ACS chemical biology

Reviews

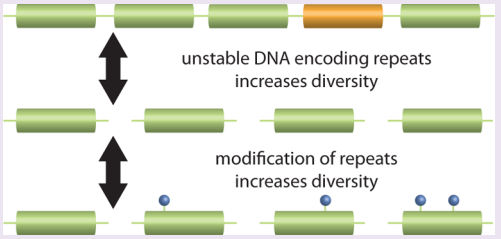
pubs.acs.org/acschemicalbiology

Chemically Modified Tandem Repeats in Proteins: Natural Combinatorial Peptide Libraries

Stephen M. Fuchs*

Department of Biology, Tufts University, 200 Boston Avenue, Medford, Massachusetts 02155, United States

ABSTRACT: Many proteins composed of tandem repeats (a linear motif, directly repeated within the sequence) are substrates for post-translational modifications (PTMs). Tandem repeats are also dynamic in number, presumably due to instability in the underlying DNA sequence. These observations lead to a hypothesis that cells use a combination of PTMs and variability in repeat number to mediate protein function. Evidence of these processes co-regulating diverse aspects of cellular function can be found in all organisms from bacteria to humans, suggesting a common but poorly described mechanism for regulating and diversifying protein function. This review highlights several examples whereby protein modifications and repetitive protein domains impart diversity. Lastly, it speculates on the possibility of using chemically modified repetitive amino acid sequences to develop peptide-based biomolecules with novel functions.



There is great interest in uncovering and subsequently adapting cellular processes that create proteins with novel function(s). While the 20 natural amino acids can presumably be linked in innumerable combinations to create proteins of diverse size, structure, and functionality, additional cellular mechanisms exist to impart further chemical and structural diversity. Utilizing these diversity-enhancing processes may be key to creating proteins with novel function, a strategy that has been successful for creating biomolecules with novel or enhanced function.[1,2]

The cellular mechanisms for diversifying proteins fit loosely into two classes: genetic and chemical mechanisms. Genetic mechanisms for imparting diversity range from altering the genetic code to encode for additional amino acids (e.g., pyrrolysine and selenocysteine) to splicing of the encoding mRNA to increase the number of individual polypeptides that can be encoded by a single message.[3] Recent advances in next-generation sequencing, coupled with thorough characterization of numerous genomes, have ignited interest in new sources of genomic diversity, in particular, the repetitive mini- and microsatellite regions of the genome.[4] While the majority of satellite DNA is associated with non-coding regions of the genome, many repetitive regions do lie within genes and code for proteins.[5] Unsurprisingly, repetitive DNA encodes for repetitive amino acids. This at first might seem to be the antithesis of diversity; however, these repetitive sequences may be among the most plastic regions of the genome,[6] as will be discussed below.

In addition to genetic mechanisms, cells use chemical modification of proteins to impart new function. Rather than change the encoding amino acid sequence, these modifications introduce new functionality, often in a spatially- and temporally defined manner.[7] Chemical modification of proteins ranges from influencing peptide bond isomerization to attaching high-molecular weight chains of carbohydrates or amino acids to a single amino acid side chain. Furthermore, modifications often occur in clusters on a given protein, greatly enhancing the number of chemically distinct structures that can exist in the cell.[8,9] Presumably, all of these disparate structures could have distinct cellular roles, thus creating immense functional diversity from a single amino acid sequence.

Many proteins that are repetitive are also substrates for diverse and numerous post-translational modifications. While most repetitive proteins are structural and/or extracellular in nature, many are involved in other aspects of cell function.[5] The diversity of modifications associated within repetitive proteins encompasses phosphorylation, methylation, glycosylation, hydroxylation, and even proline isomerization (see Table 1), among others. Taken together, these observations lead to a hypothesis that cells use a combination of PTMs and protein repeat variability to mediate protein function. Evidence of these processes co-regulating diverse aspects of cellular function can be found in all organisms from bacteria to humans. This suggests a common but poorly described mechanism for regulating and diversifying protein function. This review highlights several examples to describe mechanisms whereby protein modifications and repetitive protein domains impart diversity. Lastly, it speculates on the possibility of using chemically modified repetitive amino acid sequences to develop peptide-based biomolecules with novel functions.

CHEMICAL MODIFICATIONS OF PROTEINS IMPART DIVERSITY

One of the most common mechanisms for diversifying protein function is modification of the polypeptide chain itself. Many modifications, such as methylation, acetylation, or phosphorylation, change the chemical functionality on the side chains of

Received: September 23, 2012
Accepted: November 18, 2012
Published: November 18, 2012

   dx.doi.org/10.1021/cb3005066 | ACS Chem. Biol. 2013, 8, 275–282

Table 1. Modified Tandem Repeats in Proteins

protein	repeat consensus	associated PTMs	significance	ref
		Structural Proteins		
collagen	PPG	hydroxylation	structure stabilization	48
elastin	VGVAPG	hydroxylation	structure	76
DSPP	SS[DN], SD	phosphorylation	calcium coordination/ biomineralization	51
fibroin	SXSXSX	phosphorylation	water solubility	73
		Cell-Surface and Extracellular Proteins		
SRRPs	SAS[AEV]SAST	glycosylation	structure	77
mucins	HGVTSAPDTRPAPGSTAPPA, and others	glycosylation	immune recognition, tumor specific	78, 79
proteins with internal repeats (PIRs)	SQ[IV][STGNH]DGQ[LIV] Q[AIV][STA]	glycosylation	structure stabilization?	80
extensin (and other plant hydroxyproline-rich glyocproteins)	PSPPKHPYHYKSPPPPS	hydroxylation, glycosylation, tyrosine cross-linking	structure stabilization, cell wall assembly	81
trypanosome procyclic acidic repetitive proteins	GPEET	phosphorylation	protection	82, 83
		Nuclear Factors		
Rpb1	YSPTSPS	phosphorylation	protein recruitment, transcription	39
RNA-binding proteins (e.g., FMRP, Npl3, Gar1)	RGG	methylation	RNA binding	84, 85
yeast Spt5	S[TA]WGG[QA]	phosphorylation	protein recruitment, transcription elongation	86

amino acids via enzyme catalysis (known as post-translational modifications (PTMs). Additionally, enzymes such as the prolyl isomerases do not chemically alter the composition of amino acids but alter the *cis*/*trans* geometry of peptide bonds.[10] Not to be overlooked are numerous non-enzymatic, chemical modifications such as nitration/nitrosylation, oxidation, and deamidation, which also change the structure of amino acids, play important roles in regulating protein interactions and stability and also contribute to disease states.[11] (A thorough description of the many ways PTMs and chemical modifications of proteins alter protein structure and function can be found in ref 3).

