Glycoproteins

DOI: 10.1002/anie.200806246

The Linear Assembly of a Pure Glycoenzyme

*Benjamin G. Davis**

carbohydrates · glycoproteins · native chemical ligation

Post-translational modification (PTM),[1] the alteration of a protein after its biosynthesis, often after it has been folded, takes place typically through the alteration of the functional groups in residue side chains. However, PTM is unlike transcription and translation; since PTM is not a "templated" process, it is often unpredictable and can give rise to complex mixtures of PTM protein, and the different components in those mixtures often have different properties. Such mixtures make it difficult to fully understand the PTM products that we obtain from biology and their resulting structure–activity relationships. One potential solution to this difficulty is chemical assembly.[2]

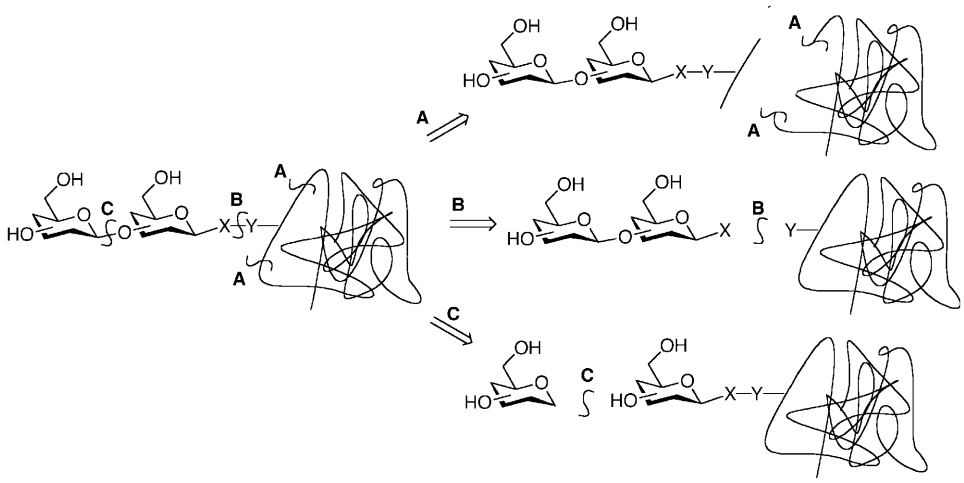

Figure 1. Strategies for glycoprotein synthesis.

By far the most diverse of the modifications is protein glycosylation; such glycosylation can take place not only post-translationally but also co-translationally. It has been estimated that some 70% of cell surface proteins in humans are glycosylated, and in many cases we do not have a clear idea of the function of these glycosylations. Consequently, methods for the synthesis of pure glycoproteins[3,4] has become a primary goal in chemistry, and has, indeed, been described by some as one of the great unsolved challenges for organic synthesis.[5]

The synthesis of modified proteins, such as glycoproteins, may be broken down into disconnections and corresponding assembly according to three general strategies (modes A–C in Figure 1).[3] Mode C involves the alteration of an existing modification, often through site-selective methods that make use of the selectivity of biocatalysis, although such methods by definition must rely on an existing site of modification and a preexisting modification; in the field of glycoprotein synthesis this is often referred to as "glycoprotein remodeling". Mode B, the site-selective convergent installation of a modification, utilizes, typically, a prefolded protein platform, often derived from expression in a straightforward expression system such as *E. coli*; modifications, even complex modifications, are attached which are often accessed through target synthesis. Mode A provides an alternative strategy and involves the linear assembly of modified amino acids or peptides in a growing peptide chain. This could, in theory, correspond, for example, to the use of a modified (glycosylated) amino acid building block in solid-phase peptide synthesis (SPPS); however, the peptide chains accessible by SPPS (typically $<$ 50–100 residues) fall somewhat short of those found in all but the smallest proteins, and the method is more effective in delivering glycopeptide[3,6] "fragments" of proteins.

