A New Strategy for the Synthesis of Glycoproteins

Zhiwen Zhang,[1] Jeff Gildersleeve,[1] Yu-Ying Yang,[1] Ran Xu,[1] Joseph A. Loo,[2] Sean Uryu,[1] Chi-Huey Wong,[1] Peter G. Schultz[1]*

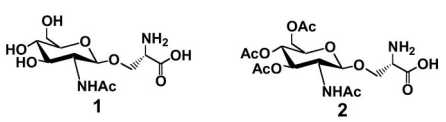
Posttranslational modifications of proteins regulate many biological processes, including metabolism, signal transduction, and gene expression. The synthetic challenges associated with generating homogeneous populations of selectively modified proteins, however, have hindered detailed studies of the effects of these modifications on protein structure and function. Here, we report an approach to the cotranslational synthesis of selectively glycosylated proteins in which the modified amino acid is genetically encoded. We show that myoglobin containing β–*N*-acetylglucosamine (GlcNAc)–serine at a defined position can be expressed in *Escherichia coli* in good yield and with high fidelity. The β-GlcNAc moiety can be recognized by a saccharide-binding protein, or subsequently modified with a galactosyltransferase to build more complex carbohydrates. This approach should be generally applicable to other posttranslational modifications such as protein phosphorylation, acetylation, and methylation.

Glycosylation is one of the most common posttranslational modifications of proteins in eukaryotes, and it affects a wide range of protein functions, from folding and secretion to biomolecular recognition and serum half-life (*1*). Although there have been substantial advances in our understanding of the effects of glycosylation, we are just beginning to understand the specific roles of oligosaccharide chains and the relationships between their structures and functions (*2*). A major challenge is that glycoproteins are typically produced as a mixture of glycoforms, making it difficult to isolate unique glycoforms from natural sources. A variety of methods have been developed to synthesize structurally defined glycoforms, but all of them impose severe restrictions on the size, quantity, and/or quality of the glycoprotein produced (*3*–*6*). Here, we report a strategy to produce unique glycoforms in *E. coli* by evolving an orthogonal synthetase-tRNA pair that genetically encodes a glycosylated amino acid in response to the amber codon, TAG. The genetic incorporation of this and other saccharide-modified amino acids directly into proteins should considerably enhance our ability to analyze and manipulate glycoprotein structure and function.

Previously, we have developed methods that make possible the systematic addition of amino acids with novel chemical and physical properties to the genetic codes of *E. coli* (*7*–*14*) and yeast (*15*). In this approach, a nonsense suppressor tRNA synthetase–tRNA pair that does not cross-react with endogenous tRNAs and synthetases is evolved to uniquely insert an unnatural amino acid. This strategy should also allow the cotranslational incorporation of glycosylated, phosphorylated, or methylated amino acids into proteins (*16*), avoiding the need for selective enzymatic or chemical modification of proteins. To test this notion, we attempted to genetically encode GlcNAc-modified serine **1** in *E. coli* (Scheme 1). This posttranslational modification is ubiquitous in nearly all eukaryotes and is involved in the regulation of cell signaling, protein trafficking, and cell growth (*17*, *18*). Because saccharide derivatives with free hydroxyl groups are transferred poorly across cell membranes (*19*), the acetylated derivative tri-acetyl-β-GlcNAc-serine **2** was used in these experiments. It has been shown that acetylation of the hydroxyl groups of sugars facilitates transport across cell membranes and that the hydroxyl acetyl groups can be deacetylated by nonspecific cytosolic esterases once inside the cell (*18*).