Beyond the diversity in chemical structure that results from modifications, chemical modification of proteins can modulate their function in numerous ways (Figure 1A). For example phosphorylation, prenylation, and ubiquitination often dictate subcellular localization, allowing a single amino acid sequence to display differential function in varying cellular environments.[12,13] Similarly, modifications such as phosphorylation, hydroxylation, and acetylation induce conformational changes. In the case of enzymes, this often leads to activation or repression of catalytic activity. Perhaps the largest role for PTMs is in mediating a multitude of interactions between proteins and other biomacromolecules, including carbohydrates, nucleic acids, lipids, and even other proteins. Ubiquitination, phosphorylation, and proteolysis all regulate the cellular stability of proteins as well.[14,15] Consequently, PTMs including glycosylation, methylation, acetylation, and phosphorylation play integral roles in the regulation of all cellular processes, from gene expression and chromatin structure to cell−cell adhesion.[5,16]

Advances in proteomics indicate that PTMs do not act in isolation. Many proteins are heavily decorated with modifications with one estimate stating that as many as 1 in 10 amino acids within a protein may be modified.[17] The role of combinatorial PTMs is best studied in the case of histones where a large set of PTMs on the N-terminal tails of the four histone proteins dictate the recruitment of a host of proteins important for DNA-templated processes (replication, transcription, DNA repair) to chromatin.[18] However, combinations of modifications are also essential for the regulation of proteins such as p53, nuclear receptors, and cell surface factors.[9] In fact, more than 10 years ago, it was proposed that the diversity at the cell surface dictated by PTMs is a primary force dictating the difference between humans and apes.[19]

In total there are more than 400 known chemical modifications of proteins.[20] Protein modifications contribute to structural diversity at individual amino acids. They also impact all facets of protein structure and function and thus represent a vital mechanism for cells to expand upon the functionality of its proteins.

SHORT TANDEM REPEATS IN PROTEINS ARE VARIABLE

As many as one-fifth of all proteins contain regions of repetitive amino acids within their sequence.[5] Repetitive proteins are involved in all cellular processes, although they are most commonly extracellular proteins or proteins involved in maintaining cellular structure or cell−cell interactions. Tandem repetitive domains are most often thought of as simple homopolymeric runs of amino acids, such as the polyglutamine-containing proteins closely associated with neurodegenerative disorders such as Huntington's disease and spinocerebellar ataxias.[21] However, as many as 20% of all proteins contain a stretch of sequence that consists of a perfect or imperfect (containing substitutions) tandemly repeated motif of 2 to >50 amino acids.[5,22]

The biology of repetitive domains has gained notoriety as the length and variability of homopolymeric repeats, such as polyglutamines, is often linked to the pathologic severity.[23] Repetitive protein variability is likely derived from instability within the DNA sequence that codes for the amino acids. Often referred to as microsatellite DNA (2−10 bp) or minisatellite DNA (∼10−100 bp), repetitive DNA sequences are prone to unusual secondary structure formation, replication errors and double-strand breakage.[24,25] These genomic insults trigger DNA repair processes that are often mutagenic, leading to variability in

dx.doi.org/10.1021/cb3005066 | ACS Chem. Biol. 2013, 8, 275−282

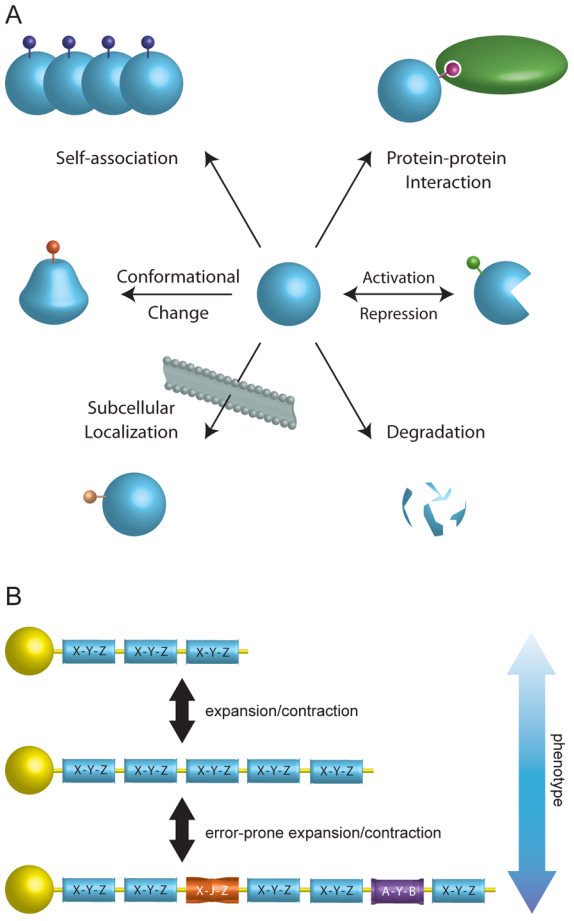


Figure 1. Protein diversity is driven by post-translational modifications and repetitive domains. (A) Post-translational modifications regulate all aspects of protein function. The macromolecular interactions, conformation, subcellular localization, stability, and activity of a protein (blue) can all be modulated through attachment of a chemical modification (small spheres) to a side chain of the protein. (B) Instability within the DNA encoding repetitive amino acid sequences results in expansion and contraction within the number of repeats. Expansion/contraction can also be mutagenic, giving rise to repeats with altered sequence. Variability in repeat length and sequence can give rise to a range of phenotypes.

repeat number as well as high mutational frequencies in regions surrounding the repeat (Figure 1B).[5,26] In fact, Lobachev and co-workers suggest repetitive regions have mutation frequencies approaching 10^{-4} in wild-type cells.[27] While it might seem that highly mutagenic repetitive DNA sequences would be deleterious to the genome, it is clear that both intergenic and protein-coding repetitive sequences play important roles within the cell. Variability in non-coding DNA repeats can tune the expression of genes and has been discussed in great detail elsewhere.[5,26] Likewise, protein-coding repeats are plastic, presumably to modulate function. For example, the repetitive domain of FLO1 facilitates interactions with other cells expressing FLO1, resulting in yeast aggregates (similar to biofilm formation but also important in the wine and beer making industries). Verstrepen and Fink demonstrated that the extent of flocculation is mediated by the number of repeats present in FLO1.[28] Expansion of repeat number within FLO1 results in a multivalent effect, where a greater number of repeats results in synthesis of more binding domains, leading to a tighter interaction between cells. Multivalency is an important concept in many areas of biology[29] and may be a driving mechanism behind the functional significance of certain repeat expansions and contractions.

