Annu. Rev. Biochem. 2000. 69:923–60

SYNTHESIS OF NATIVE PROTEINS BY CHEMICAL LIGATION*

Philip E. Dawson[1] and Stephen B. H. Kent[2]
[1]*The Scripps Research Institute, La Jolla, California 92037; e-mail: dawson@scripps.edu;*
[2]*Gryphon Sciences, South San Francisco, California 94080; e-mail: skent@gryphonsci.com*

Key Words chemical protein synthesis, thioester, protein, peptide, solid phase synthesis, polymer-supported synthesis, protein engineering

■ **Abstract** In just a few short years, the chemical ligation of unprotected peptide segments in aqueous solution has established itself as the most practical method for the total synthesis of native proteins. A wide range of proteins has been prepared. These synthetic molecules have led to the elucidation of gene function, to the discovery of novel biology, and to the determination of new three-dimensional protein structures by both NMR and X-ray crystallography. The facile access to novel analogs provided by chemical protein synthesis has led to original insights into the molecular basis of protein function in a number of systems. Chemical protein synthesis has also enabled the systematic development of proteins with enhanced potency and specificity as candidate therapeutic agents.

CONTENTS

*At the time of the invitation, as now, Stephen Kent is President and Chief Scientist at Gryphon Sciences. Gryphon Sciences is focused on the development and sale of enhanced protein therapeutics using chemical protein synthesis. The core technology of the company is largely the subject matter of the chapter we have submitted.

0066-4154/00/0707-0923/$14.00

INTRODUCTION: Protein Science in the Postgenome Era

An important current objective in biomedical research is to understand the molecular basis of the numerous and intricate biological activities of proteins and therefore to be able to predict and control these activities. The importance of this goal is dramatically increased today because of the explosive success of the genome-sequencing projects, which have revealed hundreds of thousands of new proteins, but only as predicted sequence data (1). For the biologist, elucidation of the biological function of a predicted protein molecule is thus a challenge of great significance. In the words of Freeman Dyson, "[In the post-genome era], proteins will emerge as the big problem and the big opportunity. When this revolution occurs, it will have a more profound effect than the Human Genome Project on the future of science and medicine" (2).

For the past 20 years, most studies of the molecular basis of protein action have been carried out by recombinant DNA-based expression of proteins in genetically engineered cells (3). From its introduction, this powerful method revolutionized the study of proteins by enabling the production of large amounts of proteins of defined molecular composition and by allowing the systematic variation of the amino acid sequence of proteins (4). Expression of proteins in engineered cells is now a mature technology, and its scope and limitations are well understood: (*a*) Small proteins (i.e. $<$30 kDa) are easier to express than large, multidomain proteins; (*b*) folding of large-protein molecules can also be a challenge; (*c*) product

heterogeneity is frequently a problem, caused by uncontrolled processing of the nascent polypeptide in the cell; and (*d*) the overexpression of proteins that are toxic to the cell, such as proteases, can be difficult (5).

Additionally, because the cell is used as a protein factory, such molecular biology studies are inherently limited to the 20 genetically encoded amino acids. Efforts have been made to use cell-free synthesis to expand the repertoire of ribosomal synthesis to include noncoded amino acids as building blocks (6, 7). These attempts to incorporate other amino acids have had very limited success—obtaining adequate amounts of pure protein from the cell-free translation systems can be a significant challenge (8), and many unnatural amino acids are simply not compatible with ribosomal polypeptide synthesis (9).

Chemical synthesis is an attractive alternative to biological methods of protein production. The use of synthetic chemistry promises the unlimited variation of the covalent structure of a polypeptide chain with the objective of understanding the molecular basis of protein function. Chemistry also promises the ability to systematically tune the properties of a protein molecule in a completely general fashion.

This vision was one of the prime imperatives of organic chemistry in the time of Emil Fischer at the beginning of the 20th century. In a 1905 letter to Adolf Baeyer, Fischer wrote, "My entire yearning is directed toward the first synthetic enzyme. If its preparation falls into my lap with the synthesis of a natural protein material, I will consider my mission fulfilled" (10). In the decades since then, the challenge of applying the methods of chemistry to the study of protein action has stimulated numerous advances in synthetic methods. Historically, these advances included the use of novel reversible protecting groups (11), novel activation methods for the formation of covalent bonds (12), and even polymer-supported synthesis (13), all of which sprang from the drive to apply the science of chemistry to the study of proteins.

DOMAINS: Building Blocks of the Protein World

Because proteins are large molecules, applying chemical synthesis to them is a considerable challenge. Furthermore, the biological functions of proteins originate in the tertiary structure of the protein molecule—that is, in the precise three-dimensional folded structure of the polypeptide chain. The typical protein molecule is ∼30 kDa in size and consists of two ∼15-kDa domains (14–16); each domain has a polypeptide chain length of ∼130 (±40) amino acids (14–16). Protein domains are defined as autonomous units of folding and, frequently, of function (17, 18). As such, domains are the building blocks of the protein world. The challenge confronting the chemist is, first, the total synthesis of folded domains and then the ability to stitch these domains together to build complex protein molecules.

CHEMICAL PROTEIN SYNTHESIS: The State of the Art in 1990

Since last reviewed in this journal (19), total chemical synthesis of native proteins has made a number of important contributions to biomedical research. It is notable that the Kent laboratory at the California Institute of Technology used total chemical synthesis based on predicted gene sequence data to carry out pioneering studies of human immunodeficiency virus 1 (HIV-1) protease enzyme (20). The existence of this virally encoded aspartyl proteinase had been postulated based on an analysis of viral nucleic acid sequence data, and molecular genetic studies had indicated that its action in processing the gag-pol polyprotein was essential to the viral life cycle (21). For this reason, the HIV-1 protease was, early on, proposed as an important target for drug development. The first preparations of the enzyme of defined molecular composition were produced by chemical synthesis (22), using a highly optimized version of stepwise solid-phase peptide synthesis (19). This work proved that the active form of the HIV-1 protease was a homodimer consisting of two identical 99-residue polypeptide chains, and it showed that the chemically synthesized enzyme accurately processed the putative cleavage sites in the viral gag-pol translation product (22).

In a strikingly important contribution, total chemical synthesis was also used to prepare large amounts of homogeneous enzyme for the determination of the original crystal structures of the HIV-1 protease molecule (Figure 1). The structure of the unliganded synthetic enzyme (23) corrected a seriously flawed low-resolution structure (24) that had been obtained by using protein derived from recombinant expression in *Escherichia coli*. Even more significantly, use of chemically synthesized enzyme provided the first high-resolution cocrystal structures of the HIV-1 protease molecule complexed with substrate-derived inhibitors (25–27). These structural data were made freely available to the research community and formed the foundation for the successful worldwide programs of structure-based drug design (28) that led to the development of the highly effective protease inhibitor class of acquired immune deficiency syndrome therapeutic agents (29).

SYNTHETIC-PEPTIDE CHEMISTRY: Useful but Bounded

Despite successful syntheses of the HIV-1 protease (22) and of a limited number of other proteins (30–35), at the start of the decade of the 1990s total chemical synthesis, by the standard methods of peptide chemistry of even a small protein molecule remained a daunting task, often requiring large teams and taking years to complete, with no guarantee of success. The routine, reproducible preparation of synthetic polypeptides of defined chemical structure was limited to products of ~50 amino acid residues (19; Figure 2). This size limitation applied equally to synthesis by solution or by solid-phase methods, but for differing reasons.

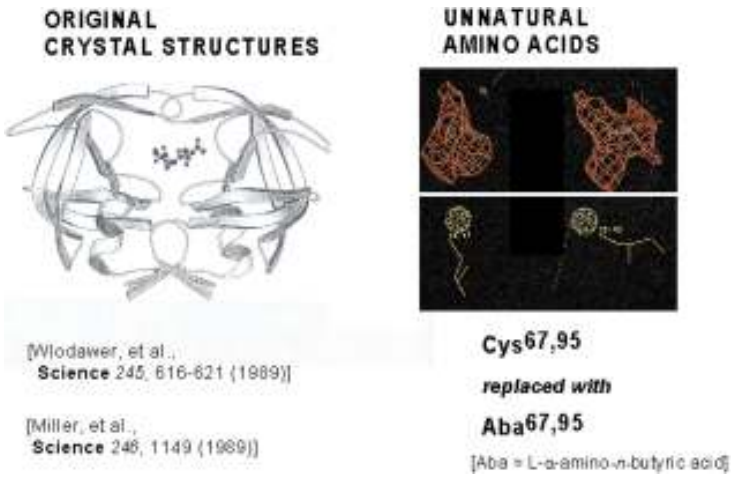


Figure 1 Crystal structures of chemically synthesized HIV-1 protease. These were the original high-resolution structures (23, 25–27) of this protein and guided the subsequent drug design programs. The synthetic protein preparation used for X-ray crystallography contained L-α-amino-*n*-butyric acid residues in place of the two Cys residues in each subunit. (*Left*) Molscript representation of the synthetic enzyme in complex with the substrate-derived inhibitor MVT101 (25). (*Right upper panel*) 2Fo-Fc electron density map for the side chains of the unnatural amino acids used to replace the two Cys residues in each subunit of the synthetic enzyme (23). (*Lower panel*) Side chains of the L-α-amino-*n*-butyric acid residues superimposed on the mercury atoms from Cys-containing enzyme (24) that has been crystallized in the same space group. This shows that the side chains of the unnatural amino acid have the same conformation as the natural Cys side chains. (Adapted from References 23 and 25).

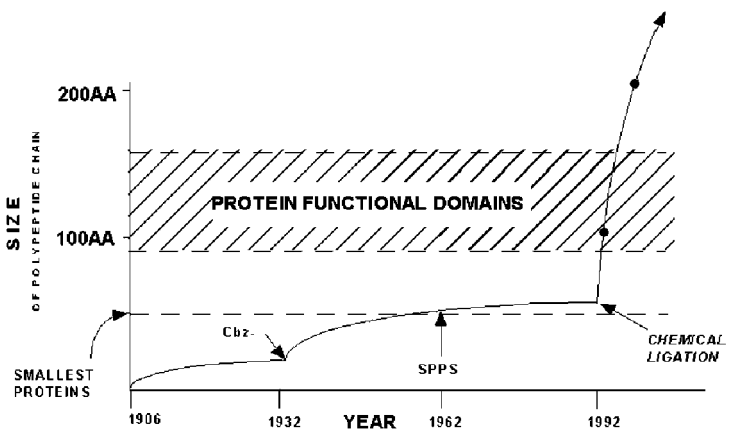


Figure 2 Historical progress in the size of synthetically accessible polypeptides.

Classical solution synthetic chemistry involves the preparation of fully protected peptide segments and their subsequent condensation in organic solvents for the convergent synthesis of large polypeptides (36). The problems associated with this classical approach have been summarized (19). These limitations include the laborious and technically demanding preparation of the protected segments, the lack of general, highly resolving methods for the purification of protected segments, and the inability to directly characterize fully protected peptides—even by modern analytical methods.[1] In addition, it became apparent that fully protected polypeptide chains frequently had only limited solubility in organic solvents that are useful for peptide synthesis. This poor solubility made such protected peptide segments difficult to work with, and the low concentrations attainable for reacting segments often led to slow and incomplete reactions (37, 38).

By contrast, unprotected peptide segments usually have good solubility properties, are more easily handled, and can be directly characterized. The most efficient way of making unprotected peptides is stepwise solid-phase peptide synthesis (SPPS). This ingenious chemical synthesis method, the progenitor of all polymer-supported organic chemistry, was introduced in 1963 by Merrifield (13). Both the principles and the practical aspects of SPPS have been thoroughly described (19). By the end of the 1980s, it was possible by highly optimized stepwise SPPS (19) to make, in good yield and high purity, essentially any peptide $\leq$50 amino acids in length. Reverse-phase high-pressure liquid chromatography methods could be routinely used to purify these synthetic products and to evaluate their homogeneity (39). More recently, electrospray mass spectrometry has provided a straightforward general method for the precise characterization of the covalent structure of unprotected synthetic peptides (40). Despite the extraordinary power of solid-phase peptide synthesis, lack of quantitative reaction eventually leads to the formation of significant levels of resin-bound byproducts. It is this statistical accumulation of coproducts that limits the ultimate size of high-purity polypeptides of defined covalent structure that can be effectively prepared in this way.

Thus, synthetic peptide chemistry, whether by stepwise SPPS (19) or by solution methods (36), can provide routine access to polypeptide chains of $\sim$50 amino acids. This corresponds to only the very smallest proteins and protein domains.

A number of attempts were made to take advantage of the ability to make, characterize, and handle unprotected peptides (41–44). Noteworthy is the development of enzymatic ligation methods for the preparation of large polypeptides from synthetic peptide segments, with ligase enzymes specifically engineered for this purpose by the methods of molecular biology (45). Ironically, the principal obstacle to general utility of enzymatic ligation has proven to be the limited

[1]For example, electrospray mass spectrometry has become one of the most useful tools for determining the covalent structure of peptides (40); this powerful method involves direct ionization of an analyte from aqueous solution. The efficacy of this ionizaton depends on the presence of multiple ionizable groups in the molecule under study. The lack of such groups in fully protected peptides precludes direct analysis by electrospray mass spectrometry.

solubility of even the unprotected peptide segments, under the physiological conditions compatible with the enzymes used (46). Despite considerable efforts and some notable successes (47), such methods have not found widespread use.