A series of positive and negative selections were used to isolate from a library of active site mutants, an orthogonal *Methanococcus jannaschii* tyrosyl tRNA synthetase (TyrRS) (*7*) that specifically charges the corresponding *M. jannaschii* suppressor ($mutRNA^{Tyr}_{CUA}$) (*7*) with β-GlcNAc-serine in *E. coli*. On the basis of the x-ray structure of the homologous *Bacillus stearothermophilus* TyrRS, two libraries were constructed in which residues in the tyrosine-binding site were randomized. The first was encoded by plasmid pBK-lib-m, in which the residues Tyr^{32}, Ala^{67}, His^{70}, Gln^{155}, Asp^{158}, and Ala^{167} were randomized, and the second was encoded by plasmid pBK-lib, in which residues Tyr^{32}, Glu^{107}, Asp^{158}, Ile^{159}, and Leu^{162} were randomized. The combined library had about 2.6×10^9 independent clones. This library was then subjected to a positive selection, based on suppression of an amber codon introduced at Asp^{112} in the chloramphenicol acetyltransferase gene, to enrich TyrRS mutants capable of incorporating the glycosylated amino acid. A negative selection, based on suppression of three amber codons in the toxic barnase gene, was then used to delete from the selected clones those mutant TyrRSs that incorporate endogenous amino acids. After five rounds of positive selection and four rounds of negative selection, three clones emerged with the following mutations: sugar candidate (S)1-90 (Glu^{107}Pro, Asp^{158}Cys, Ile^{159}Tyr, Leu^{162}Arg), S4-5 (Tyr^{32}Gly, Glu^{107}Gly, Asp^{158}Cys, Ala^{167}His), and S1-5 (Glu^{107}Cys, Asp^{158}His, Ile^{159}Asp, Leu^{162}Met). All of these clones appear to be highly selective for β-GlcNAc-serine, because the replacement of **2** with 1 mM serine, α-tri-acetyl-GalNAc-threonine, α/β-tri-acetyl-GalNAc-serine, or β-tetra-acetyl-Glu-asparagine does not permit cell growth above 30 μg/ml of chloramphenicol. These in vivo genetic results suggest that the newly selected mutant TyrRSs have excellent specificity toward β-GlcNAc-L-serine.

To test the efficiency and fidelity of the incorporation of **2**, a mutant myoglobin gene (Gly4TAG) containing an amber codon at the fourth position and a C-terminal His_6 tag was generated (*12*). When the mutant synthetase S1-90 was coexpressed with the $mutRNA^{Tyr}_{CUA}$ and Gly4TAG myoglobin genes in the presence of **2** in minimal media, 1 mg/liter of the full-length mutant myoglobin was produced (Fig. 1). For comparison, 5.5 mg/liter of wild-type myoglobin was produced under similar conditions, indicating a good level of suppression for S1-90. In the absence of either S1-90, $mutRNA^{Tyr}_{CUA}$, or **2**, no expression of full-length myoglobin was observed by silver-stained SDS–polyacrylamide gel electrophoresis (PAGE) (Fig. 1). Matrix-assisted laser desorption/ionization–time-of-flight (MALDI-TOF) analysis afforded an average protonated mass of the His_6 tag–purified mutant myoglobin of 18,430.1 daltons, which agrees within 86 parts per million with the theoretical mass of myoglobin containing $Glc(OH)_3NAc$-serine



Scheme 1.

[1]Department of Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA. [2]Department of Biochemistry and Biological Chemistry, University of California at Los Angeles, 405 Hilgard Avenue, 402 Paul D. Boyer Hall, Los Angeles, CA 90095, USA.

*To whom correspondence should be addressed. E-mail: schultz@scripps.edu

without methionine ($M_{theoretical}$ = 18,428.5 daltons) (Fig. 2) (*20*). In addition, no signals corresponding to either the *O*-acetylated glycomyoglobin or the wild-type myoglobin were observed. The mass spectrum data confirm a high degree of specificity for the incorporation of GlcNAc-serine into myoglobin (≥96%).

