BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 22 2012, pages 2905–2913
doi:10.1093/bioinformatics/bts541

Structural bioinformatics

Advance Access publication September 4, 2012

Post-translational modifications induce significant yet not extreme changes to protein structure

Fuxiao Xin and Predrag Radivojac*

School of Informatics and Computing, Indiana University, Bloomington, IN, USA

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: A number of studies of individual proteins have shown that post-translational modifications (PTMs) are associated with structural rearrangements of their target proteins. Although such studies provide critical insights into the mechanics behind the dynamic regulation of protein function, they usually feature examples with relatively large conformational changes. However, with the steady growth of Protein Data Bank (PDB) and available PTM sites, it is now possible to more systematically characterize the role of PTMs as conformational switches. In this study, we ask (1) what is the expected extent of structural change upon PTM, (2) how often are those changes in fact substantial, (3) whether the structural impact is spatially localized or global and (4) whether different PTMs have different signatures.

Results: We exploit redundancy in PDB and, using root-mean-square deviation, study the conformational heterogeneity of groups of protein structures corresponding to identical sequences in their unmodified and modified forms. We primarily focus on the two most abundant PTMs in PDB, glycosylation and phosphorylation, but show that acetylation and methylation have similar tendencies. Our results provide evidence that PTMs induce conformational changes at both local and global level. However, the proportion of large changes is unexpectedly small; only 7% of glycosylated and 13% of phosphorylated proteins undergo global changes >2 Å. Further analysis suggests that phosphorylation stabilizes protein structure by reducing global conformational heterogeneity by 25%. Overall, these results suggest a subtle but common role of allostery in the mechanisms through which PTMs affect regulatory and signaling pathways.

Contact: predrag@indiana.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 31, 2012; revised on August 3, 2012; accepted on August 27, 2012

1 INTRODUCTION

Post-translational modifications (PTMs) refer to *in vivo* biochemical processing events of a protein after its synthesis (Walsh, 2006). It is speculated that nearly every protein undergoes some form of PTM (Lodish, 2004) and >400 types of PTMs have been reported so far, spanning all domains of life. Different PTMs display different physicochemical properties (Mann and Jensen, 2003); thus, the same protein may exhibit different functions upon different modifications (Jungblut *et al*., 2008). As a result, the high diversity of PTMs, combined with their reversibility and enzymatic control, makes them a vital component of molecular recognition, signal transduction and protein degradation (Deribe *et al*., 2010; Uy and Wold, 1977; Walsh *et al*., 2005; Wold, 1981). Dysregulation of PTMs and mutation of PTM sites are implicated in a number of diseases (Vidal, 2011), from various monogenic disorders (Li *et al*., 2010) to complex diseases such as cancer (Bode and Dong, 2004; Krueger and Srivastava, 2006; Radivojac *et al*., 2008), heart disease (Van Eyk, 2011) and neurodegenerative disorders (Gong *et al*., 2005; Thomas *et al*., 2004).

The mechanisms through which PTMs regulate protein function are of great interest to biologists. Most PTM events introduce additional chemical groups to residue side chains with the potential to alter the energy landscape of a protein and subsequently lead to conformational changes observed in crystal structures. Various examples have shown that this structural change is essential for the modified protein to display new functionalities as in the case of phosphorylation (Blasie *et al*., 1990; Edreira *et al*., 2009; Giannopoulos *et al*., 2009; Lee and Koland, 2005; Menet and Rosbash, 2011), glycosylation (Arnold *et al*., 2007), acetylation (Gu and Roeder, 1997) and sumoylation (Geiss-Friedlander and Melchior, 2007). Additional mechanisms include change of binding affinity or creation of binding sites (Deribe *et al*., 2010; Nishi *et al*., 2011; Schaller and Parsons, 1995; Toh *et al*., 2001).

The most extensively studied PTM is phosphorylation that, with some exceptions, adds a phosphoryl group to serine, threonine or tyrosine residues in eukaryotes and to histidine or aspartic acid residues in prokaryotes. The phosphoryl group has a double negative charge under physiological conditions and is anticipated to affect the energy landscape of the modified protein (Stock and Da Re, 2000). In their review, Johnson and Lewis (2001) analyzed 17 pairs of phosphorylated and non-phosphorylated structures to characterize the structural consequences of phosphorylation. They showed that the dominant structural response was an adjustment of protein conformation to accommodate for the electrostatic effects between the phosphate and surrounding charged atoms. However, the types and extent of structural changes were highly diverse: they observed both local and long-range changes; both association and disassociation of protein complexes and both order-to-disorder and disorder-to-order transitions. In one extreme case, phosphorylation of Ser14 in glycogen phosphorylase results in a 50Å shift of Ser14 itself. In addition, this phosphorylation event alters the tertiary structure of enzyme's catalytic site that is around 50 Å away from Ser14. However, there are also situations in which

*To whom correspondence should be addressed.

Table 1. Number of clusters and sites identified for each PTM after various stages of data filtering

PTMs	Initial data			After removing clusters without both RMSD^u and RMSD^m		After removing clusters without RMSD^u or RMSD^m	
	No. of clusters	No. of sites	No. of disordered sites	No. of clusters	No. of sites	No. of clusters	No. of sites
Glycosylation	175	303 (269N,17S,17T)	9	121	236 (205N,17S,14T)	64	136 (115N,13S,8T)
Phosphorylation	70	89 (47S,10T,22Y,10H)	28	54	55 (32S,4T,11Y,8H)	32	30 (16S,3T,5Y,6H)
Acetylation	16	17 (1A,5C,6K,5S)	6	14	13 (1A,4C,4K,4S)	10	8 (1A,2C,3K,2S)
Mono-methylation	15	17 (2N,1C,4H,1I,7K)	3	13	10 (2N,3H,1I,4K)	8	7 (3H,1I,3 K)

The numbers in parentheses provide breakdown over different amino acid residues. Four of the acetylation sites were *N*-terminal.

phosphorylation and other PTMs introduce no detectable conformational change. We found multiple such cases in this study; for example, *Pseudomonas putida* benzoylformate decarboxylase (1bfdA is phosphorylated; 3fsjX is not); *Zea mays* polyamine oxidase (1b37C is glycosylated; 1h83C is not) or human lysine methyltransferase SET7 (2f69B is methylated; 3m59B is not). In each of these cases, the global root mean-square deviation (RMSD) between two structures was $\leq$0.13 Å and the local RMSD, within 6 Å of the modification site, was $\leq$0.05 Å.