The linear coupling of such glycopeptide fragments would result in chain lengths approaching or even matching those of naturally occurring glycoproteins; one method for such mode A linear assembly, native chemical ligation (NCL), provides a powerful tool for such glycopeptide coupling.[7–12] This process, which has been excellently reviewed elsewhere,[13–16] typically involves the chemoselective reaction of the N-terminal cysteine residue on one peptide with the C-terminal thioester of another. The ligation essentially proceeds by transthioesterification followed by a spontaneous and essentially irreversible $S \rightarrow N$ acyl shift to give a native peptide bond. This methodology was first introduced for protein synthesis by Kent and co-workers in the 1990s[8] based

[*] Prof. B. G. Davis
Department of Chemistry, University of Oxford
Chemistry Research Laboratory
Mansfield Road, Oxford, OX1 3TA (UK)
Fax: (+44) 1865-28-5002
E-mail: ben.davis@chem.ox.ac.uk

on observations by Wieland et al.[17] in the 1950s and has since been refined to enhance its utility.[10] The fact that NCL can be carried out in aqueous media in the absence of protecting groups has seen its growing application to the syntheses of larger glycopeptides and glycoproteins, and some impressive examples have been reported.

The NCL strategy has been particularly enhanced by a variant referred to as expressed protein ligation (EPL),[14] which has been used to incorporate cysteines at the C or N terminus of bacterially expressed peptides.[14,18,19] This can be used to generate either the peptide thioester components (from the Cys side-chain thioester formed from intein extrusion) or N-terminal cysteine peptide components needed for NCL through expression hosts. In this way EPL allows ready access, for example, to the larger protein scaffolds in a more economic manner than standard SPPS methods. In an early, simple but illustrative example, the mannan-binding protein (MBP) sequence was expressed in *E. coli* as a fusion to the N terminus of a widely used intein from the mutated *Saccharomyces cerevisiae* VMA gene; this fusion construct generates protein that also bears a chitin-binding domain for ready purification. Once expressed, this intein portion self-spliced the binding domain, and the resulting peptidothioester was used in NCL with small acyl acceptor peptides including Cys-Asn(GlcNAcβ).[20] Nowadays, the applicability and availability of these inteins is such that commercially available expression vectors exist, and such "kits" (e.g., the IMPACT system) have been used by a number of groups; for example, Imperali, Hackenberger, and co-workers employed an intein in a semisynthetic route to the immunity protein Im7.[21]

Amongst several recent reports, in one of the leading examples to date, Macmillan and Bertozzi used EPL to construct three well-defined model GlyCAM-1 (a 132-residue protein) glycoproteins, in the first reported modular syntheses of biologically relevant glycoproteins.[22] The mucin-like GlyCAM-1 glycoprotein serves as a ligand during leukocyte homing and comprises two mucin domains separated by a central unglycosylated core domain. In this report, semisynthetic variants were obtained displaying: the glycosylated N-terminal domain (1), the glycosylated C-terminal domain (2), and the protein with glycosylation in both domains (3). The N-terminal glycosylated domain (1) was obtained by NCL between a glycosylated thioester peptide and GlyCAM-1 41–132, again expressed from an IMPACT CN intein fusion vector. The N-terminal cysteine-bearing fragment of GlyCAM-1 (residues 41–132) was expressed as an intein/chitin-binding-domain fusion protein and purified on chitin beads; it was subsequently cleaved from its C-terminal intein/chitin-binding domain using the factor Xa protease. The C-terminal glycoform required a procedure inverse to that used in the preparation of the N-terminal glycosylated domain (1) with a bacterially derived thioester (GlyCAM-1 1–77)[22] and a synthetic N-terminal cysteine glycopeptide (78–132), constructed by both SPPS and NCL. The final N- and C-glycosylated variant (3) was constructed from this same glycopeptide 78–132 and ligated to a bacterially expressed central core unit (C41–S77) using a small amount of 2-mercaptoethanesulfonic acid (MESNA). The resulting intermediate was subsequently ligated in a similar fashion to that used for the N-terminal glycosylated domain (1) with factor Xa to produce the final N- and C- glycosylated glycoform (3) with an impressive presentation of 13 *N*-acetylglucosamines in predetermined positions.[23]