MODIFICATIONS AND REPETITIVE REGIONS SYNERGIZE TO IMPART BIOLOGICAL COMPLEXITY

A striking feature of many repetitive protein domains is that they are also rich in post-translational modifications (see Table 1). To clarify, it is well-known that many modifying enzymes recognize short amino acid motifs within proteins. While many motifs may be found multiple times in a particular protein, such as the notable ARKS motif in histones (which are methylated, acetylated, and phosphorylated),[30] these motifs would not necessarily constitute a repetitive domain as defined here. Rather the propensity to be influenced by underlying DNA instability within these regions, in combination with modifications, imparts the great potential diversity or complexity into these simple protein sequences. For example, a seven amino acid repeat that could be modified at two locations by a single post-translational modification would have four discrete chemical structures (see Figure 2B). If this sequence were also tandemly repeated, the number of possible chemical structures is 4^n, where n is the number of repeats. Thus, the biological complexity that can be achieved by a simple amino acid sequence can be astounding, as is demonstrated by the examples below.

The C-Terminal Domain Repeat of Eukaryotic RNA Polymerase II. RNA polymerase II (RNAPII), the enzyme primarily responsible for mRNA synthesis in eukaryotic cells, possesses a long stretch of tandemly repeating amino acids at its C-terminus (Figure 2). This C-terminal domain (CTD) is composed of ~20−52 highly conserved repeats of a seven amino acid sequence, YSPTSPS. The CTD is a target for diverse modifications including phosphorylation at 5 residues (Y1, S2, T4, S5, S7),[16,31,32] O-GlcNAcylation at S5 and S7,[33,34] proline isomerization at the amide bond preceding P6,[35] and even lysine methylation[36] and ubiquitination[37] in some degenerate repeats within the human and murine CTD.

How did such a complicated combination of PTMs and repeats come into existence and why would it be evolutionarily maintained? It is plausible to imagine an ancient RNAPII with a short or nonexistent CTD. As organisms evolved and became more complex, and needed to accommodate more protein factors on the CTD, the instability within the CTD allowed the cell to adapt − building a longer CTD, capable of binding more proteins. Indeed, Stiller and co-workers have examined the CTDs of diverse eukaryotes and there is a general trend with longer CTDs being associated with more complex organisms.[38] Additionally, early mutagenesis studies on the CTD in yeast revealed that only 8 of 26 repeats are essential for growth, however, these strains acquire spontaneous mutants with longer CTD domains.[39]

As described above, one of the primary functions of PTMs is to modulate protein−protein interactions. In the case of the CTD, phosphorylation (mainly of serine) within the repeat increases the functionality of the CTD by creating additional binding platforms (Figure 2A). Indeed it is well established that different RNAPII-associating factors bind to the CTD in different phosphorylated forms.[16,40] Several kinases act on the CTD during different phases of transcription (e.g., initiation, elongation, termination). As such, phosphorylation of the CTD imparts temporal control on the binding of factors. For example, factors that bind Ser5 phosphorylated-repeats tend to associate early in the transcription cycle, factors involved in

dx.doi.org/10.1021/cb3005066 | ACS Chem. Biol. 2013, 8, 275−282

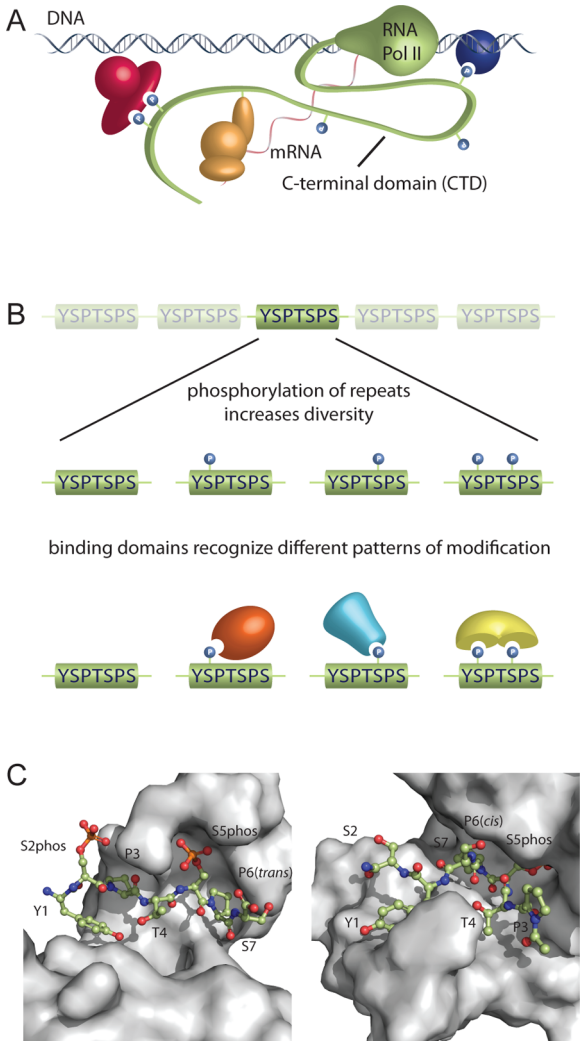


Figure 2. Modifications within the C-terminal domain (CTD) repeat of RNA polymerase II. (A) Model of RNA polymerase on DNA with differentially modified repeats within the CTD recruiting different protein factors. (B) Phosphorylation is most prevalent at Ser2 and Ser5 within the CTD repeat. A single repeat therefore can have at four discrete structures, which can be differentially recognized by protein domains. (C) Proline within the CTD repeat can adopt a *cis*- or *trans*-peptide bond, which results in two distinct structures that can be differentially recognized. Ser5 phosphorylation is recognized differently by Pin1 (left) and Ssu72 (right) depending on the conformation at Pro6.[42,43]