In addition to the analysis of experimentally determined structures, computational approaches have also been explored (Narayanan and Jacobson, 2009). Common strategies include molecular dynamics and conformational sampling. However, both of these strategies are limited by several factors, including computational requirements necessary for modeling micro- to millisecond events on large molecules, assumptions on the scale of conformational change or influence of a particular force field (Lwin and Luo, 2006; Narayanan and Jacobson, 2009). Recent studies have tested the accuracy of computational models by predicting the structure of the phosphorylated molecule based on the structure of the unmodified molecule and then comparing the predicted with the actual phosphorylated structure (Groban *et al*., 2006; Shen *et al*., 2005). The results of these and other studies (Latzer *et al*., 2008) suggest that such methods may be accurate enough to provide valuable insights into the structure–function relationship.

Despite the recent progress in understanding the structural impact of PTMs, much of the focus has been on individual proteins. However, with the rapid growth of protein structure data as well as the presence of multiple structures corresponding to the same amino acid sequence, larger scale studies focused on characterizing the overall trends of the structural impact are becoming realistic. This is further facilitated by the results of recent work in which multiple X-ray structures of the same protein in Protein Data Bank (PDB) (Berman *et al*., 2000) were reported to be similar to those observed in solution using nuclear magnetic resonance (Lange *et al*., 2008), suggesting that different X-ray structures of the same protein can in principle be used to study and understand protein conformational flexibility.

In this study, we systematically analyze groups of protein structures (corresponding to the same sequence) in their modified and unmodified forms to address questions regarding the universality, extent and signatures of structural changes upon PTM. Our work provides evidence that PTMs, similar to ligand binding, induce generally small but statistically significant conformational changes.

2 METHODS

2.1 Data collection and experimental protocol

Protein structures and sequences corresponding to the SEQRES fields were obtained from PDB. RNA, DNA and ligand sequences were discarded and only polypeptide sequences were retained. CD-HIT (Yang *et al*., 2010), which can cluster a sequence database at a given sequence identity threshold, was used to form clusters of PDB chains corresponding to identical sequences. Only clusters with two chains or more were kept. We then examined corresponding PDB files to find clusters in which chains had different MODRES profiles.

We analyzed four PTMs in this study: glycosylation, phosphorylation, acetylation and methylation. Regular expression patterns 'GLYCO', 'PHOSPHO', 'ACETYL' and 'METHYL' were used to retrieve clusters that may contain any of the PTM types. Then, the chemical component dictionary from PDB was consulted to retain proteins with appropriate modification descriptors (the list of descriptors is shown in Supplementary Table S1). Only protein structures with resolution $\leq$2.5 Å and *R*-value $\leq$0.3 were retained. The final dataset contained 276 clusters, each with at least one modified and one unmodified structure; see Table 1 for a detailed breakdown. The average sequence identity between clusters was 19.2% (median was 18.6%), and the average number of structures for each cluster was 7.9 (median was 4). The experimental protocol is illustrated in Figure 1.

2.2 Calculation of RMSD, hydrogen bonds, crystal contacts and salt bridges

Although all PDB chains in a CD-HIT cluster had the same amino acid sequence, calculating the RMSD between pairs was not straightforward because missing residues in the corresponding 3D coordinate (ATOM) fields led to situations in which two structures were not directly superimposable. We established residue correspondences in each pair of structures by performing a global alignment between the two sequences concatenated from the 3D coordinate fields, allowing for gaps but not for mismatches. Then, a least-squares fitting of aligned amino acids was used to calculate the RMSD. Only Cα atoms were used for RMSD calculations.

Local structural environments were defined as concentric shells using radii 6, 12, 18 and 24 Å from a PTM site or its counterpart in an