In a recent issue of *Angewandte Chemie*, two back-to-back papers by the Unverzagt group described progress in a powerful synthesis of a single glycoform of an enzyme.[40,41] They chose an excellent model system with which to demonstrate these strategic principles. Ribonuclease (RNase) is a 124-residue protein that has been an archetype of glycoprotein function by virtue of its single N-glycosylation site at Asn34. In prescient work by Rudd, Dwek, et al. in 1994[24] one glycosylated form (a so-called glycoform) of RNase (RNase-B) was chosen as an enzymatic model for the dissection of the effects of glycosylation upon the function of the underlying protein scaffold to which it is attached. Through capillary electrophoresis they were able, somewhat heroically, to separate different glycoforms of RNase-B and to show that different sugars gave rise to different hydrolytic activities on the same protein primary sequence. Owing to the four disulfide bonds found in RNase, it has also often been the subject of detailed studies on disulfide scrambling and even in strategies for enzyme-assisted peptide ligation (including glycosylated variants)[25] and enzymatic remodeling.[26]

A glycosylated fragment (30–68) of RNase had previously been successfully prepared by Unverzagt and co-workers using SPPS and NCL methodology, at the time the first example of the synthesis of a complex-type N-linked glycopeptide using NCL.[27] An Fmoc-protected asparagine, glycosylated with a complex unprotected biantennary heptasaccharide, was introduced using 1-benzotriazolyloxy-tris(pyrrolidino)phosphonium (PyBOP) in the presence of *N*,*N*-diisopropypethylamine (DIPEA) onto a pentapeptide attached to solid support by a Rink amide "safety-catch" linker.[28] Activation of the glycosyl asparagine in situ gave the highest coupling yields, and free hydroxy groups could be capped without activating the safety-catch linker. The resulting glycopeptide was further extended by SPPS and released from the safety-catch linker by treatment with sodium thiophenolate. The resulting thioester was coupled to protein fragment RNase40–68 by NCL. This novel linker construct facilitated rapid analysis by LC–MS through the acidic cleavage of the Rink amide linker; thereby the standard two-step cleavage reaction often necessitated by SPPS was avoided.

Although there are several ways for accessing the N-terminal cysteine peptides (the acyl acceptor components in NCL) such as the use of TEV protease,[29] cyanogen bromide,[30] or factor Xa,[23] the use of inteins pursued in Unverzagt's RNAse assembly is still relatively rare. In the work described he also uses the commercially available IMPACT system, but, as is often the case, expression in *E. coli* led to the formation of insoluble protein as inclusion bodies. As a result of this intracellular precipitation of probably misfolded protein, the intein system did not self-cleave. Often the way around this problem in protein production is to resolubilize using denaturant (e.g. guanidinium chloride) and then dilute away the denaturant to give soluble refolded

protein. Here, after several attempts at this more traditional method, Unverzagt et al. have found a novel approach that seems to be particularly applicable to thiol-rich proteins. Thus, they used a carboxyethylmethanethiosulfonate (MTS) reagent[31] to efficiently cap the seven thiols in the protein fragment as mixed disulfides under the conditions of resolubilization and refolding; the intein aspartimide formation proved still to be active and give them a great yield of cleaved disulfide-modified RNase fragment. It should be noted that the MTS reagents used here greatly outstripped all other methods that were tried to form such disulfide-capped products.