CHEMICAL LIGATION OF UNPROTECTED PEPTIDE SEGMENTS

As recently as 1991 (48), the challenge remained: namely, to develop methods that enable the general application of the tools of chemistry to the world of the protein molecule. It was evident (41–44, 48) that a truly useful approach to chemical protein synthesis would be based on the ability to routinely make unprotected peptides $\leq$50 amino acid residues in length and would consist of a practical way to stitch such synthetic peptides together to give polypeptides of any desired length, and hence the corresponding folded protein molecules.

Based on this premise, in the early 1990s the principle of chemoselective reaction (49) was adapted to enable the use of unprotected peptide segments in chemical protein synthesis (50). This novel "chemical ligation" approach relied on a conceptual breakthrough, the principles of which are shown in Figure 3.

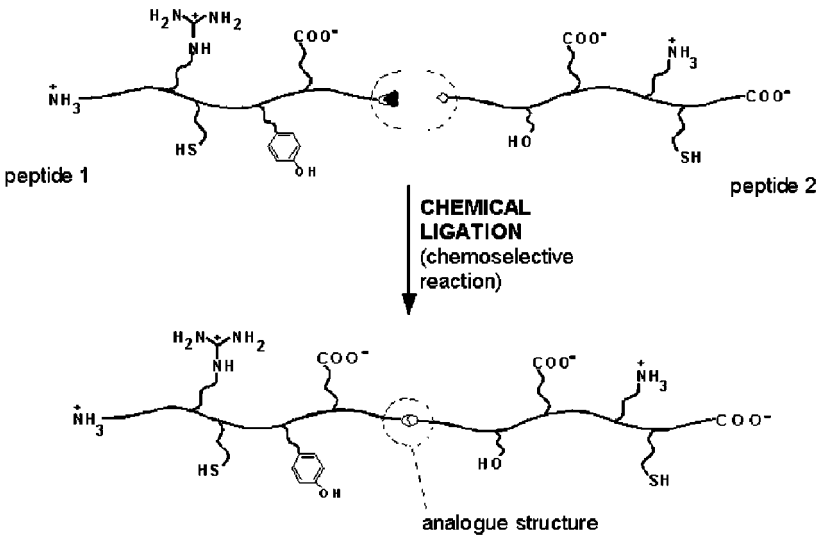


Figure 3 Principles of chemical ligation (48, 50). Uniquely reactive functionalities are incorporated into each peptide by chemical synthesis. Mutual chemoselective reaction of these moieties allows the use of completely unprotected peptide segments, which are prepared by standard means and can be readily purified and characterized by sensitive, high-resolution methods. Reaction is carried out in aqueous solution in the presence of chaotropes, such as 6 M guanidine-HCl, to increase the concentration of reacting segments and speed up the reaction. The product polypeptide is obtained directly in final form.

TABLE 1 Chemistries used for the synthesis of native proteins by chemical ligation of unprotected peptide segments

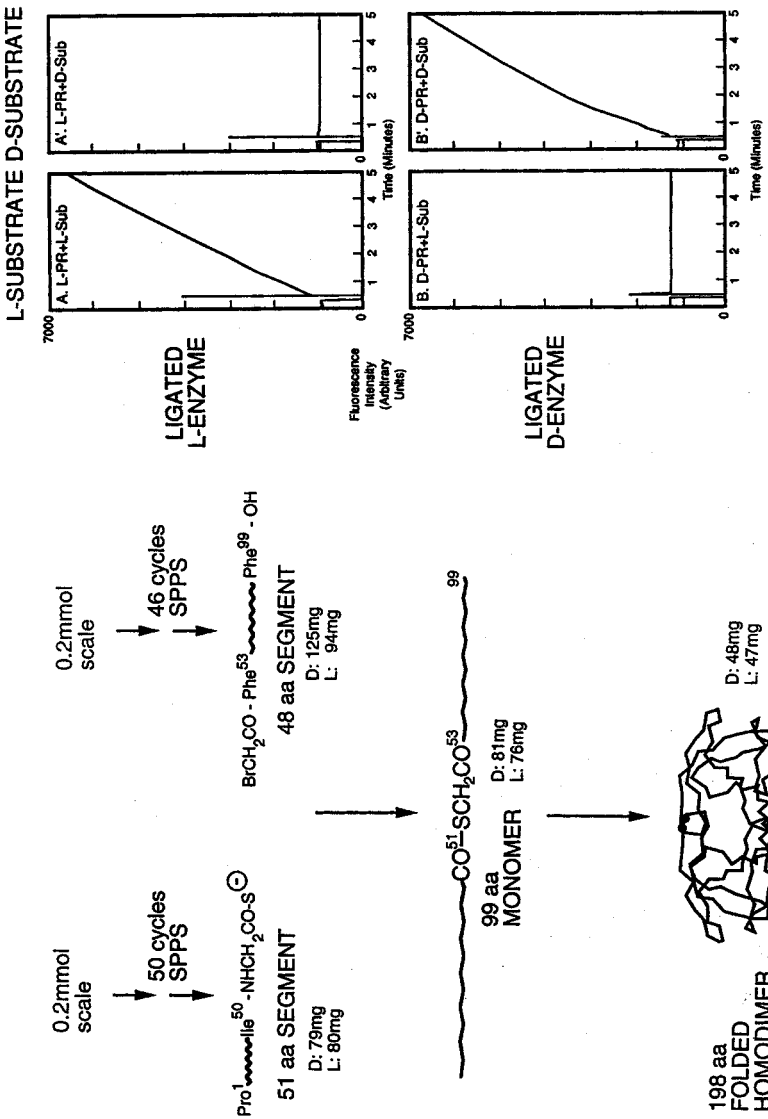
Chemistry	Reference
1. Thioester-forming ligation	50
2. Oxime-forming ligation	53
3. Thioether-forming ligation	59
4. Directed disulfide formation	85
5. Thiazolidine-forming ligation	60, 61
6. Peptide bond-forming ligation	62

In essence, the use of unique, mutually reactive functional groups not normally found in peptides enabled the site-specific ligation of completely unprotected peptide segments for the synthesis of large polypeptide chains. Reactions were designed to be carried out in aqueous solution, and a chaotropic agent such as 6 M guanidine-HCl was used to increase the solubility of the reacting peptide segments, thereby allowing the use of higher peptide concentrations to accelerate the ligation reactions.

This chemical ligation method has proven to be simple to implement, highly effective, and generally applicable (51). A variety of ligation chemistries has been used (Table 1), and the chemical ligation of unprotected peptide segments has provided access to a range of protein targets.

The price paid for such unprecedented synthetic convenience, at least in the initial stages of development of the method, was the formation of an unnatural structure at the site of ligation between two peptide segments (50). However, these unnatural structures are often well-tolerated within the context of a folded protein, and numerous examples exist of fully active protein molecules that are chemically synthesized in this way. Some early examples of proteins prepared by the chemical ligation method include (*a*) enzymatically active HIV-1 protease (50); (*b*) the mirror image enzyme D-HIV-1 protease, which was prepared by a thioester-forming chemical ligation (52; Figure 4) and its high-resolution crystal structure determined (20); (*c*) the facile total synthesis of proteinlike TASP molecules of unusual topology (53–55); (*d*) the synthesis of backbone-engineered variants of the HIV-1 protease (56) to investigate the mechanism of the enzyme (Figure 5); (*e*)

Figure 4 Total synthesis of mirror image forms of the HIV-1 protease enzyme molecule (52). (*Left*) Unprotected ~50-residue peptide segments are reacted by thioester-forming chemical ligation to give the 99-residue polypeptide chain of the HIV-1 protease monomer. Folding gave excellent yields of the homodimeric enzyme molecules. (*Right*) Reciprocal chiral specificity of the mirror image enzyme molecules, exemplified in a fluorogenic assay. The ligated L-enzyme acted only on the L-substrate, whereas the ligated D-enzyme acted only on the D-substrate.


L-SUBSTRATE D-SUBSTRATE
A: L-PR+L-Sub
A': L-PR+D-Sub
B. D-PR+L-Sub
B': D-PR+D-Sub
7000
0
1 2 3 4 5
Time (Minutes)
Fluorescence Intensity (Arbitrary Units)
LIGATED L-ENZYME
LIGATED D-ENZYME
0.2mmol scale
50 cycles SPPS
51 aa SEGMENT
D: 79mg
L: 80mg
Pro1 ~~~ Ile50 -NHCH2CO-S⊖
0.2mmol scale
46 cycles SPPS
48 aa SEGMENT
D: 125mg
L: 94mg
BrCH2CO - Phe53 ~~~ Phe99 - OH
99 aa MONOMER
D: 81mg
L: 76mg
CO51-SCH2CO53
99
198 aa FOLDED HOMODIMER
D: 48mg
L: 47mg
[(CO-S)51-53]HIV-1 PR

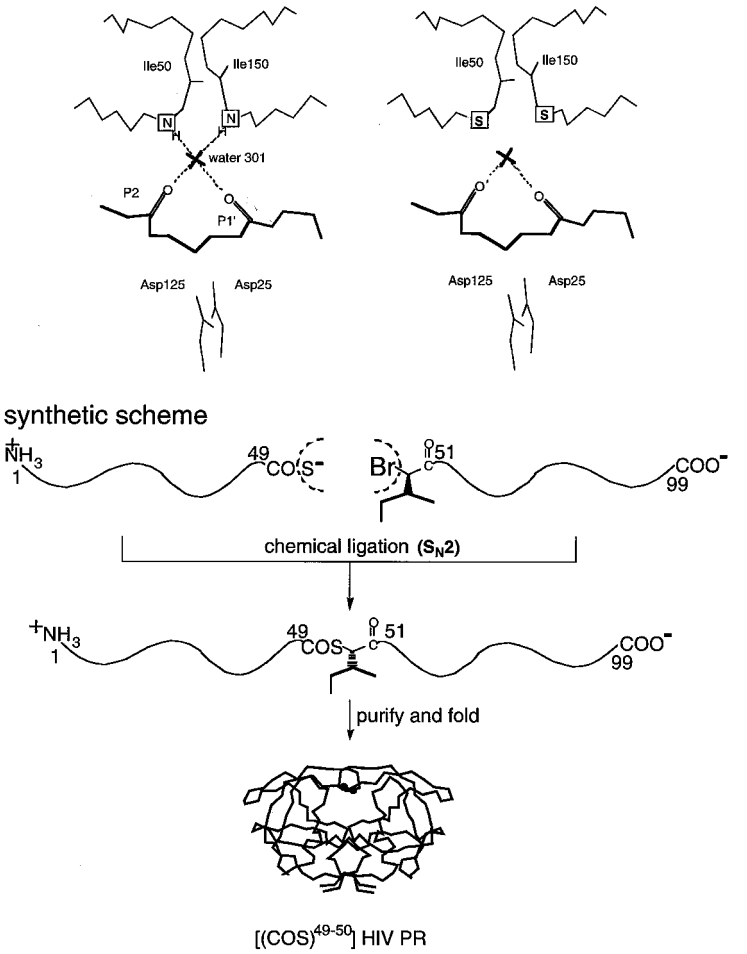


Figure 5 Backbone-engineered HIV-1 protease by chemical ligation (56). (*Top*) Design of the variant enzyme. (*Top left*) H bonding of "water 301" by the amide —NH— of Ile50 at the tip of each flap structure. (*Top right*) Sulfur atoms replacing these —NH— moieties, thus deleting the H-bonding potential. (*Bottom*) Synthetic scheme. Nucleophilic thioester-forming ligation, with inversion of configuration at the 'D-Ile50' chiral center, to give the desired 99-residue polypeptide, which is folded to form the homodimeric enzyme molecule.

the synthesis of fully functional covalent heterodimers of b/HLH/Z transcription factors (57; Figure 6); and (*f*) the synthesis of receptor mimetics (58).

These and other syntheses performed by the chemical ligation method demonstrated that proteins could now be made in high yield and good purity from unprotected peptide building blocks and that unnatural analogs could be readily prepared to investigate new aspects of protein structure and function.