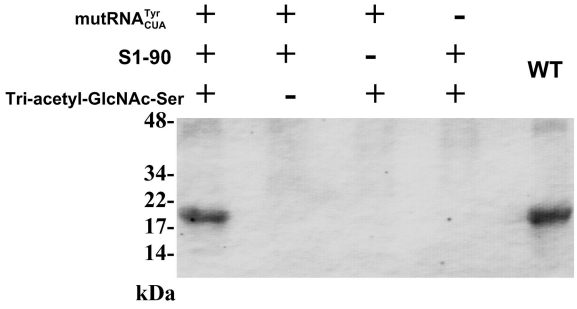


Fig. 1. Expression of the Gly4→**1** mutant myoglobin (~18.5 kD). Proteins were purified by Ni^{2+}-affinity chromatography and resolved by SDS-PAGE. The gel was silver stained. Lane 1: Myoglobin was expressed in the presence of the orthogonal tRNA, synthetase S1-90, and **2**. The band at ~18 kD corresponds to the full-length myoglobin. Lane 2: Proteins eluted after expression in the presence of the orthogonal tRNA and the synthetase S1-90 but in the absence of substrate **2**. Lane 3: Proteins eluted after expression in the presence of the orthogonal tRNA and substrate **2** but in the absence of synthetase S1-90. Lane 4: Proteins eluted after expression in the presence of the synthetase S1-90 and substrate **2** but in the absence of the orthogonal tRNA. Lane 5 contains the purified wild-type (WT) myglobin for comparison.

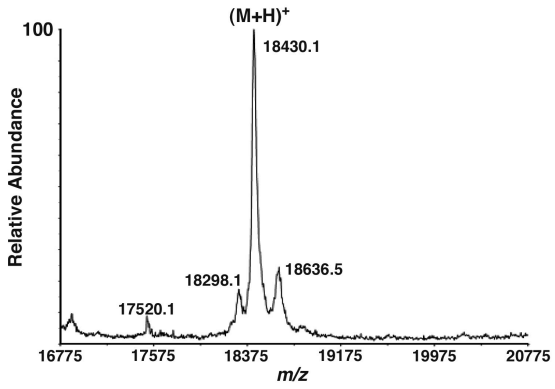


Fig. 2. MALDI-TOF analysis of the molecular weight of the Gly4→**1** mutant myoglobin. m/z, mass-to-charge ratio; $(M + H)^+$, protonated molecular ion.