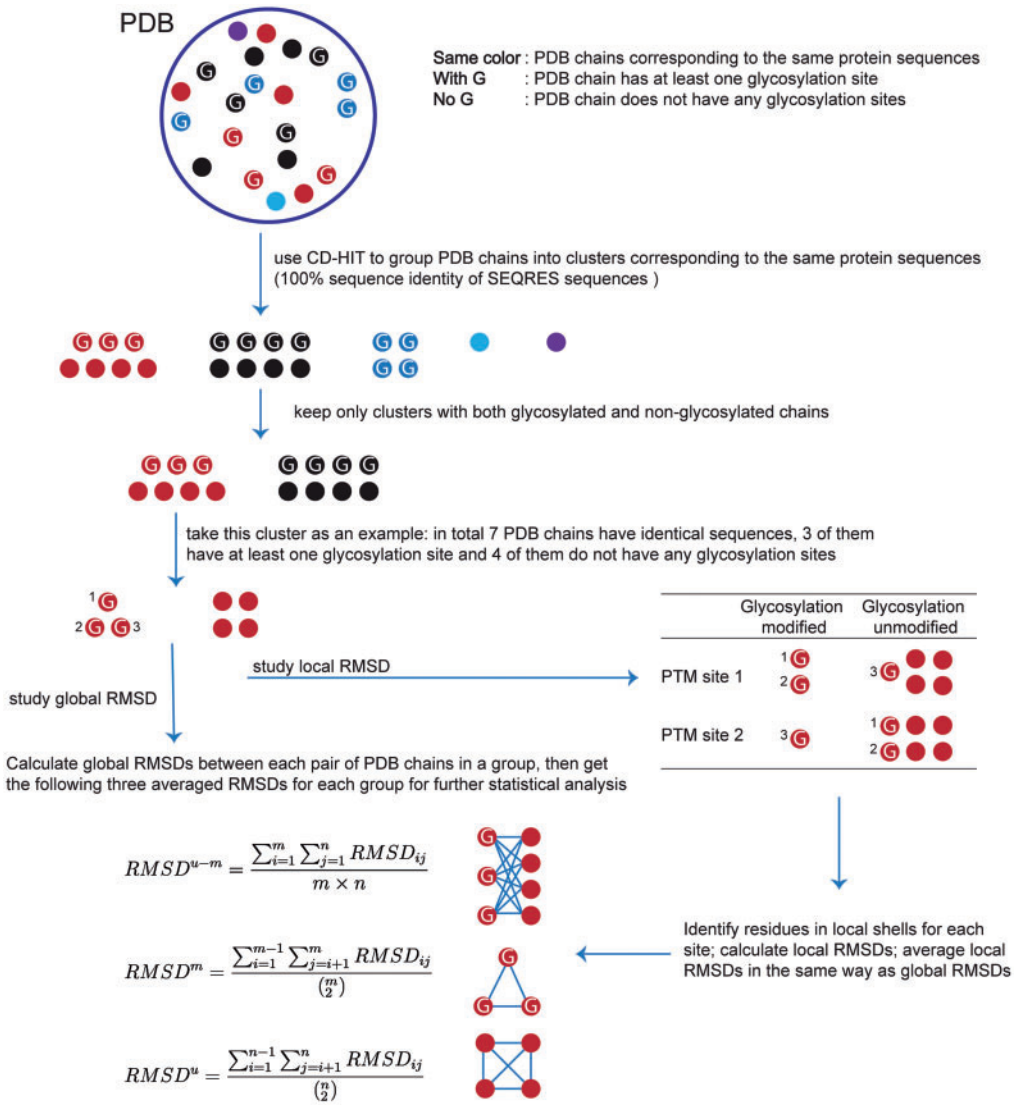

Fig. 1. Experimental procedure for collecting RMSD data for statistical analysis. Considering a protein cluster with $m = 3$ modified and $n = 4$ unmodified structures, we define the between-group RMSD ($RMSD^{u-m}$), where $RMSD_{ij}$ is the RMSD between i-th modified structure and j-th unmodified structure. The two within-group RMSDs, $RMSD^{m}$ and $RMSD^{u}$, were calculated separately on the sets of modified and unmodified structures. The two within-group RMSDs and one between-group RMSD were collected for each PTM site in a cluster and were then subject to a paired t-test to explore whether the between-group RMSD is significantly greater than the within-group RSMD corresponding to the unmodified structures, i.e. to test whether PTMs significantly change protein structure. In these experiments, any situation in which the between-group RMSD was greater than the within-group RMSD (i.e. $RMSD^{u-m} > RMSD^{u}$) was interpreted as structural change upon modification. On the other hand, comparisons between $RMSD^{u}$ and $RMSD^{m}$ were used to suggest potential stabilizing or destabilizing effect upon modification. For example, if the conformational heterogeneity upon modification increases, a particular modification event has destabilized the protein, which could then be supported by the fact that $RMSD^{m} > RMSD^{u}$

unmodified chain (the average longest distance between any two Cα atoms of the protein structures involved in this study was 62 Å). Local RMSD was calculated in a similar way as global RMSD but only included amino acids within the local environment distance cutoffs.

Hydrogen bonds were calculated using HBPLUS (McDonald and Thornton, 1994), but the bonds between main chain atoms and water molecules were excluded since they were not expected to contribute to a change in the number of hydrogen bonds when modified and unmodified chains were compared. Crystal contacts were calculated using CryCo (Eyal *et al.*, 2005) with a default threshold distance of 10 Å. A salt bridge was reported when a positively charged atom was within 4 Å of a negatively charged atom.

2.3 Statistical analysis

The paired t-test was used for hypothesis testing, with the significance level set to 0.05. Generalized linear model (GLM) fitting was used to explore the influence of various parameters on the structural effects upon PTM. We briefly summarize the GLM framework below (Agresti, 2007).

GLM is a generalization of the standard linear model in which the target variable y is modeled as a linear combination of the predictor variables (features) $x_1, x_2, \ldots, x_n$, that is $y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n + e$, where $a_{i\in\{0,1,n\}}$ are real valued coefficients and e is a stochastic error term modeled using a normal distribution $\mathcal{N}(0, \sigma^2)$. In GLM, the target variable is modeled as $y = g(a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n + e)$,

where the inverse function of $g(\cdot)$, $f(\cdot)$, is called the link function. The GLM can be re-written using the link function as $f(E[y]) = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$, where $E[y]$ is the expectation of y. GLM provides a convenient mechanism to model the error using distributions other than normal (e.g. a binomial distribution is used in logistic regression). An identity link function reduces GLM to a standard linear regression model in which $e:\mathcal{N}(0,\sigma^2)$.

The quality of the fit is tested using deviance, defined as $D = -2(\log(L) - \log(L_0))$, where L is the likelihood for the fitted model and L_0 is the likelihood for the full model which is a strict memorization of each data point (Agresti, 2007). Deviance can be viewed as a log-likelihood ratio statistic and thus it approximately follows a χ^2 distribution with the degree of freedom equaling the number of data points minus the number of parameters in the model. Deviance can be used to calculate P-values in order to evaluate goodness-of-fit. Small P-values indicate poor fits, whereas ones that are not very small indicate adequate fits.

In our experiments, we seek to understand which predictor variables contribute to the observed difference between modified and unmodified structures within the same CD-HIT cluster (recall that all structures in a cluster correspond to the same amino acid sequence). To accomplish this, we used a vector-space representation in which each data point pertained to either one cluster (in calculations of global structural changes) or one PTM site (in calculations of local structural changes). The predictor variables included the following: (1) the difference between the indicator variables between modified and unmodified chains; since we compare a group of modified with a group of unmodified structures within a cluster, this value will be 1 for all data points ($x_1 = 1$); (2) the difference between the average number of crystal contacts in two groups (x_2); (3) the difference between the average number of chains in protein complex for two groups (x_3); (4) the difference between the average protein crystallization temperature (x_4); (5) the difference between the average pH (x_5); (6) the difference between the average protein structure resolutions (x_6) and (7) the difference between the average number of ligands in the protein complex (x_7).