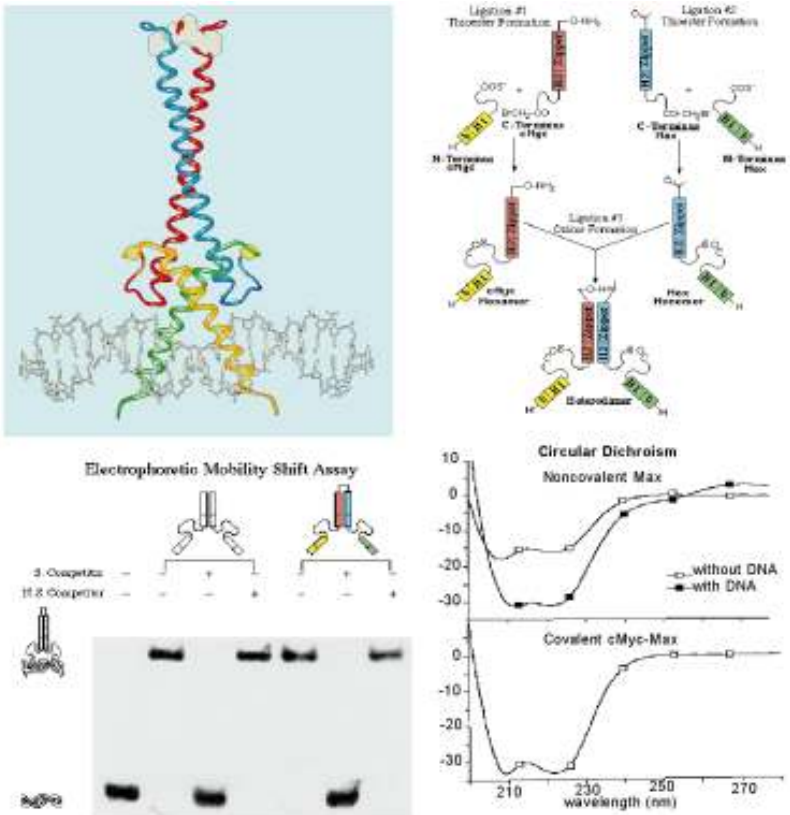


Figure 6 Total synthesis of a covalent heterodimeric transcription factor, cMyc-Max, by convergent chemical ligation (57). (*Upper left*) Molecular model of the covalent construct, bound to cognate duplex DNA. (*Upper right*) Synthetic scheme—each B/HLH/Z domain was assembled by thioester-forming chemical ligation of two peptide segments; these polypeptide products were then covalently linked by oxime-forming chemical ligation to yield a synthetic protein construct with two N terminals and no C terminal. (*Lower right*) Circular dichorism measurements showed that the covalent cMyc-Max construct folded correctly and was preordered even in the absence of cognate DNA. (*Lower left*) The covalent cMyc-Max heterodimer was active in a gel shift assay for DNA binding. Adapted from Reference 57 and Ferré-D'Amare AR (1995. PhD thesis).

NATIVE CHEMICAL LIGATION

The original ligation chemistries (50, 53, 59–61) gave a nonpeptide bond at the site of ligation. In 1994, based on the original principles of the chemical ligation method (48, 50), Dawson et al introduced an ingenious extension of the chemistries used for the chemoselective reaction of unprotected peptide segments—native chemical

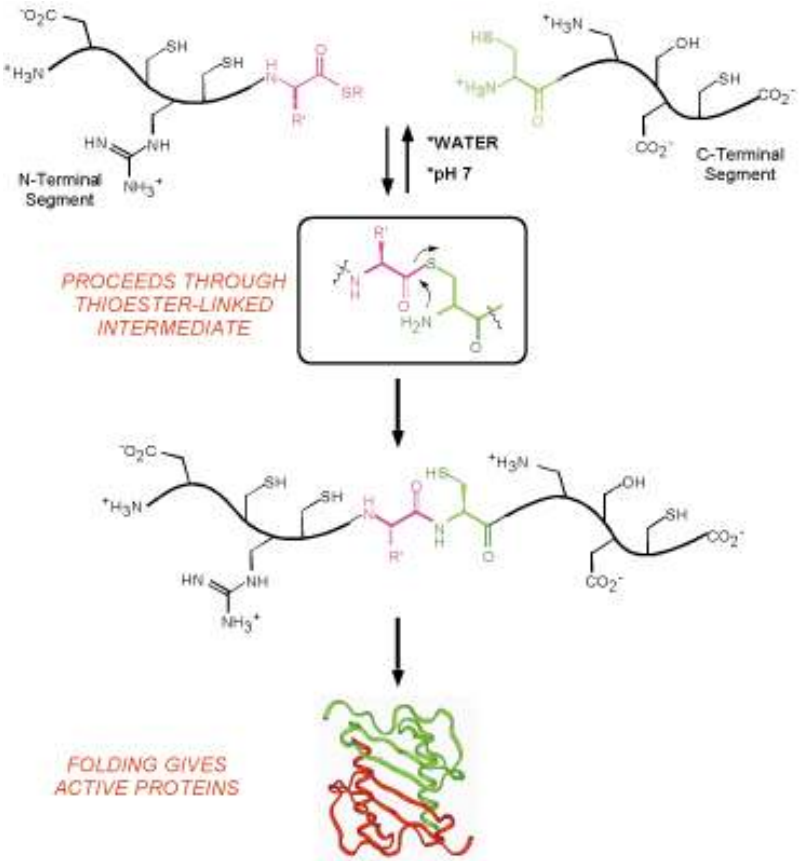


Figure 7 Native chemical ligation (62). Unprotected peptide segments are reacted by means of reversible thiol/thioester exchange to give thioester-linked initial reaction products. Uniquely, the thioester-linked intermediate involving an N-terminal Cys residue (*boxed*) is able to undergo nucleophilic rearrangement by a highly favored intramolecular mechanism; this step is irreversible (under the conditions used) and gives a polypeptide product that is linked by a native amide (i.e. peptide) bond. Only a single reaction product is obtained, even in the presence of additional Cys residues in either segment. The product polypeptide is subsequently folded to give the desired synthetic protein molecule.

ligation (62). In this method, simply mixing together two peptide segments that contain correctly designed, mutually reactive functionalities led to the formation of a single polypeptide product containing a native peptide bond at the ligation site. This highly chemoselective reaction is performed in aqueous solution at neutral pH under denaturing conditions. The chemical principles underlying the native chemical ligation method are shown in Figure 7.

The essential feature of native chemical ligation is the (transient) formation of a thioester-linked product, as was the case in the original method (50) for the

synthesis of proteins by chemical ligation. In the native chemical ligation method, however, this initial thioester-linked product is not isolated; rather, it is expressly designed to undergo spontaneous rearrangement, via intramolecular nucleophilic attack, to give the desired amide-linked product (62, 63). The result is a completely native polypeptide chain that is obtained directly in final form.

A feature of the native chemical ligation method is that ligation occurs at a unique N-terminal Cys residue. It does not matter how many additional internal Cys residues are present in either segment (62, 64). No protecting groups are necessary for any of the side-chain functional groups normally found in proteins, and quantitative yields of the ligation product are obtained.

Where this exquisite selectivity originates is important; it lies in the use of reversible thiol/thioester exchange reactions to form the thioester-linked intermediate ligation products (62, 63). The exchange is promoted by suitable thiol catalysts and is freely reversible under the neutral aqueous conditions used for the reaction. Intramolecular nucleophilic attack to form the amide bond at the ligation site is irreversible under the same conditions, so that, over the time course of the reaction, all of the freely equilibrating intermediates are depleted by the irreversible reaction step, giving a single polypeptide ligation product. Typical data from a native chemical ligation reaction are shown in Figure 8. Detailed studies of mechanistic aspects of the native chemical ligation reaction have been published (63, 65).

Formation of a native peptide bond at the ligation site has been unequivocally demonstrated by a variety of methods, including chemistry (62), NMR (66), and X-ray crystallography (67; Figure 9). A remarkable feature of the native chemical ligation of unprotected peptide segments is the absence of racemization in the coupling reaction. Detailed studies have been carried out, and no racemization was detected in the ligation product to a limit of $<1\%$ D-amino acid content (68).

BIOCHEMICAL PEPTIDE LIGATION

Protein Splicing

This cellular processing event occurs post-translationally at the polypeptide level in certain classes of protein molecules, to generate a truncated final product that results from excision of the central portion of the initial polypeptide produced on the ribosome. Intein-mediated protein splicing is a biochemical[2] reaction that

[2]It can be expected that examples of chemical ligation of protein domains will be discovered in vivo, making use of biochemical mechanisms other than intein-mediated protein splicing. In a variety of phyla, the cell already makes use of polypeptide thioesters in numerous biochemical processes. These processes include the action of cysteine proteinases (69); the ubiquitination of proteins targeted for catabolic destruction (70); the nonribosomal synthesis of peptides (71); and in the complement-mediated response to foreign pathogens (72). Given the obvious utility to the cell of cutting and pasting protein domains at the polypeptide level, it is reasonable to assume that nature will have worked out ways of taking advantage of its existing "tool kit" to accomplish this task by chemical ligation.

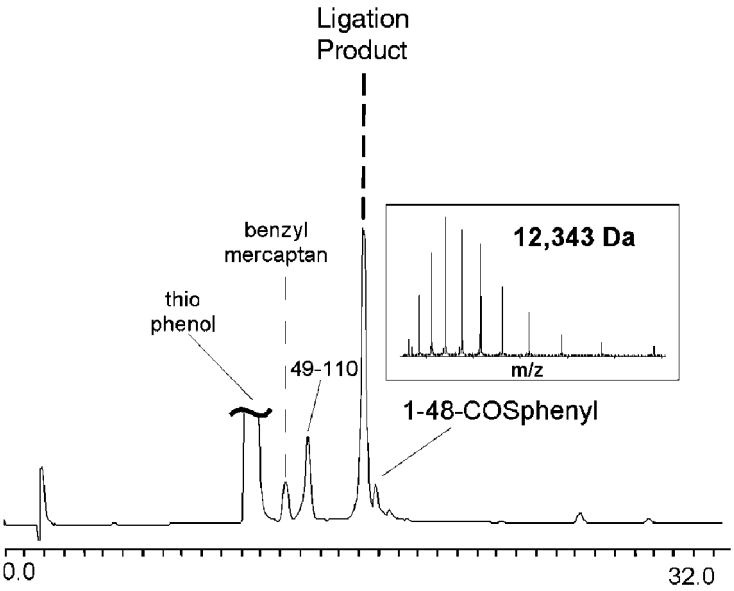


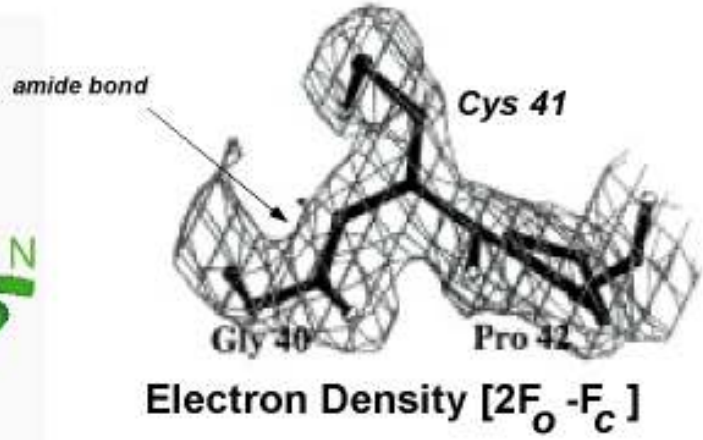
Figure 8 Raw analytical HPLC data from the synthesis by native chemical ligation of the 110-residue polypeptide chain of the enzyme barnase (63). The two unprotected peptide segments, (1-48)$^{\alpha}$COSbenzyl and Cys49-110, were reacted in aqueous solution at pH 7 in the presence of a thiol catalyst. Data after 7 h show a nearly quantitative reaction to form a single product. (*Insert*) Electrospray mass spectrometric data for the ligated 110-residue polypeptide, M_w 12,343.

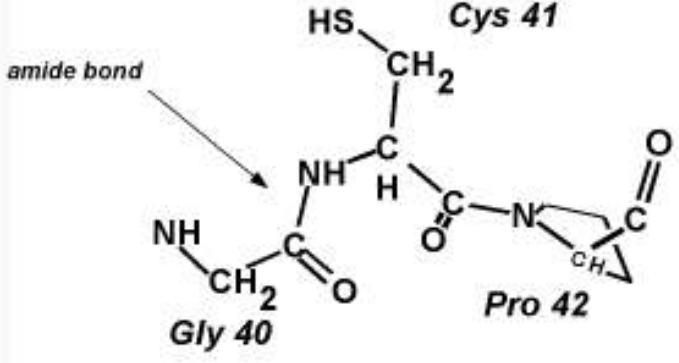
has considerable utility in its own right, and it is the subject of another chapter in this same volume (73). It is interesting that the publication of the native chemical ligation method (62) in 1994 strongly influenced the subsequent elucidation of critical aspects of the biochemical mechanism of natural protein splicing, as described by Xu et al (74). Based in part on an appreciation of the thioester-mediated acyl shift mechanism that had already been defined for the synthetic native chemical ligation reaction (62), protein splicing was shown to proceed via (intein-mediated) formation of (thio)ester-linked intermediates, followed by nucleophilic attack to form the final amide-linked spliced polypeptide (73).

In fact, the mechanisms of native chemical ligation and intein-mediated protein splicing are quite distinct in certain critical respects, despite the shared features

Figure 9 Crystal structure of the synthetic protein Eglin C. (*Left*) Tserine protease inhibitor Eglin C (*orange*) complexed with recombinant subtilisin (*green*) (67). (*Right*) Structure of the bond formed at the site of native chemical ligation in the synthesis of the Eglin C protein. The newly formed amide bond is defined by continuous electron density in the $2F_o$-F_c map. Adapted from Reference 67.