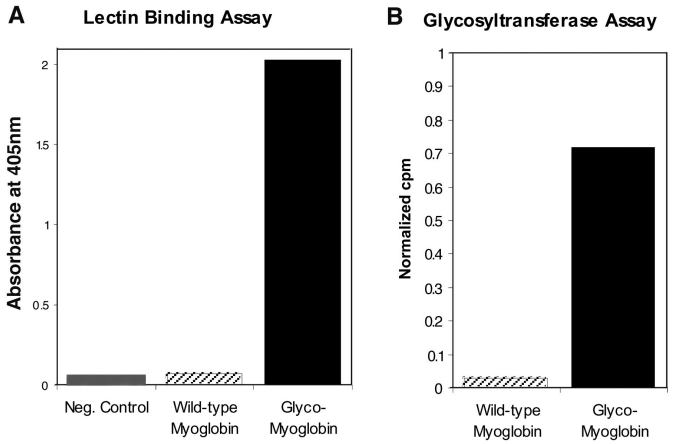


Fig. 3. Characterization of the purified mutant myoglobin containing the glycosylated amino acid. (**A**) Binding of a GlcNAc-specific lectin, BSII, to wild-type myoglobin and glycomyoglobin was tested. Absorbance values at 405 nm are shown for wild-type myoglobin, glycomyoglobin, and negative control (no lectin added). Gly4→**1** mutant myoglobin (200 ng) and wild-type myoglobin (200 ng) were immobilized in microtiter plate wells and subsequently incubated with biotinylated BSII and streptavidin-alkaline phosphatase conjugate. Wells were incubated with *p*-nitrophenyl phosphate and monitored by measuring the absorbance at 405 nm. (**B**) Quantitative analysis of the galactosyltransferase reaction was carried out in solution, and the radiolabeled galactose was normalized such that 1.0 corresponds to 100% transfer. To the solutions containing high-performance liquid chromatography–purified wild-type myoglobin (100 μg) and Gly4→**1** mutant myoglobin (100 μg) were added pyruvate kinase (5 U), UDP-glucose pyrophosphorylase (1 U), inorganic pyrophosphorylase (10 U), galactose-1-phosphate-uridyl transferase (1 U), bovine milk galactosyltransferase (2 U), glucose-1-phosphate (3 μmol), uridyl diphosphate (3 μmol), phosphoenolpyruvate (0.01 mmol), and dithiothreitol (2 μmol). After the reaction was adjusted to pH 7.2, [H^3]-galactose-1-phosphate (0.01 mmol) was added. The reaction was carried out for 48 hours at room temperature. Protein products were separated with a PD-10 Sephadex 25 column. Incorporated radiolabel was measured on a liquid scintillation analyzer.

Several additional experiments were performed to further characterize the mutant myoglobin. To demonstrate selective recognition by a GlcNAc-specific lectin, an assay similar to an enzyme-linked immunosorbent assay was used to analyze the binding of a GlcNAc-specific lectin, *Bandeiraea simplicifolia* II (BSII) (*21*), to wild-type myoglobin and glycomyoglobin (Fig. 3A). Glycomyoglobin produced a signal at least 200 times as high as that of wild-type myoglobin when it was incubated with biotinylated BSII, streptavidin-alkaline phosphatase conjugate, and *p*-nitrophenyl phosphate. In addition, this result shows that the carbohydrate has not been modified to other isomeric forms such as GalNAc and *N*-acetylmannoseamine (ManNAc), because this lectin is highly selective for GlcNAc (*21*).

Because GlcNAc is a natural substrate from which more complex carbohydrates are synthesized, we also investigated whether the *O*-GlcNAc-serine residue in myoglobin could be selectively modified with a galactosyltransferase. Beta-1,4-galactosyltransferase is known to transfer galactose (Gal) from the sugar nucleotide uridine 5′-diphosphate (UDP)–Gal to the fourth position of a GlcNAc to form Galβ1,4GlcNAc (*22*). To determine if the *O*-glycosylated myoglobin can be modified with UDP-Gal, both wild-type and *O*-glycosylated myoglobin were incubated with bovine milk galactosyltransferase and radioactive UDP-[H^3]-galactose in solution (*23*). After incubation for 48 hours at room temperature, a 72% yield of disaccharide was obtained based on the radiolabel present (Fig. 3B). We are currently exploring whether it is possible to extend the oligosaccharide chains in vivo by coexpressing glycosyltransferases in *E. coli*.

These studies demonstrate that β-GlcNAc-L-serine can be cotranslationally incorporated into proteins in *E. coli* with excellent specificity and good yield. The incorporated β-GlcNAc-serine may serve as a primary glycosylation site to which saccharides can be added sequentially with glycosyltransferase (*23*). The approach represents a new strategy for the preparation of homogeneous glycoproteins in *E. coli*, which should facilitate the studies on the structures and functions of glycosylated proteins, and if scalable, provide a unique strategy for producing homogeneous glycoforms of therapeutic glycoproteins.

References and Notes

1. R. A. Dwek, *Chem. Rev.* **96**, 683 (1996).
2. C. R. Bertozzi, L. L. Kiessling, *Science* **291**, 2357 (2001).
3. P. Sears, C.-H. Wong, *Science* **291**, 2344 (2001).
4. M. Wacker *et al.*, *Science* **298**, 1790 (2002).
5. B. G. Davis, *Chem. Rev.* **102**, 579 (2002).