transcription elongation recognize repeats phosphorylated at both Ser2 and Ser5, and mRNA processing and transcription termination factors associate with the CTD in its Ser2-phosphorylated form.[40,41] In addition, proline isomerization plays an important role in regulating repeat binding. As show in Figure 2C, recognition of Ser5 phopshorylation by two different proteins, Ssu72 and Pin1, is determined by the *cis*/*trans* conformation of proline at position 6 within the repeat.[42,43] Within the past few years *cis*-proline was first demonstrated to facilitate the binding of the phosphatase Ssu72, and new modifications at Tyr1, Thr4, and Ser7 were uncovered (reviewed in ref 41), suggesting there are still many unanswered questions regarding the complicated regulation of the RNAPII CTD. However, it does appear that the CTD (at least in yeast) is under two levels of regulation: one at the level of DNA, to control the number of repeats, and one at the post-translational level, modifying the CTD to accommodate numerous protein factors with diverse functions with both spatial and temporal precision.

Collagen Structure. Collagen is the most abundant protein in mammals, making up more than 25% of total protein.[44] Collagen acts as the main component of connective tissue, and this is in large part because of its long fibrillar structure. This structure is derived from the ability of individual collagen peptides to form polyproline helices, which interact with each other to form a stable triple-helical structure.[45] There are diverse classes of collagens, but the consensus motif is a repeating unit of the tripeptide X_{aa}-Y_{aa}-Gly where the X_{aa} position is most commonly proline and the Y_{aa} position often includes the modified amino acid (2*S*,4*R*)-4-hydroxyproline (Hyp), which is catalyzed by the enzyme prolyl-4-hydroxylase.[46]

Like the RNAPII CTD, hydroxylation of collagen has been shown to be important for the binding of certain protein factors (Glycoprotein VI).[47] However the primary role of hydroxyproline in collagen is to modulate the structure of collagen itself. As shown in Figure 3, individual collagen polypeptides form

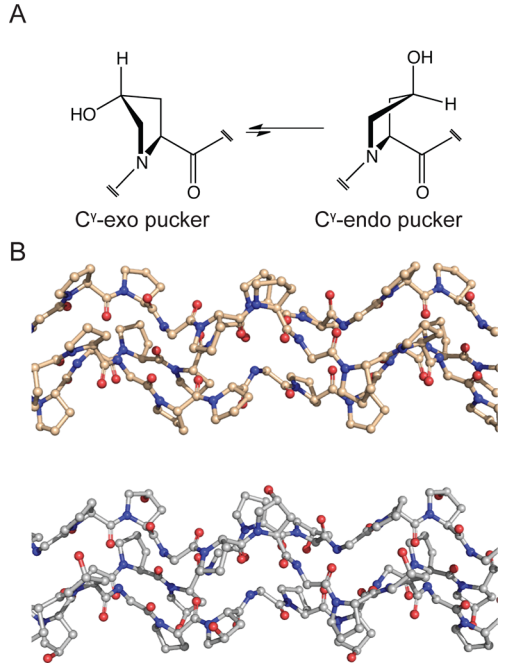


Figure 3. Collagen structures is stabilized by proline hydroxylation. (A) Hydroxyproline (Hyp) favors a C^{γ}-exo pucker, which stabilizes triple-helical collagen when in the Y_{aa} position of a X_{aa}-Y_{aa}-Gly repeat sequence. (B) Structure of triple helical Pro-Pro-Gly repeats (top) and Pro-Hyp-Gly repeats (bottom).[87]

polyproline type II helices and assemble with two other collagen polypeptides to form a left-handed triple helical structure.[45] Numerous studies have reported that hydroxyproline in the Y_{aa} position, but not the X_{aa} position, increases the stability of the collagen triple helix. While many factors play into the overall stabilization of collagen structure, the ability of individual strands to form polyproline helices is related to stability. The pyrrolidine ring of proline can adopt two conformations, a C^{γ}-endo or C^{γ}-exo conformation (Figure 3A). Polyproline helices are stabilized with X_{aa} in a C^{γ}-endo conformation and Y_{aa} in a C^{γ}-exo conformation.[48] Hydroxyproline greatly prefers the C^{γ}-exo conformation, whereas proline has an approximately 2:1 prerefence for the C-endo conformation. Thus, model peptides

dx.doi.org/10.1021/cb3005066 | *ACS Chem. Biol.* 2013, 8, 275–282

with Pro-Hyp-Gly repeats form considerably more stable triple helices than Pro-Pro-Gly repeats of equal length (Figure 3B).

Organisms use hydroxyproline to "tune" the stability of collagen depending on its given function. For example, collagen associated with cold-water fish has ~50 hydroxyproline residues per 1000 residues, whereas collagen from their warm-water counterparts has ~80 hydroxyproline residues. This less than 2-fold change in hydroxyproline content nonetheless results in a >20 °C increase in thermostability.[49] Similarly, hydroxylysine, another modification common within collagen in bone, is important for intramolecular cross-linking between tropocollagen strands.[50] Thus, cells use hydroxyproline density within collagen repeats to dictate the function of individual collagen fibrils.