C
N
C
N


amide bond
Cys 41
Gly 40
Pro 42
Electron Density [2F$_O$ -F$_C$]


amide bond
HS
Cys 41
CH$_2$
NH
C
H
C
O
N
C
O
NH
C
O
CH$_2$
CH
Gly 40
Pro 42

of involvement of (thio)ester-linked intermediates and the final step of an *S*- or *O*- to *N*-acyl shift to form the amide-linked product. Both the mechanistic basis of selectivity and the thermodynamic driving force for the ligation reaction differ between the two processes. In intein-mediated splicing, the precise site of joining the N-extein and C-extein peptides is biochemical in origin and arises from spatial juxtaposition of the reacting residues, which is brought about by the folded conformation of the intein protein domain. By contrast, in native chemical ligation, initial reaction products may involve every thiol functionality in the reacting peptide segments (64); the exquisite selectivity originates in freely reversible thiol/thioester exchange among these initial products, followed by irreversible rearrangement of just one intermediate to give a single, defined reaction product (62).

The thermodynamic driving force also has distinct origins in the two processes. The starting point for intein-mediated splicing is amide bonds within a single polypeptide chain, and the reaction yields a ligated product and a peptide fragment (73). Thus, the driving force for the biochemical splicing reaction is actually the same as for solvolytic cleavage of peptide bonds—the generation from an uncharged amide of an ionized moiety (perhaps two) with favorable solvation properties. There may also be a contribution from the distorted (high energy) state of the starting amide bond induced by the folded structure of the intein-containing protein (73). For native chemical ligation, the reaction of a peptide-thioester with an amine to form an amide (i.e. peptide) bond is strongly favored on simple enthalpic grounds. Because of these critical mechanistic differences in thermodynamic driving force and selectivity of reaction, it is an oversimplification to describe native chemical ligation as "biomimetic" (75).

Use of defective-intein expression systems as a route to the preparation of peptide-thioesters for use in native chemical ligation is discussed on p. 951.

Conformationally Assisted Ligation

In some cases, folding conditions can be used to accelerate the rate of native chemical ligation (76). Many proteins can be cut into two or more polypeptides that can be reconstituted to form a nativelike conformation. In these cases, the revealed N and C terminals of the peptide fragments are located in close proximity at the site of chain scission. This greatly increases the collision frequency, and a weakly activated C-terminal group such as a thioester can be used to religate the fragments. Total synthesis of proteins using this approach has been demonstrated with the chymotrypsin inhibitor CI2 (Figure 10). When two synthetic peptide segments spanning the CI2 molecule, one incorporating a C-terminal thioester and the other an N-terminal cysteine, are mixed together under folding conditions, conformationally-assisted ligation proceeds in <2 min, compared with several hours for chemical ligation under denaturing conditions (76). In suitable systems, the peptide-$^{\alpha}$thioester segment can even be mixed under folding conditions with a version of the other peptide segment without a Cys at the N terminal, and conformationally-assisted ligation still proceeds in a matter of hours.

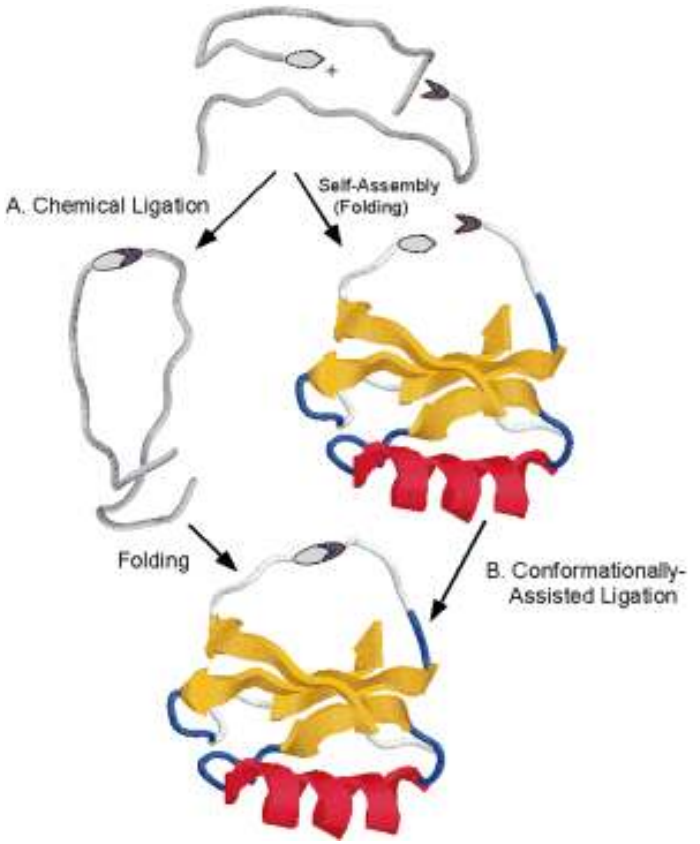


Figure 10 Conformationally assisted chemical ligation, exemplified for the chymotrypsin inhibitor C12. (*A, left*) Thioester-mediated chemical ligation at Cys under standard denaturing conditions occurs over several hours. (*B, right*) The same ligation reaction under folding conditions, in which the two segments associate to increase the collision frequency between the reacting functionalities, is complete within 3 min (76).

This conformationally assisted chemical ligation extends previously developed semisynthetic approaches that used other weakly activated ester groups for C-terminal activation of a peptide segment (77).

SCOPE OF NATIVE CHEMICAL LIGATION FOR THE SYNTHESIS OF PROTEINS

The broad scope of the native chemical ligation method for the total synthesis of proteins is summarized in Figure 11 and by the data shown in Table 2.

The first applications of native chemical ligation were to small, Cys-rich proteins such as disulfide-cross-linked secretory proteins or the zinc-finger proteins.

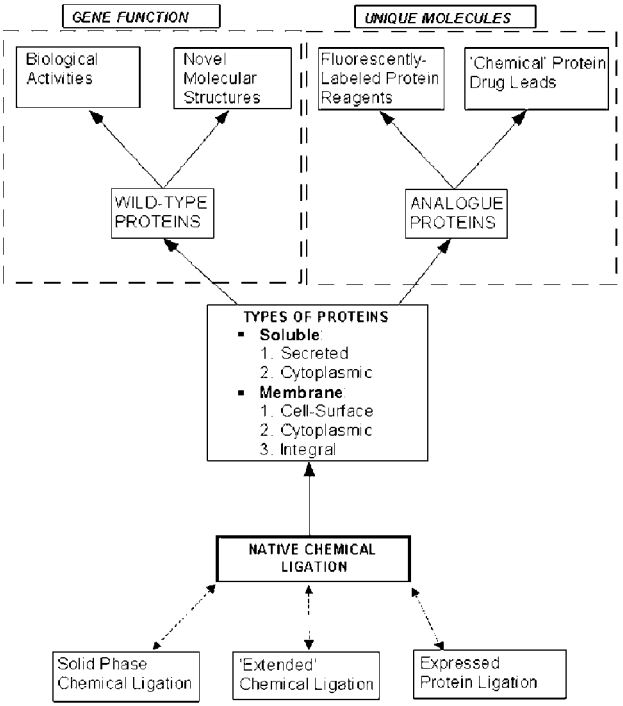


Figure 11 Applications of native chemical ligation.

In all, >300 biologically active proteins from >20 different families have been successfully prepared by total chemical synthesis with this method. These are still early results in what will surely be more widespread application of the method, but they demonstrate routine synthetic access to single-domain proteins and suggest that native chemical ligation will provide the basis of a general synthetic access to the world of proteins.

FOLDING SYNTHETIC PROTEINS

The activity of a protein molecule originates in the precise tertiary structure of the folded polypeptide chain. To complete the total synthesis of a functional protein molecule, the synthetic polypeptide chain that corresponds to the sequence of the protein must be folded to form the correct three-dimensional structure. Our intriguing experience to date has been that chemically synthesized polypeptide

TABLE 2 Selected proteins prepared by total synthesis using native chemical ligation[a]

Protein class/families[b]	Protein molecular mass (kDa)	Polypeptide size (aa)
Secretory		
Chemokines	8–10	~70
Cytokines	15–20	~160
BMPs	~25	2 × 115
Ser PR inhibitors	6–8	58–70
Agouti proteins	6–12	50–112
AFP	~6	~50
Anaphlyatoxins	~8	~70
EGFs/TGF-α	~6	~50
Receptor/membrane		
β_2 microglobulin	12	99
glp1r N-term domain	14	120
Influenza m2	50	4 × 97
Intracellular		
SH2 domains	~10	~90^{+}
SH3 domains	~7	~60
b/HLH/Z	16–20	2 × 70–180
Zn-finger	~8	~70
Redox		
Desulforedoxin	8	2 × 36
Rubredoxin	6	53
Cyt b5	10	82
Enzymes		
Retroviral proteases	20	2 × 99–116
Secretory PLA2s	14	~120
MIF	39	3 × 115
Barnase	12	110

[a]Research scale synthesis typically gives 50–100 mg of each protein; each of the above proteins had the expected biochemical or biological activity; three-dimensional molecular structures were determined by NMR or X-ray crystallography for many of these proteins.

[b]BMP, bone morphogenetic protein; AFP, anti-fungal protein; EGF, epidermal growth factor; TGF-α, transforming growth factor-α; MIF, macrophage migration inhibitory factor.

TABLE 3 Some structural motifs successfully folded as synthetic proteins[a]

Chemokine fold	SH3
Ser protease inhibitor fold	PLA2 (14-kDa form)
Kringle fold	Cytokine fold
Agouti Cys-rich domain	TGF-β fold
EGF-fold	b/HLH/Z DNA-binding domains
SH2	Rubredoxin
Zn-fingers	4OT/MIF
Aspartyl protease fold	Chitin binding domains
Anaphylatoxins	Ion channels

[a]EGF, Epidermal growth factor; TGF-β, transforming growth factor β; 40T/MIF, 4-oxalocrotonate tantomerase/macrophage migration inhibitory factor.

chains fold efficiently in vitro to give fully functional protein molecules (78). Such synthetic proteins have unique, defined folds of the polypeptide chain, as shown by NMR measurements (79, 80) and by X-ray crystallography (67, 79, 81; Figures 12 and 13). Some examples of structural motifs successfully folded as synthetic proteins are given in Table 3.

Thus, correct folding of synthetic proteins is efficient, accurate, and general at the level of single domains.

A growing list of multidomain proteins have also been successfully produced by the folding of chemically synthesized polypeptide chains (e.g. see Figures 1, 6, and 14). These proteins include homodimers (22, 23), heterodimers (20, 57),

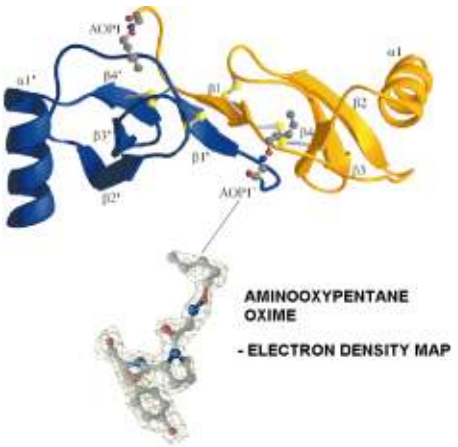


Figure 12 Crystal structure of AOP-RANTES (79). This chemically modified chemokine protein was prepared by total synthesis, using native chemical ligation. X-ray diffraction was used to determine the structure to 1.6-Å resolution. (*Top*) Ribbon structure of the crystalline dimer. (*Bottom*) $2F_o$-F_c electron density map corresponding to the unnatural aminooxypentane oxime moiety.

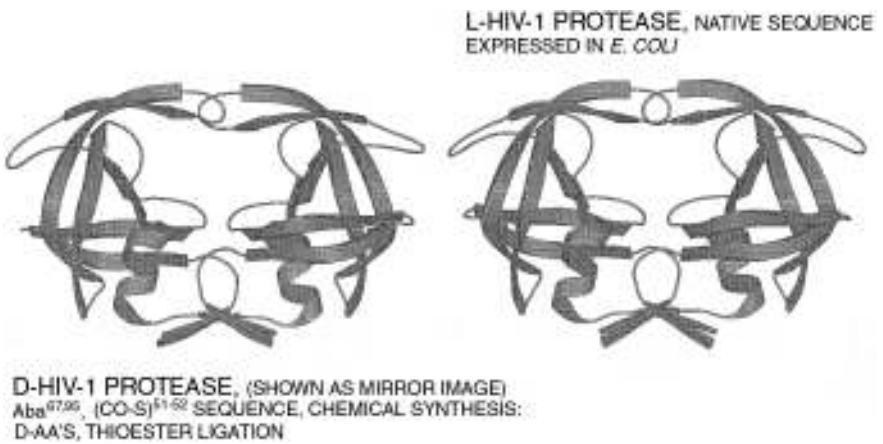


Figure 13 Three-dimensional crystal structure of the mirror image protein molecule, D-HIV-1 protease (20). The protein was prepared by thioester-forming chemical ligation of peptide segments synthesized with D-amino acids (52). (*Left*) Molscript representation of the ligated chemically synthesized D-protein molecule, displayed as the L-form for comparison purposes (20). (*Right*) Molscript representation of recombinant L-HIV-1 protease, prepared in *E. coli* (81a). The close similarity of the folded structures of the synthetic ligated D-protein and the recombinant L-protein is clearly evident.

and hexamers (82), as well as proteins containing two (57) and even three (82a, TM Hackeng, JA Fernandez, PE Dawson, SBH Kent, JH Griffin, submitted for publication) domains in a single polypeptide chain. The successful syntheses of such proteins suggests that this ability to accurately fold synthetic polypeptide chains may hold true both for single domains and for more complex proteins.

