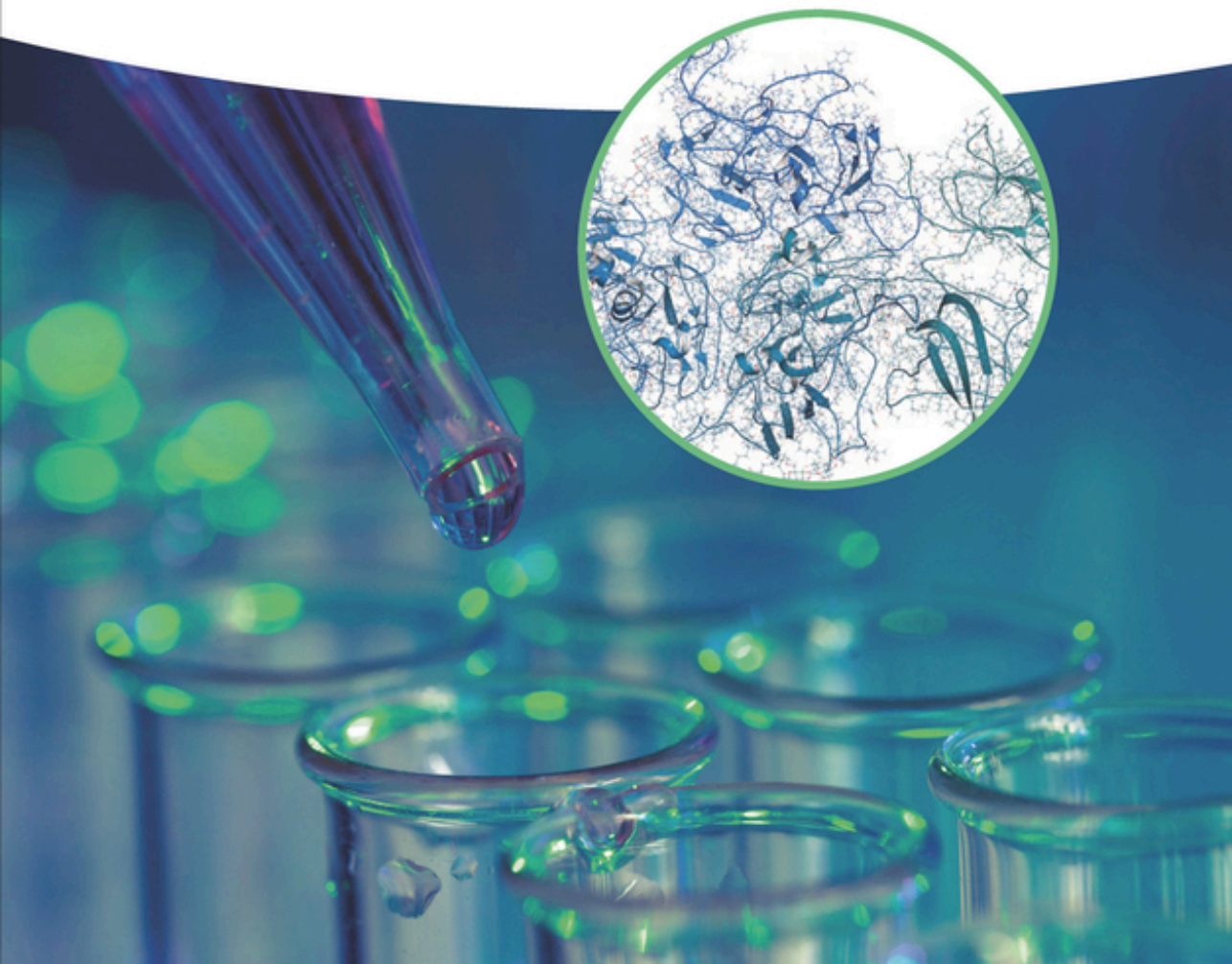
WILEY-VCH

Edited by Ashraf Brik, Philip Dawson, Lei Liu

Total Chemical Synthesis of Proteins



Total Chemical Synthesis of Proteins

Total Chemical Synthesis of Proteins

Edited by
Ashraf Brik, Philip Dawson, and Lei Liu

WILEY-VCH

Editors

Ashraf Brik
Technion - Israel Inst. of Technology
Schulich Faculty of Chemistry
Technion City
32000 Haifa
Israel

Philip Dawson
Department of Chemistry
The Scripps Research Institute
10550 N. Torrey Pines Road
United States

Lei Liu
Tsinghua University
Department of Chemistry
Hetian Building
100084 Beijing
China

Library of Congress Card No.:
applied for

British Library Cataloguing-in-Publication Data:
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

Print ISBN: 978-3-527-34660-8
ePDF ISBN: 978-3-527-82357-4
ePub ISBN: 978-3-527-82358-1
oBook ISBN: 978-3-527-82356-7

Typesetting SPi Global, Chennai, India

Printed on acid-free paper

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Chemical protein synthesis enables the generation of proteins of any architecture and facilitates the site-selective introduction of unnatural amino acids, biophysical probes, posttranslational modifications, and other functionalities. The preparation of these protein analogues enables important biochemical and biophysical studies and contributes to the fundamental understanding of proteins. In this book, we present a salient collection of articles that cover the modern methods and applications of the field, along with descriptions of various case studies in which chemical protein synthesis is leveraged to shed light on fundamental questions in protein science.