Two target variables were used for the GLM fitting: (1) $y = \log(\text{RMSD}^v) - \log(\text{RMSD}^u)$ and (2) $y = \text{RMSD}^v - \text{RMSD}^u$; and the one with the better fit was selected. Here, RMSD^v represents RMSD^{u-m} in each analysis of structural change upon modification and RMSD^m in the analysis of stabilization or destabilization of protein structure. Both types of fitting enable estimating the expected extent of structural change upon modification as follows.

Let us first consider a situation where the target variable is defined as $y = \log(\text{RMSD}^v) - \log(\text{RMSD}^u)$. Suppose that the first predictor variable is non-zero (i.e. $x_1 = 1$, with the regression coefficient a_1), while the remaining variables all equal to zero ($x_2 = x_3 = \ldots = x_7 = 0$), meaning that only PTM conditions are different between the two structure subgroups.

Since the identity link function assumes a normally distributed noise variable, the target can be expressed as

$$\log\left(\frac{\text{RMSD}^v}{\text{RMSD}^u}\right) = a_1 + \sigma Z,$$

where a_1 is the expectation of the target, σ is the standard deviation of the noise variable and Z is a standard normal variable. Both a_1 and σ can be estimated via the GLM fitting (Agresti, 2007). A normally distributed target $y : \mathcal{N}(a_1, \sigma^2)$ means that a transformed random variable e^y follows a log-normal distribution, i.e.

$$\frac{\text{RSMD}^v}{\text{RSMD}^u} = e^{a_1+\sigma Z}.$$

Because the expectation of e^y equals $e^{a_1+\frac{\sigma^2}{2}}$, the structural change upon PTM can be calculated as

$$\frac{\text{RMSD}^v - \text{RMSD}^u}{\text{RMSD}^u} = e^{a_1+\frac{\sigma^2}{2}} - 1.$$

If the target variable is $y = \text{RMSD}^m - \text{RMSD}^u$, instead of the relative difference, we can only report the expected absolute structural difference, which equals a_1.

It is important to note that because of the influence of other variables such as pH or temperature, the structural change due to PTM cannot be directly calculated from data as an average of the observed structural differences. Therefore, in cases when the fitting is adequate, the regression step estimates the proportion of the structural impact that can be attributed to the first predictor variable (x_1).

3 RESULTS

The major goal of this study was to quantify the expected structural difference between unmodified and post-translationally modified proteins and thus understand the allosteric potential of PTMs. We exploited the presence of multiple structures with identical amino acid sequence in PDB that provided us with a means to approximate conformational flexibility of each protein (Lange *et al.*, 2008). In order to work with a sufficient number of structures, we focused on the four most abundant PTM types in PDB: glycosylation, phosphorylation, acetylation and methylation. Because the datasets for acetylation and methylation were too small to carry out a reliable statistical analysis, our main focus was on glycosylation and phosphorylation.

3.1 PTMs significantly change protein local structure

We first asked to what extent the additional chemical group, or a polysaccharide molecule, impacts the local structural neighborhood around a PTM site. To understand this, for each PTM site, we studied the average RMSD between sets of modified and unmodified structures (RMSD^{u-m}) and how it relates to the RMSDs calculated on structures within each group (modified or unmodified). We then tested the hypothesis that PTMs significantly alter protein structure. Note that by using this approach, we compared RMSDs calculated on structures with similar number of atoms, as suggested previously by Gutteridge and Thornton (2005).

We studied local RMSD in four concentric shells defined by the distance (d) from the PTM site and only residues whose Cα atoms were within the shell were included in the RMSD calculation. We compared the within-group RMSD and between-group RMSD for each local environment using a one-tailed paired t-test with the null hypothesis that the within-group and between-group RMSDs are identical and the alternative hypothesis that between-group RMSD is larger than within-group RMSD. The percentage of clusters for which $\text{RMSD}^{u-m} > \text{RMSD}^u$, shown in Table 2, indicates preferences of all four PTMs for local structural re-arrangements. The P-values shown in Table 3 provide statistical support that glycosylation and phosphorylation affect protein local structure in all three layers when $d > 6$. The results are not significant for the $d \leq 6$ layer probably because there are on average <6 amino acids in this layer. Although few tests for methylation and acetylation suggest significant structural changes, potentially due to a small sample size (10 sites for methylation and 13 for acetylation), the results show a similar trend as those for glycosylation and phosphorylation. This is suggested by the observation that the majority of the cases have RMSD^{u-m} larger than RMSD^u. Supplementary Figure S1 shows the distributions of all RMSDs

Table 2. Comparison of $RMSD^{u-m}$ and $RMSD^{u}$ in the local environment of PTM sites

	Glycosyl	Phosphoryl	Methyl	Acetyl	All four
$d \leq 6$	64.0	57.8	50.0	72.7	62.9
$6 < d \leq 12$	66.1	68.2	57.1	54.6	65.7
$12 < d \leq 18$	67.7	63.4	50.0	60.0	66.4
$18 < d \leq 24$	71.4	56.4	75.0	70.0	69.0

Percentage of PTM sites where $RMSD^{u-m}$ is greater than $RMSD^{u}$ in the local structural environment. Variable d represents the distance from the PTM site.

Table 3. Comparison of $RMSD^{u-m}$ and $RMSD^{u}$ in the local environment of PTM sites

	Glycosyl	Phosphoryl	Methyl	Acetyl	All four
$d \leq 6$	1.2×10^{-4}*	0.056	0.120	0.189	0.002*
$6 < d \leq 12$	0.013*	0.011*	0.805	0.178	0.002*
$12 < d \leq 18$	5.9×10^{-7}*	0.013*	0.521	0.197	1.2×10^{-4}*
$18 < d \leq 24$	5.7×10^{-8}*	0.004*	0.165	0.190	5.2×10^{-5}*

t-test results corresponding to the values above. Each *P*-value was calculated using a paired *t*-test. *indicates *P*-values <0.05.

in the local environment for the $d \leq 6$ Å sphere (other local environments have similar distributions; Supplementary Table S2).

