β strand and parts of the flanking N- and C-loops; these eight strands form the central beta barrel of the protein. These colored regions correspond to the arrow schematics in panel C. The structure was created using pdb 2C3Z of the Ss isozyme (Schneider et al., 2005) and was rendered with UCSF chimera (Pettersen et al., 2004). (C) RheoScale scores were derived from the experimental data of Chan *et al.* (Chan et al., 2017). Structurally equivalent positions are shown on the x-axis using the numbering for the Ss isozyme. The 8 substituted units, each containing 10 amino acids, are shown with schematics at the top of the panel. The top nine rows of the heat map depict the calculated neutral, rheostat, and toggle scores for each isozyme. The bottom row shows the locations of active site positions; those in dark green are in common to all three isozymes; position 161 (light green) is in common to the Tm and Tt isozymes; position 159 (pink) participates in the active site of the Ss isozyme. As expected, all active site positions have high toggle scores and low rheostat scores.

Figure 6.

Distributions of scores calculated for positions in the three TIM isozymes. The rheostat (top), neutral (middle), and toggle (bottom) scores are shown as gray dots; the averages and standard deviations are marked with black bars for each isozyme.

Structurally Analogous Positions

Figure 7.

Distributions of scores calculated for positions in the analogous structural units in the three TIM isozymes. Each unit comprised ten amino acids, starting with the last three positions of an N-terminal loop (“ α - β loop”), the 5 positions that usually fell into a β strand, followed by two positions from the beginning of a C-terminal loop (“ β - α loop”). The scores for the 24 analogous positions (one from each of 8 units in each of three isozymes) are shown with gray dots. The averages and standard deviations for each position are marked with black bars.