Certainly, in vitro folding, in which the system contains only a single homogeneous polypeptide of defined covalent structure, is utterly distinct from and much simpler than the situation in vivo, in which the complex intracellular environment contains multiple interacting protein species at high local concentrations. In consequence of this complexity, it has recently been found that the cell possesses a sophisticated "chaperone" apparatus that is involved in protein folding in vivo (83).

In chemical protein synthesis, the folded protein molecule is formed only at the final stages of production, under carefully controlled laboratory conditions. Control of the folding process can be particularly important in the production of proteins that are toxic to the cell, such as proteolytic enzymes. Using chemistry, it is possible to keep the polypeptide unfolded and inactive until after ligation and purification, when folding can be carried out in the presence of an inhibitor. This control over enzymatic activity was one of the key features of the success of chemical protein synthesis in the early work on the HIV-1 protease (20).

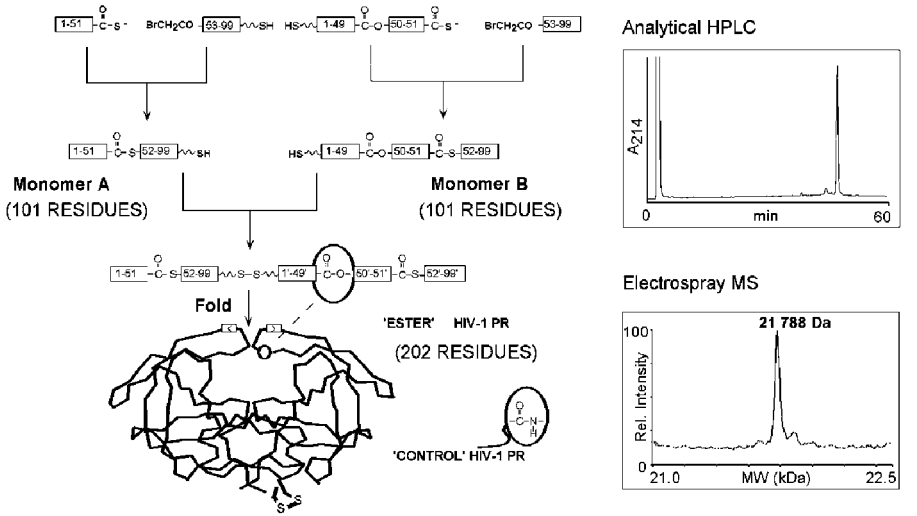


Figure 14 Total chemical synthesis of a multidomain protein: a tethered-dimer form of the HIV-1 protease. (*Left*) Convergent ligation synthetic strategy (85). Thioester-forming ligation was used to make each of the 101-residue monomers, and then directed disulfide formation was used to make the 202-residue synthetic dimer. The —NH— moiety of Ile50, in one monomer only, was substituted for by an —O— by incorporation of an ester at that position in chemical synthesis of the peptide segment (20). (*Right*) Analytical data. (*Top right*). HPLC, showing high-purity product and autolysis fragments consistent with the full enzymatic activity of the backbone-engineered 22-kDa protein. (*Bottom right*) Electrospray mass spectrometry data showing the high purity and correct observed mass of the synthetic construct.

CASE STUDIES IN THE APPLICATION OF CHEMICAL PROTEIN SYNTHESIS

Noncoded Amino Acids

Using chemistry to make proteins, it is straightforward to introduce an almost unlimited range of "unnatural" amino acids at any specific site(s) in a protein molecule. A classic example is Low's investigation of the "second shell" effects on the redox potential of an iron-sulfur protein, by systematic substitution of noncoded amino acids (84). Demonstrating the power of the chemical protein synthesis method, large amounts of each protein analog were made, purified, and fully characterized. Another example is the incorporation of a thienyl-Ala in place of a His residue in the enzyme PLA2 to establish the critical function of an imidazole side-chain functionality in the action of that enzyme (65). With chemical synthesis, multiple substitutions can be readily made at any position of the polypeptide chain, enabling virtually unlimited combinations of number, type, and position of noncoded amino acids to be incorporated into a protein molecule (85). Such studies can be very informative as to the structural basis of protein function and are made

straightforward by the synthesis of native proteins by chemical ligation of unprotected synthetic peptide segments, yet they are extremely difficult or impossible by recombinant DNA methods.

Noncoded amino acids are frequently found in native proteins in vivo. These arise from specific post-translational enzymatic modification of coded amino acid residues. One common modification of this type is γ-carboxy-glutamic acid (Gla), found for example in the eponymous Gla domains in plasma proteins. This modified amino acid is not produced biosynthetically in bacteria or yeasts, which rules out simple expression, so chemical synthesis offers an attractive route to the preparation of proteins containing Gla domains.

An example is human plasma protein S, a 635-amino-acid (aa) plasma protein that acts as an anticoagulant cofactor. This multidomain protein consists of an N-terminal Gla domain that contains 11 Gla residues, which is followed by a thrombin-sensitive region, three epidermal growth factor domains, and a sex hormone-binding globulinlike region. A polypeptide construct containing the first three domains, Gla (1–46 aa)–thrombin-sensitive region (47–76 aa)–epidermal growth factor-1(77–116 aa), has been synthesized from three segments, using native chemical ligation (TM Hackeng, JA Fernandez, PE Dawson, SBH Kent, JH Griffin, submitted for publication). Folding of this polypeptide chain produced a three-domain protein, microProtein S, that displayed anticoagulant cofactor activity.

Precise Covalent Modification

The ability to prepare native proteins by total synthesis, using chemical ligation of unprotected peptide segments, provides a convenient and general route to site-specific modification of the protein molecule. The full range of synthetic peptide and peptidomimetic chemistry (86) is at the command of the researcher who wants to make precise and controlled changes in a protein's covalent structure. Such changes are not limited by the genetic code or by the strictures of the ribosomal machinery. With chemical synthesis, virtually any conceivable covalent modification can be introduced at will anywhere in the protein molecule. An early example of the utility of this approach was the total chemical synthesis of (BTD) HIV-1 protease (87), a protein in which the Gly-Gly sequence found in a β-turn in the native protein (20) was replaced by the sterically constrained bicyclic compound BTD, a rigid mimetic of type II′ β-turn geometry (88). The resulting enzyme showed full activity and a significantly enhanced thermostability (87).

Site-Specific Tagged Proteins

Chemical synthesis enables the specific labeling of a protein molecule at unique site(s). Such specific modification is less likely to perturb the structure or activity of the protein than uncontrolled reaction with labeling reagents that stochastically target all amino or other particular functional groups in the protein. Fluorescently tagged proteins are extremely useful tools for biology and drug discovery, and synthesis of native proteins by chemical ligation allows the facile incorporation of

fluorescent dye molecules at any desired position in a protein molecule. In a recent example, a fluorescent Trp analog was incorporated by chemical synthesis into the Ras-binding domain of the protein Raf (89). The binding properties of the native domain were maintained, and the unique fluorescent label permitted the study of extremely fast kinetics of protein-protein binding.

With chemical protein synthesis, it is possible to tune the fluorescent properties of the labeled protein to the task at hand. For example, dye chelator-labeled proteins have been made for time-resolved fluorescence studies, in which it is possible to largely eliminate background emission by use of suitably "time-gated" detection (C Hunter, G Kochendoerfer, 89a).

Finally, total chemical synthesis allows the ready introduction of affinity tags, such as biotin, at precise sites in the protein molecule, while preserving biological activity, again something that is straightforward with chemistry.

Backbone Engineering

Another intriguing example of site-specific modification of the protein molecule, enabled by chemical ligation, is the covalent modification of the polypeptide backbone itself. This type of modification is not readily achieved, if possible at all, by genetic-engineering means. For example, a functionally important peptide bond (i.e. backbone amide) in the HIV-1 protease molecule was site-specifically replaced by a thioester moiety in each monomer of the homodimeric protein molecule, to investigate the direct involvement of that specific peptide bond in the mechanism of action of the enzyme, as suggested by the X-ray crystallographic data (20, 56). This approach was extended to the construction by total chemical synthesis of a 22-kDa covalent tethered dimer of the HIV-1 protease (20, 85), in which only one monomer was site specifically backbone engineered (Figure 14). The results of these studies showed that the two flap regions of the homodimeric native HIV-1 protease molecule work analogously to the single flap moiety in the two-domain, single-polypeptide chain, cell-encoded aspartyl proteinases (90).

A similar backbone engineering approach, in which specific amide —NH— moieties were replaced by —O— atoms, has been used to investigate the contribution of individual backbone H bonds to protein-protein interactions (91). More recently, an engineered backbone structure was introduced into bovine pancreatic trypsin inhibitor, by replacing one Cys residue involved in forming a disulfide bond with an *N*(ethylmercaptan)Gly, to investigate the effects of such a substitution on the folding, activity, structure, and stability of the resulting protein molecule (92). In this novel protein analog, the side chain of the Cys residue has effectively been moved to the backbone amide N atom.

Protein Medicinal Chemistry

Synthetic access enables the systematic application of the principles of medicinal chemistry to the protein molecule itself. An example is the total chemical synthesis of the potent anti-HIV molecule AOP-RANTES (79) (Figure 12). This chemical

protein analog was used as a lead compound in a successful program, based on chemical ligation, to develop even more potent anti-HIV molecules (J Wilken, D Thompson, H Gaertner, O Hartley, R Fish, JM McDonnell, Q Xu, D Fushman, D Cowburn, N Heveker, J Picard, SBH Kent, R Offord, manuscript in preparation). The chemical protein analog NNY-RANTES, which resulted from the first phase of this program, is >30-fold more effective as an anti-HIV compound and has been shown to prevent HIV infection at low nanomolar concentrations in the huPBL-SCID mouse model for acuired immune deficiency syndrome (93). NNY-RANTES is the most potent known anti-HIV compound. It is believed to work by inhibiting receptor recycling (94), thus clearing CCR5 from the surface of peripheral blood cells, a mechanism distinct from current clinical therapies for acquired immune deficiency syndrome.

Rapid Access to Functional Gene Products

In the past few years, an important new application has emerged for chemical protein synthesis—to enable rapid access to functional wild-type protein molecules directly from gene sequence data (Figure 15). Success of the genome projects has resulted in the discovery of >100,000 new proteins (1). However, these newly discovered molecules are known only as predicted open reading frames in genome sequence databases—the biomedical researcher rarely has access even to the cDNA clone corresponding to a particular gene, let alone the protein itself. For example, the recent elucidation of the complete DNA sequence of the genome of *Caenorhabditis elegans* resulted in the identification of 19,090 open reading frames encoding ~7.5 million amino acid residues of polypeptide sequence (95)! The probable roles of many of these predicted proteins can be tentatively assigned by analogy to proteins of known function, using bioinformatics. Nevertheless, the precise biochemical properties of a mature gene product can only be assessed at the level of the protein molecule itself.

Synthesis of native proteins by chemical ligation of unprotected peptides can provide access in a matter of days to large (10^{+}mg) amounts of functional protein molecules of exquisite homogeneity, based directly on gene sequence data. Secretory proteins, which are generally small and rich in Cys residues, are particularly suited to facile preparation by native chemical ligation. As described above, over the past 3 years, >300 proteins and protein analogs have been prepared by this method (78). These synthetic proteins have been used in a wide range of biomedical research investigations, resulting not only in the definition of gene function but frequently in the elucidation of novel biology (96).

Structural Biology

Facile access to the large (i.e. multiple tens of milligram) amounts of high-purity preparations produced by chemical protein synthesis can be of great utility for studies of protein structure by NMR spectroscopy and by X-ray crystallography. New methods for NMR spectroscopy have considerably enhanced the speed with which the structure of small (i.e. <200-aa-residue) proteins can be determined.

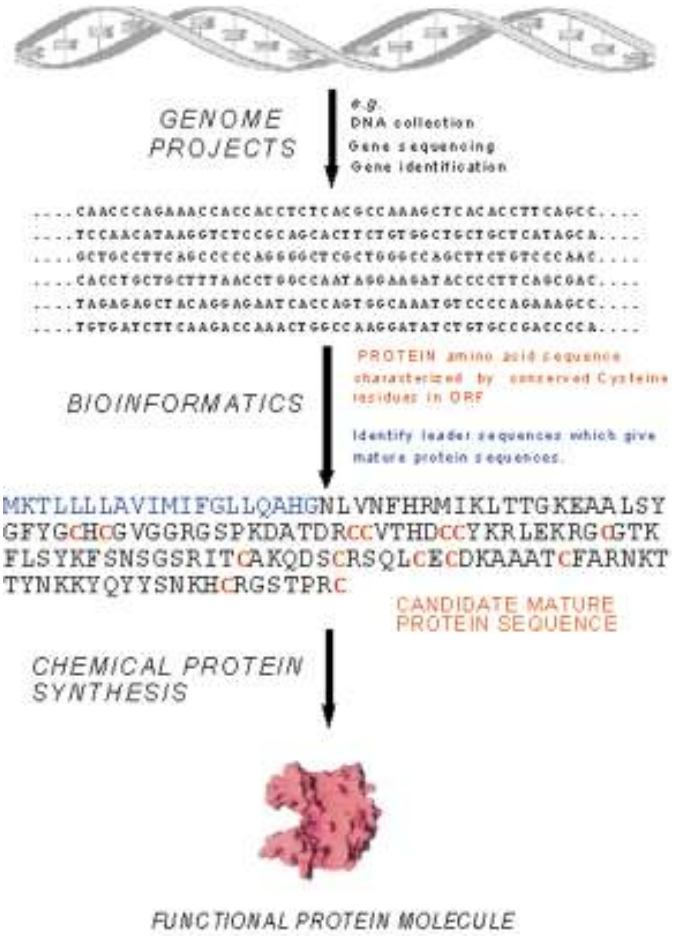


Figure 15 From gene sequence direct to functional protein, using chemical protein synthesis.