Previous studies suggested that crystallization conditions, protein complex formation and crystal packing may influence protein structure (Mohan *et al.*, 2009; Palaninathan *et al.*, 2008) and result in a difference between protein crystal structure and its structure *in vivo* (Eyal *et al.*, 2005). We therefore explored the GLM fitting to seek explanatory variables associated with structural changes. It can be observed from Table 4 that fitting was adequate (large *P*-values for the goodness of fit). Furthermore, the *P*-values of coefficients for PTM were significant for both glycosylation and phosphorylation as well as for the four PTM types together. Although some other factors also influenced the fit, PTM was a contributing factor explaining the observed structural change. The results suggest that glycosylation on average increases local RMSD ($d \leq 6$ Å) by 0.074 Å, while phosphorylation increases it by 0.651 Å. Note that each of the statistical tests shown in Table 3 was performed on a separate dataset and thus does not require correction for multiple hypothesis testing. Similarly, the GLM fitting in Table 4 was run to test the hypothesis that a particular PTM is a significant explanatory variable for the observed change in structure, as opposed to a 'discovery mode' in which one seeks to identify and report any subset of explanatory variables for a particular phenomenon.

An interesting question arises regarding the percentage of cases with large conformational changes upon PTM. We find that changes >0.5 Å occur in 8.1% and 20.0% of cases for glycosylation and phosphorylation, respectively. Similar percentages were also observed for acetylation and methylation.

Table 4. Comparison of $RMSD^{u-m}$ and $RMSD^{u}$ in the local environment of PTM sites

	Glycosylation		Phosphorylation		All four PTMs	
	Coeff	*P*	Coeff	*P*	Coeff	*P*
PTM	0.074	0.012*	0.651	0.013*	0.185	0.004*
No. of crystal contacts	0.005	0.824	0.102	0.089	–0.079	0.028*
No. of chains	–0.001	0.979	0.106	0.240	0.003	0.949
Temperature	–0.001	0.423	–0.004	0.363	–0.000	0.926
pH	0.003	0.902	–0.393	0.022*	–0.061	0.323
Resolution	–0.315	0.007*	0.242	0.661	0.093	0.698
No. of ligands	–0.016	0.436	–0.314	0.049*	–0.031	0.478
Goodness of fit	1.000		0.953		1.000	

Generalized linear model fitting for the observed local structural change between $RMSD^{u}$ and $RMSD^{u-m}$ when $d \leq 6$. The target variable was $RMSD^{u-m} - RMSD^{u}$. The crystal contacts were counted only in the local environment.

Table 5. Comparison of $RMSD^{u}$ with $RMSD^{u-m}$ and $RMSD^{u}$ with $RMSD^{m}$ in global environment of PTM sites

	Glycosyl	Phosphoryl	Methyl	Acetyl	All four
$RMSD^{u-m}$, $RMSD^{u}$	70.6	58.7	70.0	61.5	66.2
$RMSD^{m}$, $RMSD^{u}$	50.0	62.5	37.5	60.0	43.3

Percentage of PTM sites where $RMSD^{u-m}$ is greater than $RMSD^{u}$ in the entire protein structure.

Table 6. Comparison of $RMSD^{u}$ with $RMSD^{u-m}$ and $RMSD^{u}$ with $RMSD^{m}$ in global environment of PTM sites

	Glycosyl	Phosphoryl	Methyl	Acetyl	All four
$RMSD^{u-m}$, $RMSD^{u}$	5.7×10^{-4}*	0.004*	0.478	0.129	1.8×10^{-4}*
$RMSD^{m}$, $RMSD^{u}$	0.489	0.025*	0.236	0.398	0.050*

t-test results corresponding to the structural changes above.
*Indicates *P*-values <0.05.

3.2 PTMs significantly change protein global structure

We next investigated whether PTMs induce structural change at a global protein level. Although our general approach is similar to that in Section 3.1, in this case all comparisons were carried out at the level of unique protein chains instead of at the level of PTM sites. Thus, some of the protein structures unavoidably contained more than one modified residue. As shown in Table 1, the data contained 121 protein chains (clusters) for glycosylation (236 sites), 54 chains for phosphorylation (55 sites), 14 chains for acetylation (13) and 13 for methylation (10).

Table 7. GLM fitting results for global $RMSD^u$ and $RMSD^{u-m}$

	Glycosylation		Phosphorylation		All four PTMs	
	Coeff	*P*	Coeff	*P*	Coeff	*P*
PTM	0.336	4.0×10^{-4}*	0.421	0.079	0.445	8.8×10^{-7}*
No. of crystal contacts	–0.002	0.533	–0.005	0.507	–0.001	0.641
No. of chains	–0.030	0.647	–0.053	0.770	–0.002	0.981
Temperature	–0.001	0.735	0.004	0.410	0.001	0.483
pH	–0.062	0.385	0.056	0.788	–0.050	0.490
Resolution	–0.623	0.108	0.571	0.360	–0.161	0.617
No. of ligands	–0.024	0.688	–0.025	0.879	–0.028	0.629
σ of noise	0.621		0.871		0.786	
Goodness of fit	1.000		0.805		1.000	

The target variable was $\log \mathrm{RMSD}^{u-m} - \log \mathrm{RMSD}^{u}$.

Table 8. GLM fitting results for global $RMSD^u$ and $RMSD^m$

	Glycosylation		Phosphorylation		All four PTMs	
	Coeff	*P*	Coeff	*P*	Coeff	*P*
PTM	–0.005	0.959	–0.696	0.036*	–0.069	0.520
No. of crystal contacts	–0.006	0.085	0.011	0.286	–0.004	0.254
No. of chains	–0.106	0.335	0.307	0.218	–0.028	0.750
Temperature	–0.001	0.808	0.012	0.036*	0.005	0.078
pH	–0.014	0.869	–0.516	0.163	0.044	0.648
Resolution	–0.746	0.118	1.862	0.017*	0.144	0.727
No. of ligands	0.027	0.694	0.082	0.688	–0.027	0.707
σ of noise	0.647		0.900		0.841	
Goodness of fit	1.000		0.675		0.979	

The target variable was $\log \mathrm{RMSD}^{m} - \log \mathrm{RMSD}^{u}$.