6. H. C. Hang, C. R. Bertozzi, *Acc. Chem. Res.* **34**, 727 (2001).
7. L. Wang, A. Brock, B. Herberich, P. G. Schultz, *Science* **292**, 498 (2001).
8. L. Wang, A. Brock, P. G. Schultz, *J. Am. Chem. Soc.* **124**, 1836 (2002).
9. Z. Zhang, L. Wang, A. Brock, P. G. Schultz, *Angew. Chem. Int. Ed. Engl.* **41**, 2840 (2002).
10. J. W. Chin *et al.*, *J. Am. Chem. Soc.* **124**, 9026 (2002).
11. J. W. Chin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11020 (2002).
12. S. W. Santoro, L. Wang, B. Herberich, D. S. King, P. G. Schultz, *Nature Biotechnol.* **20**, 1044 (2002).
13. L. Wang, Z. Zhang, A. Brock, P. G. Schultz, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 56 (2003).
14. Z. Zhang *et al.*, *Biochemistry* **42**, 6735 (2003).
15. J. W. Chin *et al.*, *Science* **301**, 964 (2003).
16. T. Arslan, S. V. Mamaev, N. V. Mamaev, S. M. Hecht, *J. Am. Chem. Soc.* **119**, 10877 (1997).
17. L. Wells, K. Vosseller, G. W. Hart, *Science* **291**, 2376 (2001).
18. N. Lamarre-Vincent, L. Hsieh-Wilson, *J. Am. Chem. Soc.* **125**, 6612 (2003).
19. A. K. Sarkar, T. A. Fritz, W. H. Taylor, J. D. Esko, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3323 (1995).
20. The loss of the N-terminal Met is common in *E. coli*.
21. S. Ebisu, P. N. S. Iyer, I. J. Goldstein, *Carbohydr. Res.* **61**, 129 (1978).
22. K. Kamemura, B. K. Hayes, F. I. Comer, G. W. Hart, *J. Biol. Chem.* **277**, 19229 (2002).
23. K. Witte, P. Sears, C.-H. Wong, *J. Am. Chem. Soc.* **119**, 2114 (1997).

24. We thank L. Wang, S. Santoro, and A. Martin for providing plasmids for this project. Z.Z. is grateful for a National Research Service Award fellowship (GM66494). J.G. gratefully acknowledges a postdoctoral fellowship from NIH (grant 5 F32 AI10419). Funded by a grant from Department of Energy (DOE, DE-FG03-00ER45812). This is paper 15869-CH of the Scripps Research Institute.

Supporting Online Material

www.sciencemag.org/cgi/content/full/303/5656/371/DC1

Materials and Methods

Fig. S1

22 July 2003; accepted 26 November 2003

Heterodimeric GTPase Core of the SRP Targeting Complex

Pamela J. Focia, Irina V. Shepotinovskaya, James A. Seidler, Douglas M. Freymann*

Two structurally homologous guanosine triphosphatase (GTPase) domains interact directly during signal recognition particle (SRP)–mediated cotranslational targeting of proteins to the membrane. The 2.05 angstrom structure of a complex of the NG GTPase domains of Ffh and FtsY reveals a remarkably symmetric heterodimer sequestering a composite active site that contains two bound nucleotides. The structure explains the coordinate activation of the two GTPases. Conformational changes coupled to formation of their extensive interface may function allosterically to signal formation of the targeting complex to the signal-sequence binding site and the translocon. We propose that the complex represents a molecular "latch" and that its disengagement is regulated by completion of assembly of the GTPase active site.