These "disulfide-protected" fragments of RNase turned out to be invaluable in subsequent synthetic efforts by the Unverzagt group to create some excellent synthetic variants of RNase. Thus, as a first test, thiophenol-accelerated NCL of the disulfide-protected fragment over two days with an intein-derived thioester RNase1–39 gave good access to full-length RNase with concomitant "deprotection" (following treatment with 1,4-dithiothreitol). Interestingly, yields were improved (presumably because the formation of precipitating scrambled incorrectly disulfide-linked side products was avoided) through the use of a glovebox (< 10 ppm oxygen). Subsequent refolding in the presence of a glutathione redox couple yielded refolded RNases.

With access to a full-length unmodified RNase thus investigated and with the experience gained from creating smaller glycosylated fragments (30–68) of RNase (vide supra),[27] the stage was nicely set for the Unverzagt group to go after full-length glycoforms. They pragmatically chose a biantennary nonasaccharide-containing target, the sugar within which is of the so-called complex type. Not only are such sugars of greater relevance for mammalian systems, the nine-sugar amino acid building block needed here, $Gal_2GlcNAc_2Man_5GlcNAc_2$-Asn, can be accessed on a multi-milligram scale through extraction and digestion from egg yolk.[32] This access to larger quantities of a glycoamino acid building block from natural sources than are sometimes typically available from target synthesis circumvented what is often a synthetic bottleneck in glycopeptide/glycoprotein synthesis.

The initial strategy (strategy 1 in Figure 2) was to disconnect at Cys40 to yield simply two fragments, as before, the disulfide-protected fragment and a thioester, now a glycosylated RNase1–39. However, despite the fact that the group had had success preparing smaller segments with a heptasaccharide previously on their double-linker PEGA resin system,[27] no fragment more than 20 amino acids long could be accessed with the nonasaccharide. This frustration illustrates some of the problems that can arise with peptides even of quite modest lengths when they contain oligosac-