Obtaining sufficient (i.e. >10-mg) amounts of correctly folded proteins is now often the limiting step in structure determination. In a number of instances, total synthesis by chemical ligation methods has provided rapid access to high-purity protein samples in amounts useful for NMR studies (97, 98).

A recent case study of the determination by NMR of the novel structure of a chemically synthesized protein is the C-terminal Cys-rich domain of the "agouti-related" protein (80), a natural antagonist of the melanocortin-4 receptor involved in the control of human feeding behavior. In addition to small protein domains, chemical ligation approaches have contributed to the analysis of large proteins, using NMR techniques. Muir and coworkers have made use of recombinant

expression of N-terminal cysteine and thioester polypeptides (folded as domains) to label individual domains within multidomain proteins (66). By reducing spectral complexity, this approach promises to greatly simplify the NMR analysis of proteins that are >200 amino acids in size.

Chemistry also enables the precise site-specific introduction of NMR probe nuclei into the protein molecule. Thus, for the HIV-1 protease, the single γ-C atom of the active-site Asp side-chain carboxylate in each protein subunit was uniquely ^{13}C-labeled (99). NMR measurements in the presence and absence of inhibitor showed distinctive chemical shifts as a function of pH. It was possible to define the protonation state of the enzyme's catalytic apparatus and, from the unusual and dramatic chemical shifts observed, to deduce the molecular basis of the enhanced nucleophilicity of one of the two Asp side chains at the active site. It is this "super nucleophilicity" that is the defining feature of aspartyl proteinases as a class (100). This ability to precisely define at the level of a single functional group the unique molecular basis of enzymatic properties demonstrates the power of chemistry applied directly to the protein molecule itself.

Similarly, new X-ray crystallography methods have accelerated the pace of protein structure determination. In increasing instances, protein synthesis by chemical ligation has been used in conjunction with X-ray crystallography to determine the structures of novel proteins. Examples include, the chemokine SDF-1α (81), the chemical protein analog AOP-RANTES (79; Figure 12), and the mirror-image enzyme molecule D-HIV-1 protease (20; Figure 13).

Another important application of chemical protein synthesis is in the emerging genomic structural biology programs, which are aimed at the determination of the three-dimensional molecular structures of representative examples from all classes of proteins encoded in a particular genome (101). Such high-throughput structure determination will require access to great numbers of proteins in high purity and large amount. In addition, incorporation into the protein molecule of seleno-methionine residues is essential to also provide direct phase information from anomalous X-ray scattering on the same protein sample. Chemical protein synthesis by the methods described here is well suited to provide the proteins needed for genomic structural biology. A pilot study has been successfully completed in which the viral chemokine vMIP-II was prepared in [Se]Met-containing form and used for structure determination by both ^{1}H-NMR (98) and X-ray crystallography (E Lolis, submitted for publication).

CURRENT DEVELOPMENTS

Expressed Protein Ligation

From its inception, the native chemical ligation method was also envisioned for use with peptides that are produced by recombinant means (62). There are now multiple examples of the chemical ligation of recombinant peptides. These alternate

sources for suitably functionalized peptides have extended the applicability of the native chemical ligation method to include the world of peptides and domains that can be successfully produced by recombinant-DNA-based expression methods.

N-terminal cysteine recombinant peptides can be generated either by proteolytic cleavage next to a cysteine residue (102) or by an intein-based approach (103). These recombinant products can be reacted with synthetic peptide-thioesters to generate native polypeptides of hybrid biological and chemical origin. More recently, intein-based protein expression vectors have been adapted to generate polypeptide thioesters by recombinant means for use in native chemical ligation (104, 105). Interception of the partly rearranged splicing intermediate by a suitable thiol generates a recombinant peptide-thioester (Figure 16). These peptide-thioester segments can be reacted by native chemical ligation with a synthetic N-terminal Cys peptide to generate native polypeptides of hybrid chemical and biological origin (104–107). With the approaches described above, both the peptide-thioester and the N-terminal Cys peptide can be of recombinant origin. This permits the use of native chemical ligation for the mixing and matching of recombinant polypeptide segments in vitro (66).

Use of recombinant methods to generate the necessary peptide-thioester segments thus permits even molecular biologists who are not skilled in chemistry to use the native chemical ligation technique (106, 107). This "expressed protein ligation"[3] can be expected to lead to widespread use of the native chemical ligation method in biological research laboratories (109, 110).

Solid-Phase Protein Synthesis

The principles of polymer-supported organic synthesis (13, 19, 111) have been applied to the chemical ligation of unprotected peptide segments in aqueous solution [(112); Figure 17]. In solid-phase chemical ligation, unprotected peptide segments of 35–50 amino acids (i.e. ~5 kDa each) are used as building blocks to assemble the target polymer-bound polypeptide by consecutive ligation on a water-compatible polymer support. Strategies for segment condensation in both the N-to-C and C-to-N directions have been used successfully for solid-phase protein synthesis (112) and alternative linker chemistries developed (112a).

Target molecules have been constructed from as many as eight peptide segments by solid-phase chemical ligation [e.g. the polypeptide of the tissue plasminogen activator catalytic domain; M_w 25,000 (W Lu, unpublished data), and the polypeptide chain of the enzyme secretory PLA2 GV has beenassembled in a single day

[3]Sometimes erroneously referred to as "intein-mediated ligation" (106). It is important to note that, where incipient splicing of a defective intein is simply used as a way of generating a (recombinant) peptide-thioester, the ligation itself is not intein mediated; rather, the ligation reaction is standard native chemical ligation of two unprotected peptide segments (62). For an example of true intein-mediated ligation, see Reference 108.

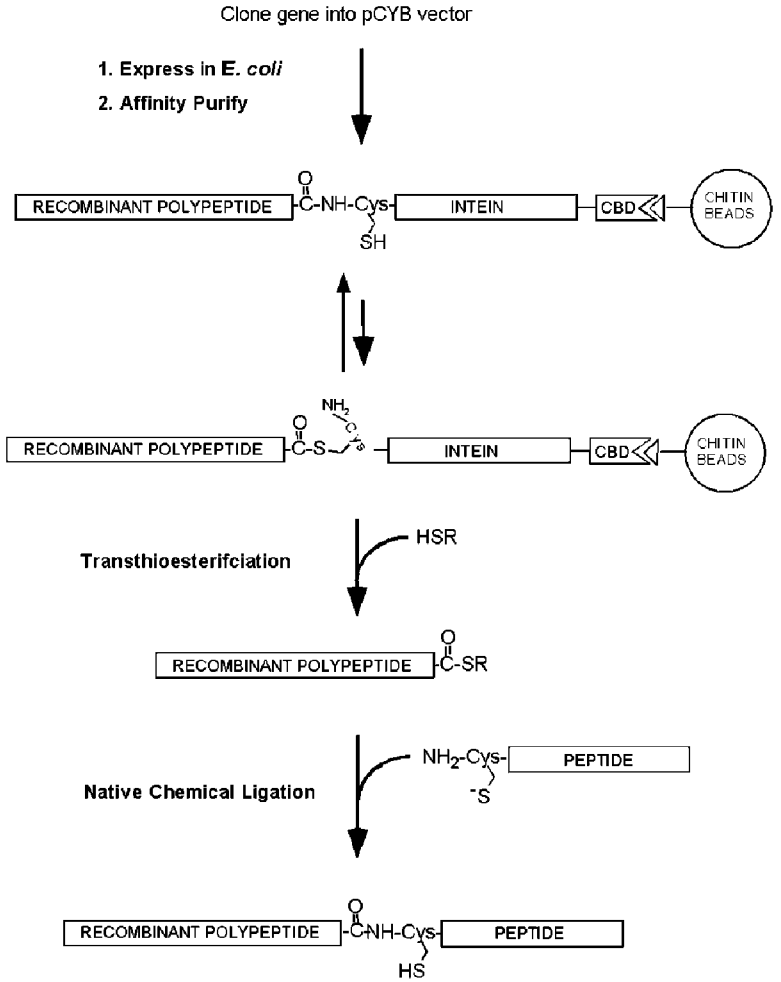


Figure 16 Expressed protein ligation (104). This process uses intein-mediated (73) preparation of a recombinant peptide-$^{\alpha}$thioester, which is then reacted with a Cys-peptide segment by native chemical ligation to prepare the desired product.

from four peptide segments (112). It can be anticipated that solid-phase chemical ligation will provide a practical chemical route to proteins that contain several hundred amino acids (Figure 18).

Membrane Proteins

An important aspect of the study of proteins which have been predicted from gene sequence data is the integral membrane class of proteins. Computer-aided analysis of the predicted open reading frames from a number of completely sequenced

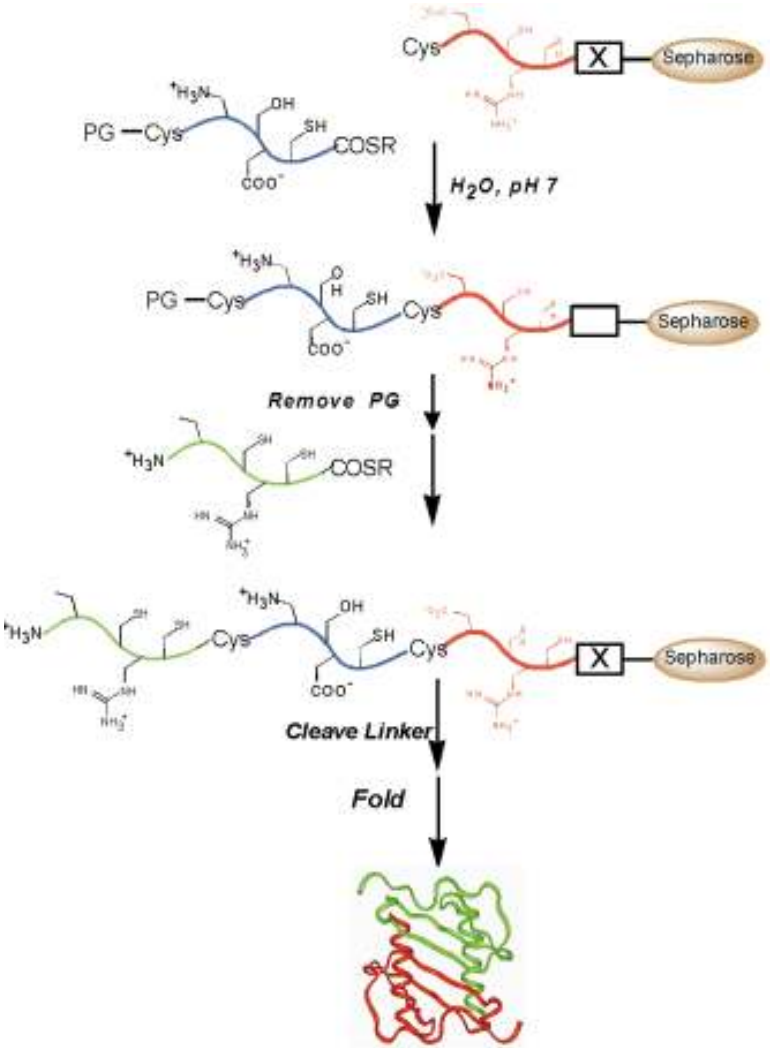


Figure 17 Solid-phase chemical ligation (112). Native chemical ligation of unprotected peptide segments and the principles of polymer-supported synthetic organic chemistry (13, 19, 111) are applied to solid-phase protein synthesis. In the example shown, the C-terminal segment of the target polypeptide is attached by a cleavable linker to a water-compatible support. The next segment as a peptide-$^{\alpha}$thioester is reacted by native chemical ligation, to give the polymer-bound ligation product. After removal of the Cys-protecting group (*PG*), successive rounds of ligation can be carried out to give the polymer-bound target polypeptide. After cleavage from the polymer support, the product is purified and folded to give the target protein molecule.