Dentin Sialophosphoprotein. Many repetitive proteins are extremely rich in serine residues. While phosphorylation of the RNAPII CTD mediates the binding of protein factors, serine phosphorylation can also mediate interactions with other molecules. For example, the protein dentin sialophosphoprotein (DSPP) is the most abundant noncollagen protein in teeth and is important for complexing inorganic calcium phosphate. DSPP has two repeating units, Ser-Ser-Asp/Asn, and Ser-Asp where approximately 90% of all serine resides within these sequences are phosphorylated.[51] George and co-workers have shown that this phosphorylation is necessary for proper mineralization.[52] Mutations within the repetitive DSPP gene are common and lead to a number of dentin diseases.[53,54] George and co-workers have proposed that the SSD repeat forms an extended structure where phosphoserine pairs would occupy both faces of a plane (see Figure 4). Presumably, these phosphate pairs act in concert to nucleate calcium phosphate mineralization. Similar molecules are thought to be important for bone mineralization and calcite formation, and as such, there is great interest in creating artificial molecules that can aid in the regrowth of these biomaterials.[55,56]

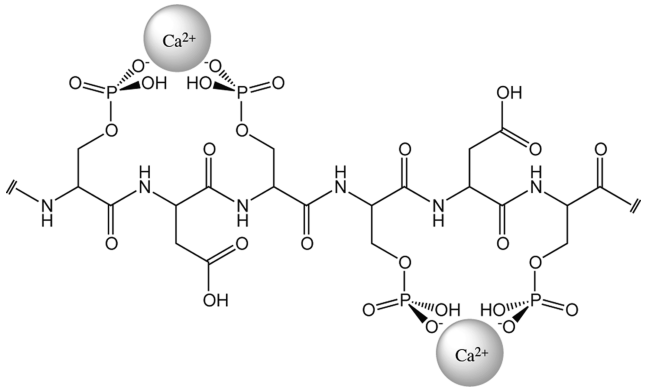


Figure 4. Model of the Ser-Ser-Asp repeat of dentin sialophosphoprotein (DSPP). More than 90% of all serine residues within DSPP are phosphorylated facilitating interactions between the protein and calcium ions presumably to drive dentin formation.[52]

FUTURE PROSPECTS

Adapting existing cellular mechanisms has been a very successful strategy for the design or selection of proteins with novel functions. Perhaps the high variability of repetitive proteins can be exploited to produce novel molecules. Furthermore, can the proclivity for chemical modification seen in repetitive sequences be utilized to even further diversify the types of peptide-based structures we can produce in cells? This may be of interest for the development of peptide-based biomaterials with a multitude of uses.

Repetitive proteins offer many potential advantages for protein engineering. The repetitiveness of the coding DNA not only provides a facile way to make extended repeats[57] but also has the advantage of being unstable and mutagenic *in vivo*.[27] Through natural cellular processes, repetitive DNA sequences will break, undergo expansion and contraction, and acquire substitutions through DNA repair (Figure 2).[23−26] This plasticity at the DNA with regard to coding sequence and repeat number would result in a highly variable amino acid sequence as well. Recent results aimed at identifying the factors responsible for repeat expansion and contraction may further provide leads in the development of designer bacterial or eukaryotic strains with enhanced mutagenic frequency.[27]

As shown for the examples above, the addition of a single PTM within a repeat sequence greatly increases the number of discrete species that exist within the cell. Multiplying this effect by either having numerous repeats or more than one PTM results in tremendous diversity. Many groups have shown it is possible to chemically synthesize peptide libraries with large numbers of modified side chains, but this process is both slow and expensive.[58,59] Is it possible to harness the power of modifying enzymes to enhance the chemical diversity within libraries of repetitive peptides *in vivo*? To do so would require an understanding of the chemical and structural parameters that control enzyme activity. For example, the cyclin-dependent kinases recognize Ser/Thr-Pro motifs, whereas the ATM/ATR kinases phosphorylate Ser/Thr-Gln motifs.[60−62] Similarly, the Y_{aa} position within the collagen repeat is a strong substrate for the enzyme prolyl-4-hydroxylase when proline is also present in the X_{aa} position.[48] However, while there are several well-described and conserved examples of substrate-consensus sequences, it remains difficult to predict sites of modifications within proteins on the basis of sequence alone. Classical studies of protease substrate specificity suggest that engineering enzymes with altered specificity is possible. Mutagenesis and selection has enabled the design of enzymes with altered promiscuity or specificity.[63,64] Strategies such as scanning peptide arrays developed by Turk make it possible to quickly screen and potentially evolve enzymes with custom specificity,[65] although relatively few examples of this are apparent in the literature.[66]

There have been several recent attempts to produce repetitive biomaterials. Many groups have tried to use the DSPP-like repeats to create peptides that will nucleate biomineralization.[67,68] Many groups have also been working to create synthetic collagen, elastin, or silks using bioengineering approaches.[69−71] These studies have not made use of the inherent variability of these repetitive regions; rather, the focus has been on better methods to clone and express repetitive DNA coding sequences.[57] A few attempts have been made to incorporate the diversity provided by protein modifications into repetitive peptides. Kaplan and co-workers utilized phosphorylation of spider silk repeats to modulate water solubility.[72] This approach is used in nature by the aquatic caddisfly, which has evolved a H-fibroin silk protein repeat (SXSXSX) that has diverged from that of related insects (GXGXGX). This adaptation presumably evolved to help the caddisfly produce silk underwater.[73] Similarly, other groups have engineered bacteria and yeast to promote hydroxyproline formation in recombinant sources.[74,75]

Chemically modification is a widely used natural mechanism for imparting diversity into protein sequences. As more is learned about the PTM enzyme specificity and DNA instability that

dx.doi.org/10.1021/cb3005066 | ACS Chem. Biol. 2013, 8, 275−282

underlie the variability found in these domains, we will be able to exploit these cellular mechanisms to develop better *in vitro* and *in vivo* systems for producing molecules with novel function.

AUTHOR INFORMATION

Corresponding Author

*E-mail: stephen.fuchs@tufts.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The author would like to thank R. White, F. Kotch, and G. Dixit for many helpful discussions and critical reading of the manuscript.

REFERENCES

(1) Day, R. N., and Davidson, M. W. (2009) The fluorescent protein palette: tools for cellular imaging. *Chem. Soc. Rev. 38*, 2887−2921.

(2) Felnagle, E. A., Chaubey, A., Noey, E. L., Houk, K. N., and Liao, J. C. (2012) Engineering synthetic recursive pathways to generate non-natural small molecules. *Nat. Chem. Biol. 8*, 518−526.

(3) Walsh, C. (2006) *Posttranslational Modification of Proteins: Expanding Nature's Inventory*, Roberts and Co. Publishers, Englewood, CO.

(4) Bennett, P. (2000) Demystified ... microsatellites. *Mol. Pathol. 53*, 177−183.

(5) Gemayel, R., Vinces, M. D., Legendre, M., and Verstrepen, K. J. (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet. 44*, 445−477.

(6) Plohl, M., Luchetti, A., Mestrovic, N., and Mantovani, B. (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)-chromatin. *Gene 409*, 72−82.