Ffh, the prokaryotic homolog of SRP54, and FtsY, the homolog of its receptor SRα, are GTPase components of the signal recognition targeting pathway that interact directly during cotranslational targeting of proteins to the membrane (*1*–*3*). They function at sequential steps, Ffh binding to the hydrophobic signal sequence as it emerges from the ribosome and FtsY interacting with Ffh to effect release of the signal peptide to the membrane translocon. Subsequently, coordinate stimulation of the guanosine triphosphate (GTP) hydrolysis activity of the two proteins leads, ultimately, to disengagement of the targeting complex (*4*–*6*). The SRP GTPases exhibit distinct properties relative to other members of the GTPase superfamily, including relatively low nucleotide affinity and rapid nucleotide exchange (*7*, *8*). The interactions of the SRP at the membrane evolve through several stages, and GTP binding and GTP hydrolysis play different roles in the engagement and disengagement of the complex (*9*–*12*); thus, GTP binding is required, but in the presence of a nonhydrolyzable GTP analog the proteins can enable targeting of a single chain but cannot recycle (*12*). Biochemical studies of the eukaryotic SRP consistent with "empty-site" behavior before formation of the targeting complex (*11*), mutational studies of prokaryotic SRP suggesting uncoupling of GTPase function from receptor interaction (*9*), and a kinked binding mode for the GTP analog GMPPNP in its complex with *Thermus aquaticus* Ffh (*10*), together support the existence of a "primed" binding mode that functions to load nucleotide without activation until assembly of the productive complex (*10*). These data suggest that the SRP GTPases may function by a logic distinct from that of the classic GTPase "switch."

We have determined the structure of the GTP-dependent heterodimeric complex of the core GTPases of Ffh and FtsY that are integral to SRP-mediated targeting. These structurally homologous SRP GTPase subunits are termed NG, because they comprise an α-helical N domain that packs against a G-domain fold similar to other GTPases (*10*, *13*–*17*). The GTPase subunits are modular; in Ffh the NG domain occurs at the N terminus of the polypeptide, and in FtsY, at the C terminus. Like other GTPases, the sequences of the SRP GTPases are characterized by four conserved motifs, I to IV, that reflect residues directly involved in nucleotide binding and hydrolysis. Several additional sequence motifs, termed ALLEADV, DARGG, and DGQ (table S1), are distinctive to the SRP GTPases. The G domain includes an insertion box subdomain (IBD), distal to the N domain interface, that contains GTPase sequence motif II and had been thought to provide the site of interaction in the targeting complex (*13*). We report the 2.05 Å resolution x-ray structure of the engagement complex of the NG domains of Ffh and FtsY from *T. aquaticus*, stabilized with the nonhydrolyzable GTP analog, β-γ methylene-guanosine 5′-triphosphate (GMPPCP) (*18*, *19*). The structure was determined by molecular replacement using the G domain of *T. aquaticus* Ffh as a search model (*20*), followed by autotracing with ARP/wARP (*21*). The resulting electron density map (fig. S1) provided an unambiguous identification of the two polypeptides present in the asymmetric unit. The crystallized proteins retain both GTP hydrolysis and reciprocal GTPase activating protein (GAP) activities (fig. S2).

The NG domains of Ffh and FtsY associate longitudinally as a symmetric heterodimeric complex, with the N, G, and IBD regions of each protein interacting primarily with their respective counterparts (Fig. 1). The pattern of contacts of each protein across the interface, their main-chain and side-chain configurations, and their ligand-binding interactions, are remarkably similar. The G domains can be superimposed by a rotation of 179° with a root mean square Cα deviation of only 1.10 Å. Two GMPPCP molecules are buried within a composite active site chamber formed at the center of the complex by the direct apposition of the two nucleotide binding sites (fig. S3). The GTP analogs are integral to the formation of the heterodimer interface. Here, we refer primarily to structure of the Ffh NG; when we refer to FtsY, we note the corresponding residue in the Ffh sequence in parentheses. With one important exception, the interactions of equivalent residues in FtsY NG are identical to those in Ffh.

Department of Molecular Pharmacology and Biological Chemistry, Feinberg School of Medicine, Northwestern University, 303 East Chicago Avenue, Chicago, IL 60611, USA.

*To whom correspondence should be addressed. E-mail: freymann@northwestern.edu