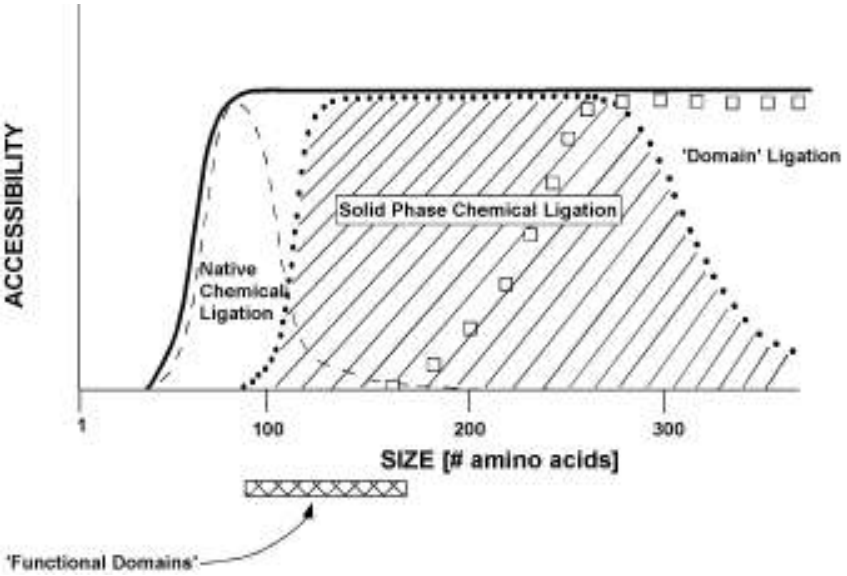


Figure 18 Size of synthetic polypeptides accessible by chemical ligation.

genomes has suggested that 20%–30% of all proteins contain membrane-spanning polypeptide sequences in the mature form of the molecule (113). Such integral membrane proteins mediate many processes in the cell, including signal transduction, ion transport, and active transport of macromolecules to name a few significant biological activities, and are thus important objectives for biomedical research. Yet integral membrane proteins are difficult to express at high levels by recombinant-DNA-based methods and have proven hard to isolate in homogeneous form in chemically defined media (114).

It is interesting that Kochendoerfer et al (115) have shown that integral membrane proteins can be synthesized in large amounts by the chemical ligation of unprotected peptide segments and isolated in high purity in media of defined chemical composition. An example is the total synthesis of the 11-kDa proton channel M2 protein of influenza A virus, which forms a tetrameric ion channel (115; Figure 19). The M2 protein had previously proven refractory to multiple attempts at expression by recombinant-DNA-based methods (W Degrado, personal communication), but was readily obtained by chemical ligation of unprotected synthetic peptides.

Glycoprotein Synthesis

Recently, the Bertozzi laboratory (116) reported the first total synthesis of a glycoprotein, using native chemical ligation in conjunction with innovative methods for the synthesis of glycopeptide-$^{\alpha}$thioesters. One of the most important applications

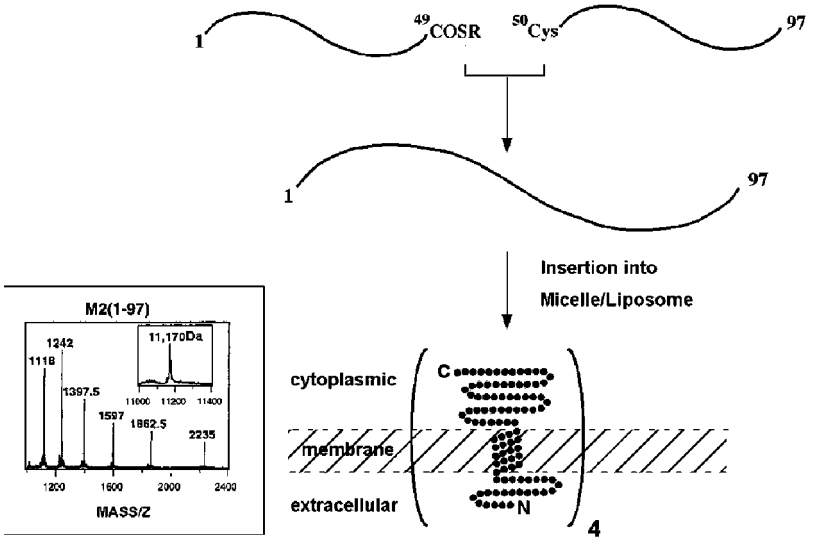


Figure 19 Chemical synthesis of an integral membrane protein (115). The 97-residue polypeptide chain of the influenza M2 protein was prepared by native chemical ligation and folded to form the active tetrameric form. (*Insert*) Electrospray mass spectrometric data showing the desired product, mass 11,170 Da.

of chemical protein synthesis will be the systematic preparation of glycoforms of gylcosylated proteins as homogeneous molecular species of defined covalent structure, to establish the role of the carbohydrate moiety in the biological function of the glycoprotein. In the near future, we can expect to see an increasing number of examples of this important capability made possible by native chemical ligation (62) and by other chemoselective reactions (117).

FUTURE DEVELOPMENTS

Ligation Sites

In its current form, native ligation chemistry uses a Cys residue at the site of formation of the new peptide bond joining two unprotected peptide segments. This means that, for a protein to be accessible by native chemical ligation, there must be no region in the polypeptide chain >50–60 aa residues without at least 1 Cys residue. Although the requirement for a Cys at the ligation site may superficially be viewed as a stringent limitation of the method, it is actually less restrictive than it at first seems. There are hundreds of protein families with interesting biological activities, encompassing many thousands of protein molecules that have native Cys residues located in positions compatible with direct application of native chemical ligation (118).

In actual practice virtually any protein molecule can be made by native chemical ligation. For proteins with no suitable Cys ligation sites in the natural sequence, it is possible to simply put a Cys wherever one is needed for ligation, usually without deleterious effects on function (63, 67; see Figure 9). The work of Muir and coworkers is illustrative of this expedient but effective approach (66, 104, 110). Their chemical ligation of recombinantly expressed polypeptide-$^{\alpha}$thioesters to synthetic peptides has typically made use of an arbitrarily introduced Cys residue at the desired ligation site, with no deleterious effects. Also, biological researchers frequently insert Cys residues into a polypeptide chain to investigate the structure-function relationships in a protein molecule (119) or as a site for the introduction of a spectroscopic probe, such as an electron spin resonance label (120). This proven utility of arbitrarily introduced Cys residues provides considerable flexibility in synthetic design for the preparation of functional protein molecules by native chemical ligation at Cys.

Additionally, it would be desirable to have the option to use thioester-mediated chemical ligation at residues other than Cys. A prototype procedure for the use of an auxiliary-functional-group approach to native amide-forming, thioester-mediated chemical ligation has been reported (121). This work defined the principles of an effective approach to ligation at non-Cys residues, but the chemistry used had to be refined and extended because severe shortcomings were observed in the original investigation, as revealed by studies in model systems (121). In this respect, recently reported work from the Dawson laboratory at The Scripps Research Institute may represent a more effective chemistry for ligation at residues other than Cys (121a), using the same auxiliary-functional-group approach.

Size of Protein Targets

To date, it has proved possible to make every protein that has been attempted by the chemical ligation of unprotected peptide segments in aqueous solution, even integral membrane proteins. However, some targets are significantly more work than others—especially if there are multiple intermediate ligation products to handle. The recently developed solid-phase protein synthesis method (see above), using polymer-supported chemical ligation (112), provides a very effective means for the ready isolation of these intermediate products and will significantly simplify syntheses requiring ligation of multiple segments.

The work of our own and others' laboratories, including the laboratories of Offord (University of Geneva, Switzerland) and Muir (Rockefeller University, New York, NY), has failed to show any inherent size limitations for application of the chemical ligation method, up to several-hundred kilodaltons in the latter case (122). Folding of chemically synthesized polypeptide chains to form native proteins, in which significant problems might have been anticipated, is usually straightforward for the domain size proteins made to date. Folding of complex multidomain proteins may or may not be as straightforward. In any event, unlike expression systems, chemical ligation allows the option of constructing complex

proteins by separately folding each domain and then stitching the folded domains together (123).

Chemical Synthesis of Peptide Segments

Virtually any target protein can be prepared by total chemical synthesis, provided that a suitable set of high-purity peptide-thioester segments is available. Ironically, for many researchers the most challenging aspect of applying the chemical ligation method to proteins is making the peptide segments. To date, the principal constraint on widespread application of the native ligation method has been the lack of methods for the facile chemical synthesis of unprotected peptide-$^{\alpha}$thioester segments. Fortunately, there is an abundance of expertise available for the chemical synthesis of peptides (86). The need to make large numbers of analogs of thousands of native proteins by chemical ligation, and hence to prepare many tens-of-thousands of peptide segments, provides an unprecedented impetus for the development of efficient methods of peptide synthesis. We can look forward with confidence to the development of radically improved methods for the rapid, cost-effective preparation of large numbers of unprotected peptide-thioester segments for use in chemical protein synthesis (124).

SUMMARY AND CONCLUSIONS

Total synthesis by the chemical ligation of unprotected peptide segments can now provide general access to native proteins of $\leq$30 kDa (Figure 18) in size. This size range encompasses the structural and functional domains that are the modular building blocks of function in the protein world, from enzymes to receptors, from signal transduction adaptor molecules to large multisubunit protein assemblies. A wide range of different proteins has already been synthesized, leading to novel biology, new three-dimensional structures, and new insights into the molecular basis of protein function. In addition, it has already been demonstrated that it is possible to stitch together, by chemical ligation folded protein domains of any size, promising general access to the world of proteins.

Perhaps the most significant future application of chemistry to proteins will be in the creation, at will, of stable post-translational modified forms of protein molecules as homogeneous entities of precise covalent structure. This will enable the dissection at the level of the protein molecule of important biochemistry, such as the intracellular signal transduction pathways. It will also enable the systematic creation of new classes of protein therapeutics with enhanced properties.

The stage is now set for the application of the tools of chemistry to the entire universe of proteins. Truly, as Edward O. Wilson has remarked. "Where nucleic acids are the codes, proteins are the *substance* of life" (125). It is no exaggeration to say that understanding the molecular basis of protein action is one of the most important challenges of our era. The ability to apply chemistry to the study of

proteins, provided by the synthetic tools described in this article, will play an important part in addressing this challenge and will have a revolutionary impact on our understanding of gene function expressed through the medium of the protein molecule.

ACKNOWLEDGMENTS

We thank our colleague Dr. Manuel Baca for his critical reading of this chapter and for the many useful suggestions that he made. This article is a perspective on the synthesis of proteins by chemical means and, at the request of the Editors, emphasizes the work of the authors' own laboratories. Every attempt has been made to cite the original literature for all of the results described. The successes of chemical protein synthesis are due solely to the hard work of our many talented colleagues. The shortcomings of this review are entirely the responsibility of the authors.

LITERATURE CITED

1. Strasberg RL, Feingold EA, Klausner RD, Collins FS. 1999. *Science* 286:455–57
2. Dyson F. 1998. *The Red Herring Mag., June.* http://www.herring.com/mag/issue55/think.html
3. Matthews BW. 1993. *Annu. Rev. Biochem.* 62:139–60
4. Smith M. 1994. *Angew. Chem. Int. Ed. Engl.* 33:1214–21
5. Cleland JL, Craik CS. 1996. *Protein Engineering: Principles and Practice.* New York: Wiley & Sons. 518 pp.
6. Hecht SM. 1992. *Acc. Chem. Res.* 25:545–52
7. Mendel D, Cornish VW, Schultz PG. 1995. *Annu. Rev. Biophys. Biomol. Struct.* 24:435–62
8. Barrett JE, Lucero CM, Schultz PG. 1999. *J. Am. Chem. Soc.* 121:7965–66
9. Cornish VW, Mendel D, Schultz PG. 1995. *Angew. Chem. Int. Ed. Engl.* 34:621–33
10. Fruton JS. 1992. A *Skeptical Biochemist*, p. 137. Cambridge, MA: Harvard Univ. Press.
11. Bergmann M, Zervas L. 1932. *Chem. Berichte* 65:1192–201
12. Sheehan JC, Hess GP. 1956. *J. Am. Chem. Soc.* 77:1067–68
13. Merrifield RB. 1986. *Science* 232:341–47
14. Berman AL, Kolker E, Trifonov EN. 1994. *Proc. Natl. Acad. Sci. USA* 91:4044–47
15. Gerstein M. 1998. *Fold. Des.* 3:497–512
16. Xu D, Nussinov R. 1997. *Fold. Des.* 3:11–17
17. Doolittle RF, Bork P. 1993. *Sci. Am.* 268:50–56
18. Doolittle RF. 1995. *Annu. Rev. Biochem.* 64:287–314
19. Kent SBH. 1988. *Annu. Rev. Biochem.* 57:957–84
20. Miller M, Baca M, Rao JKM, Kent S. 1998. *J. Mol. Struct.* (THEOCHEM) 423:137–52
21. Ratner L. 1993. *Perspect. Drug Dis. Des.* 1:3–22
22. Schneider J, Kent SBH. 1988. *Cell* 54: 363–68
23. Wlodawer A, Miller M, Jaskolski M, Sathyanarayana BK, Baldwin E, et al. 1989. *Science* 245:616–21
24. Navia M, Fitzgerald PM, McKeever BM, Leu CT, Heimbach JC, et al. 1989. *Nature* 337:615–20