(7) Trost, M., Bridon, G., Desjardins, M., and Thibault, P. (2010) Subcellular phosphoproteomics. *Mass Spectrom. Rev. 29*, 962−990.

(8) Rando, O. J. (2012) Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr. Opin. Genet. Dev. 22*, 148−155.

(9) Sims, R. J., 3rd, and Reinberg, D. (2008) Is there a code embedded in proteins that is based on post-translational modifications? *Nat. Rev. Mol. Cell. Biol. 9*, 815−820.

(10) Lu, K. P., Finn, G., Lee, T. H., and Nicholson, L. K. (2007) Prolyl cis-trans isomerization as a molecular timer. *Nat. Chem. Biol. 3*, 619−629.

(11) Cloos, P. A., and Christgau, S. (2002) Non-enzymatic covalent modifications of proteins: mechanisms, physiological consequences and clinical applications. *Matrix Biol. 21*, 39−52.

(12) Chen, C. A., and Manning, D. R. (2001) Regulation of G proteins by covalent modification. *Oncogene 20*, 1643−1652.

(13) MacGurn, J. A., Hsu, P. C., and Emr, S. D. (2012) Ubiquitin and membrane protein turnover: from cradle to grave. *Annu. Rev. Biochem. 81*, 231−259.

(14) Hunter, T. (2007) The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Mol. Cell 28*, 730−738.

(15) Komander, D. (2009) The emerging complexity of protein ubiquitination. *Biochem. Soc. Trans. 37*, 937−953.

(16) Fuchs, S. M., Laribee, R. N., and Strahl, B. D. (2009) Protein modifications in transcription elongation. *Biochim. Biophys. Acta 1789*, 26−36.

(17) Nielsen, M. L., Savitski, M. M., and Zubarev, R. A. (2006) Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol. Cell. Proteomics 5*, 2384−2391.

(18) Kouzarides, T. (2007) Chromatin modifications and their function. *Cell 128*, 693−705.

(19) Muchmore, E. A., Diaz, S., and Varki, A. (1998) A structural difference between the cell surfaces of humans and the great apes. *Am. J. Phys. Anthropol. 107*, 187−198.

(20) Khoury, G. A., Baliban, R. C., and Floudas, C. A. (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.*, DOI: 10.1038/srep00090.

(21) Hannan, A. J. (2010) TRPing up the genome: Tandem repeat polymorphisms as dynamic sources of genetic variability in health and disease. *Discov. Med. 10*, 314−321.

(22) Newman, A. M., and Cooper, J. B. (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinform. 8*, 382.

(23) Mirkin, S. M. (2007) Expandable DNA repeats and human disease. *Nature 447*, 932−940.

(24) Voineagu, I., Freudenreich, C. H., and Mirkin, S. M. (2009) Checkpoint responses to unusual structures formed by DNA repeats. *Mol. Carcinog. 48*, 309−318.

(25) Mirkin, S. M. (2006) DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. Struct. Biol. 16*, 351−358.

(26) Verstrepen, K. J., and Fink, G. R. (2009) Genetic and epigenetic mechanisms underlying cell-surface variability in protozoa and fungi. *Annu. Rev. Genet. 43*, 1−24.

(27) Zhang, Y., Shishkin, A. A., Nishida, Y., Marcinkowski-Desmond, D., Saini, N., Volkov, K. V., Mirkin, S. M., and Lobachev, K. S. (2012) Genome-wide screen identifies pathways that govern GAA/TTC repeat fragility and expansions in dividing and nondividing yeast cells. *Mol. Cell 48*, 254−265.

(28) Verstrepen, K. J., Jansen, A., Lewitter, F., and Fink, G. R. (2005) Intragenic tandem repeats generate functional variability. *Nat. Genet. 37*, 986−990.

(29) Fasting, C., Schalley, C. A., Weber, M., Seitz, O., Hecht, S., Koksch, B., Dernedde, J., Graf, C., Knapp, E. W., and Haag, R. (2012) Multivalency as a chemical organization and action principle. *Angew. Chem., Int. Ed. 42*, 10472−10498.

(30) Fischle, W., Franz, H., Jacobs, S. A., Allis, C. D., and Khorasanizadeh, S. (2008) Specificity of the chromodomain Y chromosome family of chromodomains for lysine-methylated ARK(S/T) motifs. *J. Biol. Chem. 283*, 19626−19635.

(31) Hintermair, C., Heidemann, M., Koch, F., Descostes, N., Gut, M., Gut, I., Fenouil, R., Ferrier, P., Flatley, A., Kremmer, E., Chapman, R. D., Andrau, J. C., and Eick, D. (2012) Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. *EMBO J. 31*, 2784−2797.

(32) Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., and Cramer, P. (2012) CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science 336*, 1723−1725.

(33) Kelly, W. G., Dahmus, M. E., and Hart, G. W. (1993) RNA polymerase II is a glycoprotein. Modification of the COOH-terminal domain by O-GlcNAc. *J. Biol. Chem. 268*, 10416−10424.

(34) Ranuncolo, S. M., Ghosh, S., Hanover, J. A., Hart, G. W., and Lewis, B. A. (2012) Evidence of the involvement of O-GlcNAc-modified human RNA polymerase II CTD in transcription in vitro and in vivo. *J. Biol. Chem. 287*, 23549−23561.

(35) Wu, X., Wilcox, C. B., Devasahayam, G., Hackett, R. L., Arevalo-Rodriguez, M., Cardenas, M. E., Heitman, J., and Hanes, S. D. (2000) The Ess1 prolyl isomerase is linked to chromatin remodeling complexes and the general transcription machinery. *EMBO J. 19*, 3727−3738.

(36) Sims, R. J., 3rd, Rojas, L. A., Beck, D., Bonasio, R., Schuller, R., Drury, W. J., 3rd, Eick, D., and Reinberg, D. (2011) The C-terminal domain of RNA polymerase II is modified by site-specific methylation. *Science 332*, 99−103.

(37) Li, H., Zhang, Z., Wang, B., Zhang, J., Zhao, Y., and Jin, Y. (2007) Wwp2-mediated ubiquitination of the RNA polymerase II large subunit in mouse embryonic pluripotent stem cells. *Mol. Cell. Biol. 27*, 5296−5305.