Analyses and statistical tests summarized in Tables 5–8 provide evidence that PTMs significantly change protein structure at the global level for both glycosylation ($P = 5.7 \times 10^{-4}$) and phosphorylation ($P = 4.0 \times 10^{-3}$) compared with the unmodified structures. In addition, observed structural changes are strongly related to PTM rather than any other factor. As shown in Table 7, large *P*-values for GLM fitting suggest adequate linear fitting and significant *P*-values for the coefficient of PTM but not any other explanatory factor (4.0×10^{-4} and 8.8×10^{-7} for glycosylation and all four PTMs together). The *P*-value of the PTM coefficient for phosphorylation was not significant at a 0.05 level but it was considerably smaller than *P*-values for any other factor (Table 7), suggesting the need for more data (there were only 54 data points for phosphorylation). Coefficients of GLM fitting suggest that glycosylation on average increases protein global structure RMSD by 69.7% ($a_1 = 0.336, \sigma = 0.621$). Similarly, phosphorylation increases global RMSD by 122.6% ($a_1 = 0.421, \sigma = 0.871$).

In terms of extreme changes, structural changes >2 Å were observed in only 13.0% and 6.6% of cases for phosphorylation and glycosylation, respectively. These results were similar for acetylation and methylation and were consistent with those observed at the local structure level.

3.3 Phosphorylation stabilizes protein structure at a global level

When comparing within-group RMSDs and between-group RMSDs, we compared $RMSD^m$ and $RMSD^u$ with $RMSD^{u-m}$ separately and found similar results. In order to understand the conformational flexibility between unmodified and modified forms of the proteins, we next studied the difference between $RMSD^m$ and $RMSD^u$. The results of this analysis are shown in Supplementary Figures S1 and S2 and Tables 5–8.

The distributions of the two within-group RMSDs show a similar trend observed in the comparison between within-group and between-group RMSDs (Supplementary Figs S1 and S2). The distributions of $RMSD^u$ shift toward the right-hand side compared with the distribution of $RMSD^m$ at both local and global levels, suggesting that PTMs might be able to reduce internal structural movements and thus stabilize protein structures. GLM fitting was performed on those significant comparisons to explore whether PTM was the main explanatory variable. For local RMSDs, the significant explanatory factors include the number of crystal contacts and crystallographic resolution, whereas the presence of a PTM was not significant. For global RMSD of phosphorylation, PTM was significant ($P = 0.036$), along with temperature ($P = 0.036$) and resolution ($P = 0.017$), suggesting that phosphorylation significantly stabilizes protein structure. The correlation between the temperature value, resolution and RMSD suggest that both larger difference in crystallographic resolution and larger temperature difference result in larger RMSD difference. The coefficient of PTM from GLM fitting results provides evidence that on average phosphorylation reduces the global structural difference between two proteins by 25.2% ($a_1 = -0.696, \sigma = 0.900$).

To explore the mechanism of PTM-induced structural changes, we analyzed the change in hydrogen bonds and the number of salt bridges between modified and unmodified structures. For both glycosylation and phosphorylation, we observed a significant increase in the number of hydrogen bonds in the local environment ($P = 1.57 \times 10^{-4}$ and 7.84×10^{-5} for $d \leq 6$ Å). The analysis of salt bridges showed that the phosphoryl group introduced new salt bridges in 45 of 70 protein structures (64.3%; $P = 0.036$).

4 DISCUSSION

Although the importance of PTMs as functional modulators has been established, the mechanisms through which most of the regulation is carried out are still not well understood (Walsh, 2006). In this study, we investigated the potential for allosteric regulation in PTM-mediated functional changes by quantifying structural impact upon PTM (allosteric effect is usually seen as a specific form of structural change in which binding of an effector molecule at one site in a protein alters the local structure around a functional site elsewhere in the protein, thus

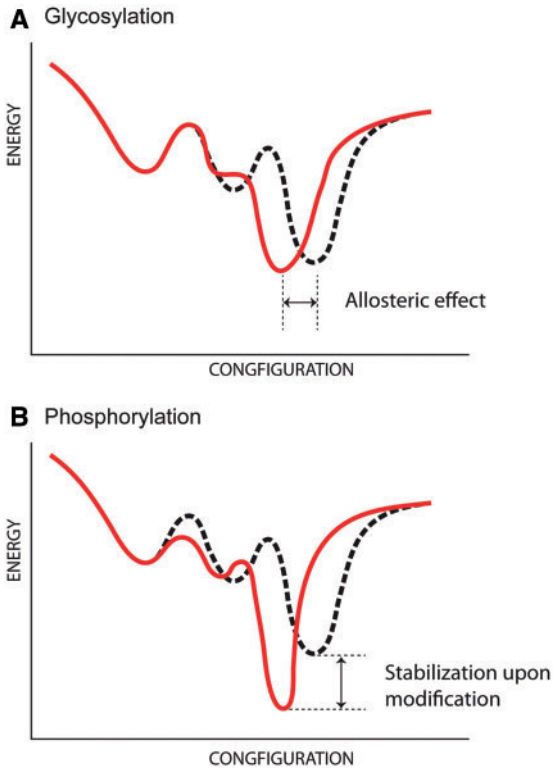

Fig. 2. The speculated energy landscape changes upon phosphorylation and glycosylation. The dotted black curve and the solid red curve correspond to unmodified and post-translationally modified structures, respectively

affecting overall protein activity; Swain and Gierasch, 2006). Our results provide quantitative evidence that PTMs induce significant conformational changes to protein structure and suggest that PTMs act in similar ways as small-molecule allosteric effectors.

We investigated four PTMs, glycosylation, phosphorylation, methylation and acetylation, and showed that all exhibit similar effects in local (Tables 2–4) and global (Tables 5–8) conformational changes. In addition, phosphorylation has showed an effect of stabilizing protein structure by introducing new hydrogen bonds and salt bridges in the local neighborhood of the modified residue. Putting these results together, we speculate that the predominant mechanism of PTM action is alteration of the energy landscape, as shown in Figure 2. Specifically, we believe that glycosylation and phosphorylation frequently lead to a conformational shift of the lowest valley in the energy landscape; however, while glycosylation likely retains the approximate abundance of the conformation with the lowest energy, phosphorylation results in an enriched abundance in the lowest energy form thus restricting conformational flexibility. Similar conclusion has been speculated for intrinsically disordered proteins (Ma and Nussinov, 2009). For phosphorylation, the stabilizing effects might be a driving force to populate protein conformations to a new state, which has been suggested for ligand binding (Hilser, 2010). Conformational changes were also found in methylation and acetylation; however, due to the problems of dataset size, only glycosylation and phosphorylation showed statistically significant differences in most experiments.