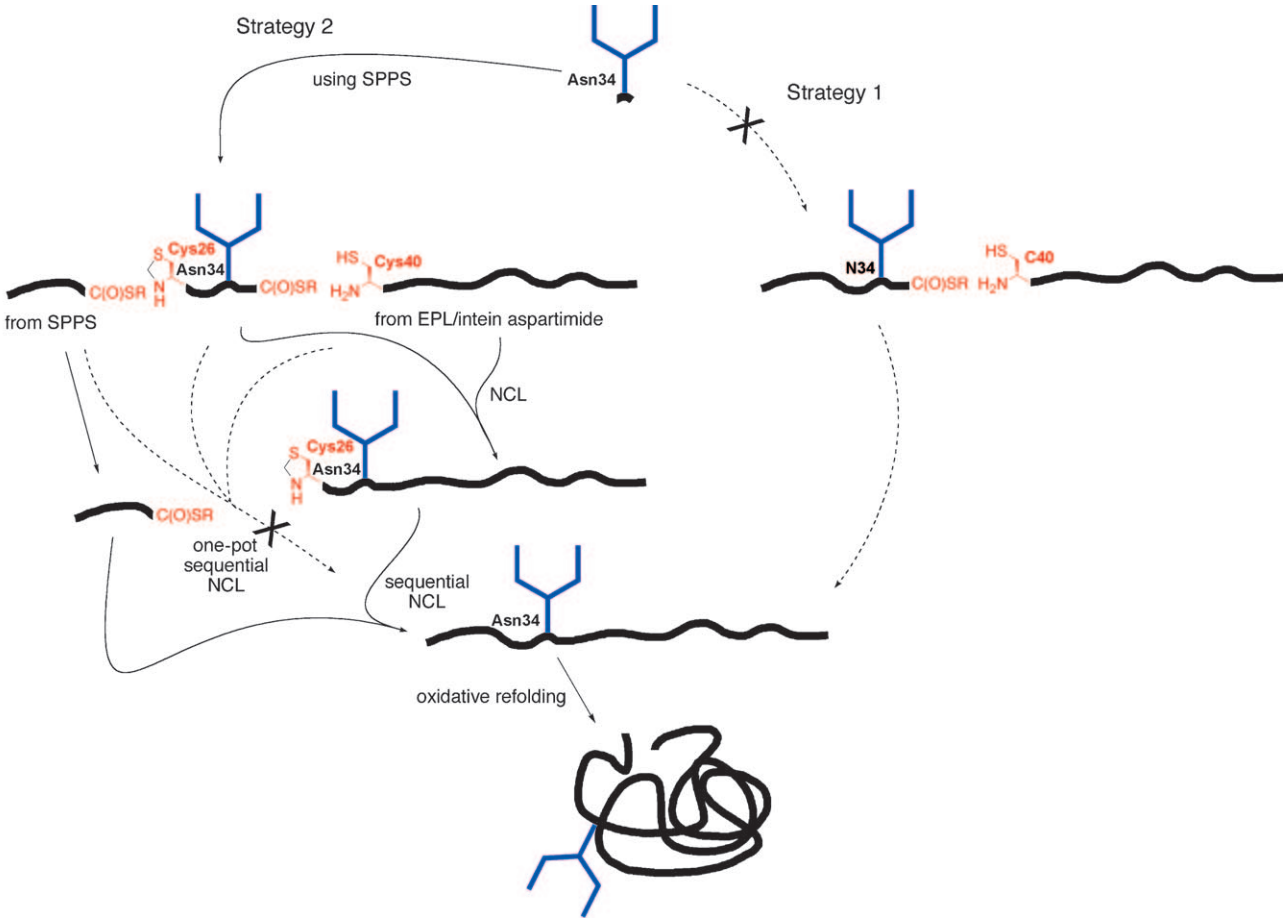

Figure 2. A labor of love: the various successful and unsuccessful strategies that Unverzagt et al. have pursued in the first NCL-mediated synthesis of a pure glycoform of enzyme.

charides. Failure necessitated a further disconnection at Cys26 and a strategy requiring three fragments in total that would be assembled by sequential NCL: RNase1–25 as a thioester for the N terminus, glycosylated RNase26–39 for the central portion, and the disulfide-protected fragment (40–124) for the C-terminal part. The assembly of more than one fragment by NCL requires some form of N-terminal protection (such as thiazolidines) to prevent homocoupling of the thioester.

The synthesis of the glycosylated central portion on the dual linker system was simplified through the use of an aminal "pseudoproline",[33] acetylation to protect the hydroxy groups of the glycan, and the replacement of potentially oxidation-sensitive Met residues by norleucine. On-resin deprotection and cleavage gave the required thiazolidine-protected thioester. The N-terminal thioester (1–25) was also assembled using an aminal pseudoproline in fair yield.

With the required fragments in hand, they then attempted the first one-pot strategy. Kent and Bang have suggested[34] that this can be achieved simply by adding methoxyamine (to deprotect the thiazoline) after the two fragments have come together, followed by further addition of the third thioester. Here, Unverzagt et al. attempted such a strategy, but they were thwarted by the, perhaps to be expected, side reaction of the good nucleophile methoxyamine with the third thioester (here thioester 1–25). As a consequence, the product of the first two fragments (glycosylated RNase 26–124) was instead isolated after thiazolidine removal in the glovebox and gel filtration. Finally, ligation with the 1–25 thioester proceeded within a day to give full-length RNase1–124 as a single glycoform. Refolding created not only an enzyme with a circular dichroism spectrum consistent with that of the native folded structure but also, importantly, hydrolytically active enzyme ($\approx 50\,\%$ the activity of RNase A).

It should be noted that the expansion of the methods available for accessing and linking glycopeptides through NCL-type strategies has recently resulted in the first purely chemical construction of other small intact glycoproteins by mode A NCL assembly. Examples have included the incorporation of a complete human complex-type sialyloligosaccharide into the 76-residue glycoprotein MCP-3,[35] based on the conventional coupling of thioesters and Cys-terminated peptides, and the synthesis of the 82-residue glycoprotein/peptide diptericin ε, using so-called sugar-assisted ligation methods.[36] These are impressive examples, but these sequence lengths are still somewhat short of those achieved here in RNase and indeed of typical glycoproteins (and the distinction of protein versus peptide is perhaps somewhat a question of semantics).