(38) Liu, P., Kenney, J. M., Stiller, J. W., and Greenleaf, A. L. (2010) Genetic organization, length conservation, and evolution of RNA polymerase II carboxyl-terminal domain. *Mol. Biol. Evol. 27*, 2628−2641.
(39) Nonet, M. L., and Young, R. A. (1989) Intragenic and extragenic suppressors of mutations in the heptapeptide repeat domain of Saccharomyces cerevisiae RNA polymerase II. *Genetics 123*, 715−724.
(40) Buratowski, S. (2009) Progression through the RNA polymerase II CTD cycle. *Mol. Cell 36*, 541−546.
(41) Heidemann, M., Hintermair, C., Voss, K., and Eick, D. (2012) Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochim. Biophys. Acta*, DOI: 10.1016/j.bbagrm.2012.08.013.
(42) Werner-Allen, J. W., Lee, C. J., Liu, P., Nicely, N. I., Wang, S., Greenleaf, A. L., and Zhou, P. (2011) cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72. *J. Biol. Chem. 286*, 5717−5726.
(43) Verdecia, M. A., Bowman, M. E., Lu, K. P., Hunter, T., and Noel, J. P. (2000) Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nat. Struct. Biol. 7*, 639−643.
(44) Di Lullo, G. A., Sweeney, S. M., Korkko, J., Ala-Kokko, L., and San Antonio, J. D. (2002) Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen. *J. Biol. Chem. 277*, 4223−4231.
(45) Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. *Science 266*, 75−81.
(46) Kivirikko, K. I., Suga, K., Kishida, Y., Sakakibara, S., and Prockop, D. J. (1971) Asymmetry in the hydroxylation of (Pro-Pro-Gly) 5 by protocollagen proline hydroxylase. *Biochem. Biophys. Res. Commun. 45*, 1591−1596.
(47) Farndale, R. W., Lisman, T., Bihan, D., Hamaia, S., Smerling, C. S., Pugh, N., Konitsiotis, A., Leitinger, B., de Groot, P. G., Jarvis, G. E., and Raynal, N. (2008) Cell-collagen interactions: the use of peptide Toolkits to investigate collagen-receptor interactions. *Biochem. Soc. Trans. 36*, 241−250.
(48) Shoulders, M. D., and Raines, R. T. (2009) Collagen structure and stability. *Annu. Rev. Biochem. 78*, 929−958.
(49) Privalov, P. L. (1982) Stability of proteins. Proteins which do not present a single cooperative system. *Adv. Prot. Chem. 35*, 1−104.
(50) Knott, L., and Bailey, A. J. (1998) Collagen cross-links in mineralizing tissues: a review of their chemistry, function, and clinical relevance. *Bone 22*, 181−187.
(51) George, A., Bannon, L., Sabsay, B., Dillon, J. W., Malone, J., Veis, A., Jenkins, N. A., Gilbert, D. J., and Copeland, N. G. (1996) The carboxyl-terminal domain of phosphophoryn contains unique extended triplet amino acid repeat sequences forming ordered carboxyl-phosphate interaction ridges that may be essential in the biomineralization process. *J. Biol. Chem. 271*, 32869−32873.
(52) He, G., Ramachandran, A., Dahl, T., George, S., Schultz, D., Cookson, D., Veis, A., and George, A. (2005) Phosphorylation of phosphophoryn is crucial for its function as a mediator of biomineralization. *J. Biol. Chem. 280*, 33109−33114.
(53) Nieminen, P., Papagiannoulis-Lascarides, L., Waltimo-Siren, J., Ollila, P., Karjalainen, S., Arte, S., Veerkamp, J., Tallon Walton, V., Chimenos Kustner, E., Siltanen, T., Holappa, H., Lukinmaa, P. L., and Alaluusua, S. (2011) Frameshift mutations in dentin phosphoprotein and dependence of dentin disease phenotype on mutation location. *J. Bone Miner. Res. 26*, 873−880.
(54) McKnight, D. A., Suzanne Hart, P., Hart, T. C., Hartsfield, J. K., Wilson, A., Wright, J. T., and Fisher, L. W. (2008) A comprehensive analysis of normal variation and disease-causing mutations in the human DSPP gene. *Hum. Mutat. 29*, 1392−1404.
(55) Grohe, B., O'Young, J., Ionescu, D. A., Lajoie, G., Rogers, K. A., Karttunen, M., Goldberg, H. A., and Hunter, G. K. (2007) Control of calcium oxalate crystal growth by face-specific adsorption of an osteopontin phosphopeptide. *J. Am. Chem. Soc. 129*, 14946−14951.
(56) Prasad, M., Butler, W. T., and Qin, C. (2010) Dentin sialophosphoprotein in biomineralization. *Connect. Tissue Res. 51*, 404−417.
(57) McDaniel, J. R., Mackay, J. A., Quiroz, F. G., and Chilkoti, A. (2010) Recursive directional ligation by plasmid reconstruction allows rapid and seamless cloning of oligomeric genes. *Biomacromolecules 11*, 944−952.
(58) Bock, I., Dhayalan, A., Kudithipudi, S., Brandt, O., Rathert, P., and Jeltsch, A. (2011) Detailed specificity analysis of antibodies binding to modified histone tails with peptide arrays. *Epigenetics 6*, 256−263.
(59) Fuchs, S. M., Krajewski, K., Baker, R. W., Miller, V. L., and Strahl, B. D. (2011) Influence of combinatorial histone modifications on antibody and effector protein recognition. *Curr. Biol. 21*, 53−58.
(60) Kim, S. T., Lim, D. S., Canman, C. E., and Kastan, M. B. (1999) Substrate specificities and identification of putative substrates of ATM kinase family members. *J. Biol. Chem. 274*, 37538−37543.
(61) O'Neill, T., Dwyer, A. J., Ziv, Y., Chan, D. W., Lees-Miller, S. P., Abraham, R. H., Lai, J. H., Hill, D., Shiloh, Y., Cantley, L. C., and Rathbun, G. A. (2000) Utilization of oriented peptide libraries to identify substrate motifs selected by ATM. *J. Biol. Chem. 275*, 22719−22727.
(62) Traven, A., and Heierhorst, J. (2005) SQ/TQ cluster domains: concentrated ATM/ATR kinase phosphorylation site regions in DNA-damage-response proteins. *BioEssays 27*, 397−407.
(63) Verhoeven, K. D., Altstadt, O. C., and Savinov, S. N. (2012) Intracellular detection and evolution of site-specific proteases using a genetic selection system. *Appl. Biochem. Biotechnol. 166*, 1340−1354.
(64) Yoo, T. H., Pogson, M., Iverson, B. L., and Georgiou, G. (2012) Directed evolution of highly selective proteases by using a novel FACS-based screen that capitalizes on the p53 regulator MDM2. *ChemBioChem 13*, 649−653.
(65) Turk, B. E., and Cantley, L. C. (2003) Peptide libraries: at the crossroads of proteomics and bioinformatics. *Curr. Opin. Chem. Biol. 7*, 84−90.
(66) Chen, C., and Turk, B. E. (2010) Analysis of serine-threonine kinase specificity using arrayed positional scanning peptide libraries. *Curr. Protoc. Mol. Biol.*, DOI: 10.1002/0471142727.mb1814s91.
(67) Choi, Y. S., Lee, J. Y., Suh, J. S., Lee, G., Chung, C. P., and Park, Y. J. (2012) The mineralization inducing peptide derived from dentin sialophosphoprotein for bone regeneration. *J. Biomed. Mater. Res. A.*, DOI: 10.1002/jbm.a.34352.
(68) Goldberg, H. A., Warner, K. J., Li, M. C., and Hunter, G. K. (2001) Binding of bone sialoprotein, osteopontin and synthetic polypeptides to hydroxyapatite. *Connect. Tissue Res. 42*, 25−37.
(69) Gauba, V., and Hartgerink, J. D. (2008) Synthetic collagen heterotrimers: structural mimics of wild-type and mutant collagen type I. *J. Am. Chem. Soc. 130*, 7509−7515.
(70) Amruthwar, S. S., and Janorkar, A. V. (2012) Preparation and characterization of elastin-like polypeptide scaffolds for local delivery of antibiotics and proteins. *J. Mater. Sci. Mater. Med.*, DOI: 10.1007/s10856-012-4749-5.
(71) Frandsen, J. L., and Ghandehari, H. (2012) Recombinant protein-based polymers for advanced drug delivery. *Chem. Soc. Rev. 41*, 2696−2706.
(72) Winkler, S., Wilson, D., and Kaplan, D. L. (2000) Controlling beta-sheet assembly in genetically engineered silk by enzymatic phosphorylation/dephosphorylation. *Biochemistry 39*, 12739−12746.
(73) Stewart, R. J., and Wang, C. S. (2010) Adaptation of caddisfly larval silks to aquatic habitats by phosphorylation of h-fibroin serines. *Biomacromolecules 11*, 969−974.
(74) Vaughn, P. R., Galanis, M., Richards, K. M., Tebb, T. A., Ramshaw, J. A., and Werkmeister, J. A. (1998) Production of recombinant hydroxylated human type III collagen fragment in Saccharomyces cerevisiae. *DNA Cell Biol. 17*, 511−518.
(75) Pinkas, D. M., Ding, S., Raines, R. T., and Barron, A. E. (2011) Tunable, post-translational hydroxylation of collagen domains in Escherichia coli. *ACS Chem. Biol. 6*, 320−324.
(76) Getie, M., Schmelzer, C. E., and Neubert, R. H. (2005) Characterization of peptides resulting from digestion of human skin elastin with elastase. *Proteins 61*, 649−657.