Although these results suggest preferences among PTMs for conformational shifts, only a small fraction of structures go through extreme changes. At a global level, glycosylation and phosphorylation introduce structural changes >2 Å in only 7–13% of cases. These results are similar to those observed for ligand binding where 9% of enzymes showed >2 Å structural changes (Gutteridge and Thornton, 2005). These results strongly suggest that despite the importance of structural change for the modified protein to modulate its function, small-to-moderate structural changes are usually sufficient.

Experiments in this study were carried out with strict controls. We only compared conformational heterogeneity between (groups of) protein structures corresponding to identical amino acid sequences. Although such a requirement greatly reduced the number of data points that can be used for statistical analyses (one data point per sequence cluster or PTM site), the approach was necessary since absolute RMSD values are not directly comparable when calculated on very different numbers of atoms. In addition, since RMSDs were calculated using Cα atoms only, side-chain alterations that may also be critical for protein function (Lee *et al.*, 2008) could not be observed. We believe this resulted in more conservative estimation of the prevalence of structural change (note that allostery may occur without any observable backbone changes; Tsai *et al.*, 2008). Another reason that the estimates of conformational changes may be conservative is the influence of the expression system when studying PTMs. In particular, non-observed *N*-linked glycosylation sites expressed in eukaryotic systems may still be glycosylated in the protein, but with the polysaccharide molecule missing from the structural model due to static disorder (Rhodes, 2006). On the other hand, proteins expressed in bacterial systems would be less likely to include such problems.

A potential limitation of this study stems from the suitability of crystallographic data for the study of conformational changes as well as the assumptions used in our statistical analysis. Although crystallographic data are generally reliable, its limitations are related to the inherent biases in PDB (Peng *et al.*, 2004) and its ability to provide high-resolution insight into conformational flexibility of macromolecules. For example, PTMs that increase flexibility of protein regions leading to order-to-disorder transition could not be analyzed in this study. Statistically, one limitation stems from GLM fitting where we included seven variables that are believed to be the most important factors leading to observed structural differences. However, other factors may also exist as well as an interplay between them. For example, a PTM can lead to protein complex formation (Nishi *et al.*, 2011), while in the GLM fitting they would be considered as independent events. Another limitation comes from the fact that a large enough dataset could not be collected to investigate the influence of modifications of different amino acid residue types. As the size of PDB increases, it will become possible to further refine the analysis.

It is important to mention that PTMs have also been linked to intrinsically disordered protein regions, i.e. regions without a single dominant conformational macro-state under physiological conditions (Radivojac *et al.*, 2007). For example, phosphorylation, ubiquitination, methylation and others have been associated to disordered regions either statistically (Daily *et al.*,

2005; Iakoucheva *et al*., 2004; Radivojac *et al*., 2010; Xie *et al*., 2007) or experimentally (Collins *et al*., 2008; Gsponer *et al*., 2008). Although such associations are certainly useful for our understanding of the mechanisms underlying PTM regulation and signaling, a large number of proteins do contain PTM sites in their structured regions. Therefore, the results obtained through our experiments are of broad importance.

Finally, in this work we provide evidence that the observed differences between modified and unmodified structures are significant and can be attributed to PTM. However, the available data do not contain intermediate structures that lead from one observed conformation to another. Thus, the structural differences between modified and unmodified structures could be explained equally well by two alternative mechanisms: structural change upon modification and conformational selection from a pre-existing structural ensemble (our preliminary analyses suggest that both may be at play). Regardless of the underlying mechanism, PTMs are associated with small but common conformational changes of their target proteins.

ACKNOWLEDGEMENTS

We thank Dr Charles E. Dann III for helpful discussions and suggestions on the project and Wyatt T. Clark for proofreading this article.

Funding: National Science Foundation (DBI-0644017 to PR); Don Brown Fellowship to FX.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (2007) *An Introduction to Categorical Data Analysis*. 2nd edn. John Wiley and Sons, Inc, Hoboken, NJ.

Arnold,J. *et al.* (2007) The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu. Rev. Immunol.*, **25**, 21–50.

Berman,H. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Blasie,J. *et al.* (1990) Large-scale structural changes in the sarcoplasmic reticulum ATPase appear essential for calcium transport. *Biophys. J.*, **58**, 687–693.

Bode,A. and Dong,Z. (2004) Post-translational modification of p53 in tumorigenesis. *Nat. Rev. Cancer*, **4**, 793–805.

Collins,M. *et al.* (2008) Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. *Mol. Cell. Proteomics*, **7**, 1331–1348.

Daily,K. *et al.* (2005) Intrinsic disorder and protein modifications: building an SVM predictor for methylation. In: *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, San Diego, CA, pp. 475–481.

Deribe,Y. *et al.* (2010) Post-translational modifications in signal integration. *Nat. Struct. Mol. Biol.*, **17**, 666–672.

Edreira,M. *et al.* (2009) Phosphorylation-induced conformational changes in Rap1b: allosteric effects on switch domains and effector loop. *J. Biol. Chem.*, **284**, 27480–27486.

Eyal,E. *et al.* (2005) The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *J. Mol. Biol.*, **351**, 431–442.

Geiss-Friedlander,R. and Melchior,F. (2007) Concepts in sumoylation: a decade on. *Nat. Rev. Mol. Cell. Biol.*, **8**, 947–956.

Giannopoulos,P. *et al.* (2009) Phosphorylation of prion protein at serine 43 induces prion protein conformational change. *J. Neurosci.*, **29**, 8743–8751.

Gong,C. *et al.* (2005) Post-translational modifications of tau protein in Alzheimer's disease. *J. Neural. Transm.*, **112**, 813–838.