Yet, for me, this first creation of an active glycosylated enzyme by Unverzagt and his team using the mode A NCL strategy is a clear landmark in glycoprotein science. Critically, as NCL becomes a more widespread technique, the admirable focus on protein function sets an important directional lead. Although the structure–activity relationships for the effect of glycosylation upon enzyme activity have been studied before using isolated samples[24] or glycoenzymes prepared by the mode B strategy,[37,38] these have either required large samples and extensive purification or have involved unnatural linkages in the proteins. The use of NCL to create synthetic proteins is not straightforward and requires expertise; issues such as the final refolding[39] from an linear construct to create an active protein with proper tertiary structure are sometimes not addressed. Nonetheless, NCL is being adopted increasingly as a method to access precise protein constructs. The choice here of a model enzyme system allowed a robust and stringent test of function. Even though Unverzagt's variants still contain some unnatural amino acids (Nle instead of Met), this synthesis has shown for the first time how a more realistic variant containing a genuine N-link (Asn-amide-to-glycan link) might be created.

Published online: May 5, 2009

[1] C. T. Walsh, S. Garneau-Tsodikova, G. J. Gatto, Jr., *Angew. Chem.* **2005**, *117*, 7508; *Angew. Chem. Int. Ed.* **2005**, *44*, 7342.
[2] B. G. Davis, *Science* **2004**, *303*, 480.
[3] B. G. Davis, *Chem. Rev.* **2002**, *102*, 579.
[4] D. P. Gamblin, E. Scanlan, B. G. Davis, *Chem. Rev.* **2009**, *109*, 131.
[5] S. J. Danishefsky, Gordon Research Conference on Natural Products, **2005**.
[6] T. Buskas, S. Ingale, G.-J. Boons, *Glycobiology* **2006**, *16*, 113R.
[7] D. Bang, B. L. Pentelute, S. B. H. Kent, *Angew. Chem.* **2006**, *118*, 4089; *Angew. Chem. Int. Ed.* **2006**, *45*, 3985.
[8] P. E. Dawson, T. W. Muir, I. Clark-Lewis, S. B. H. Kent, *Science* **1994**, *266*, 776.
[9] Y. Shin, K. A. Winans, B. J. Backes, S. B. H. Kent, J. A. Ellman, C. R. Bertozzi, *J. Am. Chem. Soc.* **1999**, *121*, 11684.
[10] D. Macmillan, *Angew. Chem.* **2006**, *118*, 7830; *Angew. Chem. Int. Ed.* **2006**, *45*, 7668.
[11] D. S. Y. Yeo, R. Srinivasan, G. Y. J. Chen, S. Q. Yao, *Chem. Eur. J.* **2004**, *10*, 4664.
[12] C. P. R. Hackenberger, D. Schwarzer, *Angew. Chem.* **2008**, *120*, 10182; *Angew. Chem. Int. Ed.* **2008**, *47*, 10030.
[13] P. E. Dawson, S. B. H. Kent, *Annu. Rev. Biochem.* **2000**, *69*, 923.
[14] T. W. T. W. Muir, *Annu. Rev. Biochem.* **2003**, *72*, 249.
[15] B. L. Nilsson, M. B. Soellner, R. T. Raines, *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 91.
[16] See reference [10].
[17] T. Wieland, E. Bokelmann, L. Bauer, H. U. Lang, H. Lau, W. Schafer, *Justus Liebigs Ann. Chem.* **1953**, *583*, 129.
[18] T. W. Muir, D. Sondhi, P. A. Cole, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6705.
[19] R. David, M. P. O. Richter, A. G. Beck-Sickinger, *Eur. J. Biochem.* **2004**, *271*, 663.
[20] T. J. Tolbert, C.-H. Wong, *J. Am. Chem. Soc.* **2000**, *122*, 5421.
[21] C. P. R. Hackenberger, C. T. Friel, S. E. Radford, B. Imperiali, *J. Am. Chem. Soc.* **2005**, *127*, 12882.
[22] D. Macmillan, C. R. Bertozzi, *Tetrahedron* **2000**, *56*, 9515.
[23] D. Macmillan, C. R. Bertozzi, *Angew. Chem.* **2004**, *116*, 1379; *Angew. Chem. Int. Ed.* **2004**, *43*, 1355.
[24] P. M. Rudd, H. C. Joao, E. Coghill, P. Fiten, M. R. Saunders, G. Opdenakker, R. A. Dwek, *Biochemistry* **1994**, *33*, 17.
[25] K. Witte, P. Sears, R. Martin, C.-H. Wong, *J. Am. Chem. Soc.* **1997**, *119*, 2114.
[26] B. Li, H. Song, S. Hauser, L.-X. Wang, *Org. Lett.* **2006**, *8*, 3081.
[27] S. Mezzato, M. Schaffrath, C. Unverzagt, *Angew. Chem.* **2005**, *117*, 1677; *Angew. Chem. Int. Ed.* **2005**, *44*, 1650.
[28] M. S. Bernatowicz, S. B. Daniels, H. Koster, *Tetrahedron Lett.* **1989**, *30*, 4645.
[29] T. J. Tolbert, D. Franke, C.-H. Wong, *Bioorg. Med. Chem.* **2005**, *13*, 909.