(77) Lizcano, A., Sanchez, C. J., and Orihuela, C. J. (2012) A role for glycosylated serine-rich repeat proteins in gram-positive bacterial pathogenesis. *Mol. Oral Microbiol. 27*, 257−269.
(78) Tarp, M. A., Sorensen, A. L., Mandel, U., Paulsen, H., Burchell, J., Taylor-Papadimitriou, J., and Clausen, H. (2007) Identification of a novel cancer-specific immunodominant glycopeptide epitope in the MUC1 tandem repeat. *Glycobiology 17*, 197−209.
(79) Irimura, T., Denda, K., Iida, S., Takeuchi, H., and Kato, K. (1999) Diverse glycosylation of MUC1 and MUC2: potential significance in tumor immunity. *J. Biochem. 126*, 975−985.
(80) Toh-e, A., Yasunaga, S., Nisogi, H., Tanaka, K., Oguchi, T., and Matsui, Y. (1993) Three yeast genes, PIR1, PIR2 and PIR3, containing internal tandem repeats, are related to each other, and PIR1 and PIR2 are required for tolerance to heat shock. *Yeast 9*, 481−494.
(81) Kieliszewski, M. J., and Lamport, D. T. (1994) Extensin: repetitive motifs, functional sites, post-translational codes, and phylogeny. *Plant J. 5*, 157−172.
(82) Acosta-Serrano, A., Vassella, E., Liniger, M., Kunz Renggli, C., Brun, R., Roditi, I., and Englund, P. T. (2001) The surface coat of procyclic Trypanosoma brucei: programmed expression and proteolytic cleavage of procyclin in the tsetse fly. *Proc. Natl. Acad. Sci. U.S.A. 98*, 1513−1518.
(83) Mehlert, A., Treumann, A., and Ferguson, M. A. (1999) Trypanosoma brucei GPEET-PARP is phosphorylated on six out of seven threonine residues. *Mol. Biochem. Parasitol. 98*, 291−296.
(84) McBride, A. E., and Silver, P. A. (2001) State of the arg: protein methylation at arginine comes of age. *Cell 106*, 5−8.
(85) Blackwell, E., and Ceman, S. (2012) Arginine methylation of RNA-binding proteins regulates cell function and differentiation. *Mol. Reprod. Dev. 79*, 163−175.
(86) Zhou, K., Kuo, W. H., Fillingham, J., and Greenblatt, J. F. (2009) Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5. *Proc. Natl. Acad. Sci. U.S.A. 106*, 6956−6961.
(87) Okuyama, K., Miyama, K., Mizuno, K., and Bachinger, H. P. (2012) Crystal structure of (Gly-Pro-Hyp)(9): implications for the collagen molecular model. *Biopolymers 97*, 607−616.

■ NOTE ADDED AFTER ASAP PUBLICATION

Figure 3 was incorrect in the version published ASAP November 28, 2012. The corrected version was re-posted on December 3, 2012.