25. Miller M, Schneider J, Sathyanarayana BK, Toth MV, Marshall GR, et al. 1989. *Science* 246:1149–52
26. Swain AL, Miller MM, Green J, Rich DH, Schneider J, et al. 1990. *Proc. Natl. Acad. Sci. USA* 87:8805–9
27. Jaskolski M, Tomasselli AG, Sawyer TK, Staples DG, Heinrikson RL, et al. 1991. *Biochemistry* 30:1600–9
28. Lam PYS, Jadhav PK, Eyermann CJ, Hodge CN, Ru Y, et al. 1994. *Science* 263:380–84
29. Coffin J.1995. *Science* 267:483–89
30. Sieber P, Eisler K, Kamber B, Riniker B, Rittel W, et al. 1978. *Hoppe-Selyler's Z. Physiol. Chem.* 359:113–23
31. Akaji K, Fujii N, Yajima H, Hayashi K, Mizuta K, et al. 1985. *Chem. Pharm. Bull.* 33:184–201
32. Zawadzke LE, Berg JM. 1993. *Proteins* 16:301–5
33. Inui T, Bodi J, Kubo S, Hishio H, Kimura T, et al. 1996. *J. Pept. Sci.* 2:28–39
34. Hojo H, Kwon Y, Kakuta Y, Tsuda S, Tanaka I, et al. 1993. *Bull. Chem. Soc. Jpn.* 66:2700–6
35. Woo DD-L, Clarke-Lewis I, Chait BT, Kent SBH. 1989. *Protein Eng.* 3:29–37
36. Sakakibara S. 1995. *Biopolymers* 37:17–28
37. Kiyam S, Fujii N, Yajima H, Moriga M, Takagi A. 1984. *Int. J. Pept. Protein Res.* 23:174–86
38. Gatos D, Athanassopoulos P, Tzavara C, Barlos K. 1999. In *Peptides 1998*, ed. S Bajusz, F Hudecz, pp. 146–48. Budapest, Hungary: Akademiai Kiado
39. Clark-Lewis I, Kent SBH. 1989. In *The Use of HPLC in Protein Purification and Characterization*, ed. AR Kerlavage, pp. 43–79. New York: Liss
40. Chait BT, Kent SBH. 1992. *Science* 257:1885–94
41. Blake J, Li CH. 1981. *Proc. Natl. Acad. Sci. USA* 78:4055–58
42. Yamashiro D, Li CH. 1988. *Int. J. Pept. Protein Res.* 31:322–34
43. Kemp DS, Carey RI. 1993. *J. Org. Chem.* 58:2216–22
44. Hojo H, Aimoto S. 1991. *Bull. Chem. Soc. Jpn.* 64:111–17
45. Chang TK, Jackson DY, Burnier JP, Wells JA. 1994. *Proc. Natl. Acad. Sci. USA* 91: 12544–48
46. Braisted AC, Judice JK, Wells JA. 1997. *Methods Enzymol.* 289:298–313
47. Jackson DY, Burnier J, Quan G, Stanley M, Tom J, Wells JA. 1994. *Science* 266:243–47
48. Stephen BH, Kent D, Alewood P, Alewood M, Baca A, et al. 1992. In *Innovation and Perspectives in Solid Phase Synthesis*, ed. R Epton, pp. 1–22. Andover, UK: Intercept
49. Vogel AI. 1989. *Textbook of Practical Organic Chemistry,* p. 13. Reading, MA: Addison-Wesley
50. Schnolzer M, Kent SBH. 1992. *Science* 256:221–25
51. Muir TW, Kent SBH. 1993. *Curr. Opin. Biotechnol.* 4:420–27
52. deLisle-Milton R, Milton SCF, Schnolzer M, Kent SBH. 1993. In *Techniques in Protein Chemistry IV*, ed. RH Angeletti, pp. 257–67. New York: Academic
53. Rose K. 1994. *J. Am. Chem. Soc.* 116: 30–33
54. Dawson PE, Kent SBH. 1993. *J. Am. Chem. Soc.* 115:7263–66
55. Rau HK, Haehnel W. 1998. *J. Am. Chem. Soc.* 120:468–76
56. Baca M, Kent SBH.1993. *Proc. Natl. Acad. Sci. USA* 90:11638–42
57. Canne LE, Ferré-D'Amare AR, Burley SK, Kent SBH. 1995. *J. Am. Chem. Soc.* 117:2998–3007
58. Muir TW, Williams MJ, Ginsberg MH, Kent SBH. 1994. *Biochemistry* 33:7701–8
59. Englebretsen DR, Garnham BG, Bergman DA, Alewood PF. 1995. *Tetrahedron Lett.* 36:8871–74
60. Liu CF, Tam JP. 1994. *J. Am. Chem. Soc.* 116:4149–53

61. Liu CF, Rao C, Tam JP. 1996. *J. Am. Chem. Soc.* 118:307–12
62. Dawson PE, Muir TW, Clark-Lewis I, Kent SBH. 1994. *Science* 266:776–79
63. Dawson PE, Churchill MJ, Ghadiri MR, Kent SBH. 1997. *J. Am. Chem. Soc.* 119:4325–29
64. Hackeng TM, Dawson PE, Griffin JH, Kent SBH. 1997. *Proc. Natl. Acad. Sci. USA* 94:7845–50
65. Hackeng TM, Griffin JH, Dawson PE. 1999. *Proc. Natl. Acad. Sci. USA* 96:10068–73
66. Xu R, Ayers B, Cowburn D, Muir TW. 1999. *Proc. Natl. Acad. Sci. USA* 96:388–93
67. Lu W, Randal M, Kossiakoff A, Kent SBH. 1999. *Chem. Biol.* 6:419–27
68. Lu WY, Qasim MA, Kent SBH. 1996. *J. Am. Chem. Soc.* 118:8518–23
69. Storer AC, Menard R. 1994. *Methods Enzymol.* 244:486–500
70. Scheffner M, Smith S, Jentsch S. 1998. In *Ubiquitin and the Biology of the Cell,* ed. J-M Peters, JR Harris, D Finley, pp. 65–98. New York: Plenum
71. Stachelhaus T, Mootz HD, Marahiel MA. 1999. *Chem. Biol.* 6:493–505
72. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. 1989. In *Molecular Biology of the Cell*, pp. 1035–36. London: Garland
73. Paulus H. 2000. *Annu. Rev. Biochem.* 69:447–96.
74. Xu M-Q, Perler FB. 1996. *EMBO J.* 15:5146–53
75. Tam JP, Yu Q. 1998. *Biopolymers* 46:319–27
76. Beligere GS, Dawson PE. 1999. *J. Am. Chem. Soc.* 121:6332–33
77. Wallace CJA. 1995. *Curr. Opin. Biotechnol.* 6:403–10
78. Wilken J, Kent SBH. 1998. *Curr. Opin. Biotechnol.* 9:412–26
79. Wilken J, Hoover D, Thompson DA, Barlow PN, McSparron H, et al. 1999. *Chem. Biol.* 6:43–51
80. Bolin KA, Anderson DJ, Trulson JA, Gantz I, Thompson DA, et al. 1999. *FEBS Lett.* 451:125–31
81. Dealwis C, Fernandez EJ, Thompson DA, Simon RJ, Siani MA, Lolis E. 1998. *Proc. Natl. Acad. Sci. USA* 95:6941–46
81a. Fitzgerald PMD, McKeever BM, VanMiddlesworth JF, Springer JP, James P, et al. 1990. *J. Biol. Chem.* 265(24):14209–19
82. Fitzgerald MC, Chernushevich I, Standing KG, Kent SBH, Whitman CP. 1995. *J. Am. Chem. Soc.* 117:11075–80
82a. Beligere GS, Dawson PE. 1999. *Biopolymers* 51:363–69
83. Sigler PB, Xu Z, Rye HS, Burston SB, Fenton WA, Horwich AL. 1998. *Annu. Rev. Biochem.* 67:581–608
84. Low DW, Hill MG. 1998. *J. Am. Chem. Soc.* 120:11536–37
85. Baca M, Muir TW, Schnolzer M, Kent SBH. 1995. *J. Am. Chem. Soc.* 117:1881–87
86. Goodman M, Felix A, Moroda L, Toniolo C, eds. 2000. *Houben-Weyl: Methods of Organic Chemistry*, Vol. E 22: *Synthesis of Peptides and Peptidomimetics*, Stuttgart, Ger.: Thieme. In press
87. Baca M, Alewood PF, Kent SBH. 1993. *Protein Sci.* 2:1085–91
88. Nagai U, Sato K. 1985. *Tetrahedron Lett.* 26:647–50
89. Sydor JR, Herrmann C, Kent SBH, Goody RS, Engelhard M. 1999. *Proc. Natl. Acad. Sci. USA* 96:7865–70
89a. Kochendoerfer GG, Hunter CL. 1999. In *Peptides 1998*, ed. S. Bajusz, F. Hudecz, pp. 182–83. Budapest, Hungary: Akademiai Kiado
90. Davies DR. 1990. *Annu. Rev. Biophys. Biophys. Chem.* 19:189–215
91. Lu W, Qasim MA, Laskowski M Jr, Kent SBH. 1997. *Biochemistry* 36:673–79
92. Bark SJ, Kent SBH. 1999. *FEBS Lett.* 460:67–76
93. Mosier DE, Picchio GR, Gulizia RJ, Sabbe R, Poignard P, et al. 1999. *J. Virol.* 73:3544–50

94. Mack M, Luckow B, Nelson PJ, Cihak J, Simmons G, et al. 1998. *J. Exp. Med.* 187:1215–24
95. The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*. *Science* 262:2012–18
96. Campbell JH, Zlotnik A, Siani MA, Thompson DA, Butcher EC. 1998. *Science* 279:381–84
97. Lu W, Starovasnik MA, Kent SBH. 1998. *FEBS Lett.* 429:31–35
98. Shao W, Fernandez E, Wilken J, Thompson DA, Siani MA, et al. 1998. *FEBS Lett.* 441:77–82
99. Smith R, Brereton IM, Chai RY, Kent SBH. 1996. *Nat. Struct. Biol.* 3: 946–50
100. Hartsuck JA, Tang J. 1972. *J. Biol. Chem.* 247:2575–80
101. Pennisi E. 1998. *Science* 279:978–79
102. Erlanson DA, Chytil M, Verdine GL. 1996. *Chem. Biol.* 3:981–91
103. Perler FB, Xu MQ, Paulus H. 1997. *Curr. Opin. Chem. Biol.* 1:292–99
104. Muir TW, Sondhi D, Cole PA. 1998. *Proc. Natl. Acad. Sci. USA* 95:6705–10
105. Kinsland C, Taylor SV, Kelleher NL, Mclafferty FW, Begley TP. 1998. *Protein Sci.* 7:1839–42
106. Roy RS, Allen O, Walsh CT. 1999. *Chem. Biol.* 6:789–99
107. Evans TC, Benner J, Xu M-Q. 1998. *Protein Sci.* 7:2256–64
108. Otomo T, Ito N, Kyogoku Y, Yamazaki T. 1999. *Biochemistry* 38(49):16040–44
109. Peters R, Sikorski R. 1998. *Science* 281:367–68
110. Cotton GJ, Muir TW. 1999. *Chem. Biol.* 6:R247–56
111. Bunin BA. 1998. *The Combinatorial Index.* New York: Academic. 322 pp
112. Canne LE, Botti P, Simon RJ, Chen Y, Dennis EA, Kent SBH. 1999. *J. Am. Chem. Soc.* 121:8720–27
112a. Brik A, Keinan E, Dawson PE. 2000. *J. Org. Chem.* 65: In press
113. Wallin E, von Heijne G. 1998. *Protein Sci.* 7:1029–38
114. England PM, Zhang Y, Dougherty DA, Lester HA. 1999. *Cell* 96:89–98
115. Kochendoerfer GG, Salom D, Lear JD, Wilk-Orescan R, Kent SBH, DeGrado WF. 1999. *Biochemistry* 38:11905–13
116. Shin Y, Winans KA, Backes BJ, Kent SBH, Ellman JA, Bertozzi CR. 1999. *J. Am. Chem. Soc.* 121:11684–89
117. Lemieux GA, Bertozzi CR. 1998. *Trends Biotechnol.* 16:506–13
118. Tsujimura A, Yasojima K, Kuboki Y, Suzuki A, Ueno N, et al. 1995. *Biochem. Biophys. Res. Commun.* 214:432–39
119. Cai K, Klein-Seetharaman J, Farrens D, Zhang C, Altenbach C, et al. 1999. *Biochemistry* 38:7925–30
120. Langen R, Cai K, Altenbach C, Khorana HG, Hubbell WL. 1999. *Biochemistry* 38:7918–24
121. Canne LE, Bark SJ, Kent SBH. 1996. *J. Am. Chem. Soc.* 118:5891–96
121a. Offer J, Dawson PE. 2000. *Org. Lett.* 2:23–6
122. Severinov K, Muir TW. 1998. *J. Biol. Chem.* 273:16205–9
123. Fitzgerald MC, Kent SBH. 1998. In *Bioorganic Chemistry: Peptides and Proteins*, ed. SM Hecht, pp. 65–99. New York: Oxford Univ. Press
124. Ingenito R, Bianchi E, Fattori D, Pessi A. 1999. *J. Am. Chem. Soc.* 121:11369–74
125. Wilson EO. 1998. In *Consilience: the Unity of Knowledge*, pp. 99–100. New York: Knopf