[30] D. Macmillan, L. Arham, *J. Am. Chem. Soc.* **2004**, *126*, 9530.
[31] R. Wynn, F. M. Richards, *Methods Enzymol.* **1995**, *251*, 351.
[32] Y. Kajihara, Y. Suzuki, K. Sasaki, L. R. Juneja, *Methods Enzymol.* **2003**, *362*, 44.
[33] T. Wöhr, F. Wahl, A. Nefzi, B. Rohwedder, T. Sato, X. Sun, M. Mutter, *J. Am. Chem. Soc.* **1996**, *118*, 9218.
[34] D. Bang, S. B. Kent, *Angew. Chem.* **2004**, *116*, 2588; *Angew. Chem. Int. Ed.* **2004**, *43*, 2534.
[35] N. Yamamoto, Y. Tanabe, R. Okamoto, P. E. Dawson, Y. Kajihara, *J. Am. Chem. Soc.* **2008**, *130*, 501.
[36] Y.-Y. Yang, S. Ficht, A. Brik, C.-H. Wong, *J. Am. Chem. Soc.* **2007**, *129*, 7690.
[37] B. G. Davis, R. C. Lloyd, J. B. Jones, *Bioorg. Med. Chem.* **2000**, *8*, 1527.
[38] R. C. Lloyd, B. G. Davis, J. B. Jones, *Bioorg. Med. Chem.* **2000**, *8*, 1537.
[39] M. Basharov, *Biofizika* **2002**, *47*, 989.
[40] C. Piontek, P. Ring, O. Harjes, C. Heinlein, S. Mezzato, N. Lombana, C. Pöhner, M. Püttner, D. Varón Silva, A. Martin, F. X. Schmid, C. Unverzagt, *Angew. Chem.* **2009**, *121*, 1968; *Angew. Chem. Int. Ed.* **2009**, *48*, 1936.
[41] C. Piontek, D. Varón Silva, C. Heinlein, C. Pöhner, S. Mezzato, P. Ring, A. Martin, F. X. Schmid, C. Unverzagt, *Angew. Chem.* **2009**, *121*, 1974; *Angew. Chem. Int. Ed.* **2009**, *48*, 1941.