Groban,E. *et al.* (2006) Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS Comput. Biol.*, **2**, e32.

Gsponer,J. *et al.* (2008) Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science*, **322**, 1365–1368.

Gu,W. and Roeder,R. (1997) Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell*, **90**, 595–606.

Gutteridge,A. and Thornton,J. (2005) Conformational changes observed in enzyme crystal structures upon substrate binding. *J. Mol. Biol.*, **346**, 21–28.

Hilser,V. (2010) An ensemble view of allostery. *Science*, **327**, 653–654.

Iakoucheva,L. *et al.* (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037–1049.

Johnson,L. and Lewis,R. (2001) Structural basis for control by phosphorylation. *Chem. Rev.*, **101**, 2209–2242.

Jungblut,P. *et al.* (2008) The speciation of the proteome. *Chem. Cent. J.*, **2**, 16.

Krueger,K. and Srivastava,S. (2006) Posttranslational protein modifications: current implications for cancer detection, prevention, and therapeutics. *Mol. Cell Proteomics.*, **5**, 1799–1810.

Lange,O. *et al.* (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, **320**, 1471–1475.

Latzer,J. *et al.* (2008) Conformational switching upon phosphorylation: a predictive framework based on energy landscape principles. *Biochemistry*, **47**, 2110–2122.

Lee,D. *et al.* (2008) Structure of Escherichia coli tyrosine kinase Etk reveals a novel activation mechanism. *EMBO J.*, **27**, 1758–1766.

Lee,N. and Koland,J. (2005) Conformational changes accompany phosphorylation of the epidermal growth factor receptor C-terminal domain. *Protein Sci.*, **14**, 2793–2803.

Li,S. *et al.* (2010) Loss of post-translational modification sites in disease. *Pac. Symp. Biocomput.*, 337–347.

Lodish,H. *et al.* (2004) *Molecular Cell Biology*. 5th edn. W.F. Freeman and Company, New York, NY, USA.

Lwin,T. and Luo,R. (2006) Force field influences in beta-hairpin folding simulations. *Protein Sci.*, **15**, 2642–2655.

Ma,B. and Nussinov,R. (2009) Regulating highly dynamic unstructured proteins and their coding mRNAs. *Genome Biol.*, **10**, 204.

Mann,M. and Jensen,O. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, **21**, 255–261.

McDonald,I. and Thornton,J. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.

Menet,J. and Rosbash,M. (2011) A new twist on clock protein phosphorylation: a conformational change leads to protein degradation. *Molecular Cell*, **43**, 695–697.

Mohan,A. *et al.* (2009) Influence of sequence changes and environment on intrinsically disordered proteins. *PLoS Comput. Biol.*, **5**, e1000497.

Narayanan,A. and Jacobson,M. (2009) Computational studies of protein regulation by post-translational phosphorylation. *Curr. Opin. Struct. Biol.*, **19**, 156–163.

Nishi,H. *et al.* (2011) Phosphorylation in protein-protein binding: effect on stability and function. *Structure*, **19**, 1807–1815.

Palaninathan,S. *et al.* (2008) Structural insight into pH-induced conformational changes within the native human transthyretin tetramer. *J. Mol. Biol.*, **382**, 1157–1167.

Peng,K. *et al.* (2004) Exploring bias in the Protein Data Bank using contrast classifiers. *Pac. Symp. Biocomput.*, 435–446.

Radivojac,P. *et al.* (2007) Intrinsic disorder and functional proteomics. *Biophys. J.*, **92**, 1439–1456.

Radivojac,P. *et al.* (2008) Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*, **24**, i241–i247.

Radivojac,P. *et al.* (2010) Identification, analysis, and prediction of protein ubiquitination sites. *Proteins*, **78**, 365–380.

Rhodes,G. (2006) *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. Academic Press, Burlington, MA, USA.

Schaller,M. and Parsons,J. (1995) pp125FAK-dependent tyrosine phosphorylation of paxillin creates a high-affinity binding site for Crk. *Mol. Cell. Biol.*, **15**, 2635–2645.

Shen,T. *et al.* (2005) The folding energy landscape and phosphorylation: modeling the conformational switch of the NFAT regulatory domain. *FASEB J.*, **19**, 1389–1395.

Stock,J. and Da Re,S. (2000) Signal transduction: response regulators on and off. *Curr. Biol.*, **10**, R420–R424.

Swain,J. and Gierasch,L. (2006) The changing landscape of protein allostery. *Curr. Opin. Struct. Biol.*, **16**, 102–108.

Thomas,M. *et al.* (2004) Androgen receptor acetylation site mutations cause trafficking defects, misfolding, and aggregation similar to expanded glutamine tracts. *J. Biol. Chem.*, **279**, 8389–8395.

Toh,K. *et al.* (2001) An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science*, **291**, 1040–1043.
Tsai,C. *et al.* (2008) Allostery: absence of a change in shape does not imply that allostery is not at play. *J. Mol. Biol.*, **378**, 1–11.
Uy,R. and Wold,F. (1977) Posttranslational covalent modification of proteins. *Science*, **198**, 890–896.
Van Eyk,J. (2011) Overview: the maturing of proteomics in cardiovascular research. *Circ. Res.*, **108**, 490–498.
Vidal,C. (2011) *Post-Translational Modifications in Health and Disease*. Protein Reviews. Springer, New York, NY, 1st edn.
Walsh,C. (2006) *Posttranslational Modification of Proteins: Expanding Nature's Inventory*. Roberts and Company Publishers, Englewood, CO.
Walsh,C. *et al.* (2005) Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem. Int. Ed. Engl.*, **44**, 7342–7372.
Wold,F. (1981) In vivo chemical modification of proteins (post-translational modification). *Annu. Rev. Biochem.*, **50**, 783–814.
Xie,H. *et al.* (2007) Functional anthology of intrinsic disorder. 3. ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J. Proteome Res.*, **6**, 1917–1932.
Yang,F. *et al.* (2010) Using affinity propagation combined post-processing to cluster protein sequences. *Protein Pept. Lett.*, **17**, 681–689.