In Chapter 1, Stephen Kent, the pioneer and leading figure in modern chemical protein synthesis, introduces the essential role of chemical protein synthesis in developing a fundamental understanding of the principles that give rise to the structures and biological functions of protein molecules. Strong emphasis was given to analytical control and documentation of the synthetic process in order to meet standards, similar to those used to characterize small organic molecules. Characterization of a synthetic protein molecule should rigorously characterize the molecular homogeneity, correct covalent structure, and defined folded structure.

New exciting advances in synthetic methodology have greatly reshaped the landscape of chemical protein synthesis over the past 10 years. In Chapter 2, Bradley L. Pentelute et al. describe the new technology of automated flow-based solid-phase peptide synthesizer (AFPS) that can incorporate each amino acid residue in as little as 40 seconds. This technology offers a template for accelerating syntheses of long polypeptide sequences using an automated flow instrument. This approach may open the chemical biology field to routine manufacture of custom-made, fully synthetic proteins. In Chapter 3, Vangelis Agouridas, Oleg Melnyk, and coworkers describe the development of *N,S(Se)*-acyl shift systems for linking peptide segments together through the most widely used method of chemical protein synthesis, namely, native chemical ligation. This strategy is especially useful for systems that require latent thioesters. Many biologically active proteins have been successfully synthesized with the SEA (bis(2-sulfanylethyl)amido) and SEAlide (*N*-sulfanylethylanilide) chemistries. In Chapter 4, Lei Liu et al. describe the development and optimization of hydrazide-based native chemical ligation, as well as the optimization of the preparation, activation, and ligation of peptide

hydrazides. The hydrazide-based method exhibits several advantages including easy preparation, high stability, good handling properties, and controllable activation. It has been successfully used in the synthesis of peptide pharmaceutics, posttranslationally modified proteins, photo-activatable protein tools, membrane proteins, and mirror-image proteins.

To further expand the scope and utility of native chemical ligation, Richard J. Payne et al. describe in Chapter 5 the efforts made to develop various thiolated amino acids, which, with the help of mild, radical-mediated desulfurization conditions, can enable ligation at 16 of the 20 proteinogenic amino acids. They also describe their seminal studies on selenocysteine (Sec), which have enabled rapid chemical protein synthesis through the diselenide–selenoester ligation (DSL) and novel strategies such as "selenol ligation auxiliary." In Chapter 6, Suwei Dong et al. describe the recent development on new strategies that aim to facilitate peptide ligations at sterically demanding sites. These ligation reactions are usually challenging, as the strong activating conditions may result in competitive hydrolysis. Furthermore, in Chapter 7, Michael S. Kay et al. describe the recent studies on how to control segment solubility in large protein synthesis. Various strategies that utilize solvent choice, isoacyl dipeptides, and semipermanent solubilizing tags are surveyed and analyzed in depth. This chapter provides very useful guides for researchers to overcome the often-encountered solubility problems in chemical protein synthesis.

With the advent of fast peptide synthesis and ligation, other limiting factors have received attentions for chemical protein synthesis. In Chapter 8, Oliver Seitz et al. describe the current state of the art in the development of methods that allow minimizing the number of high-performance liquid chromatograpohy (HPLC) purification steps. These methods typically rely on rapid and HPLC-free assembly of proteins using capture and release purification handles, opening opportunities for the chemical synthesis of protein arrays under parallel conditions. In Chapter 9, Vincent Aucagne et al. describe solid-phase peptide chemical ligation (SPCL), which entails a ligation-based assembly on a solid support to avoid the laborious intermediate chromatographic purifications. The development of new polymers, new linkers, new ligations reactions with improved kinetics, and new masking/unmasking strategies has broadened the scope of SPCL to more ambitious targets. On a different front, to expand the flexibility of peptide condensation, ligation strategies other than native chemical ligation need to be developed. In Chapter 10, Xuechen Li et al. describe the approach of serine/threonine ligation (STL) between peptide C-terminal salicylaldehyde esters and N-terminal Ser/Thr residues. This approach takes advantage of the high abundance of Ser/Thr residues in the protein sequence and has paved the way for broad applications in proteins synthesis and related constructs.

Modern total chemical synthesis of proteins can now reach approximately 200–400 amino acids. To amplify the ability to make proteins containing unnatural amino acids and posttranslational modifications, an important strategy is protein semisynthesis where a biologically produced protein or protein fragment is selectively modified through covalent chemical reactions. In Chapter 11, Philip A. Cole et al. describe the development of methods for protein semisynthesis including expressed protein ligation, cysteine modification reactions, and enzyme-catalyzed

protein/peptide ligations. These methods have been successfully employed to augment our understanding on protein posttranslational modification. As a unique example of protein semisynthesis methods, in Chapter 12, Tilman M. Hackeng et al. describe bio-orthogonal imine chemistry for protein modification. Details were provided for how to incorporate carbonyl or α-nucleophiles moieties into proteins, and how to carry out different imine chemistries such as oxime and hydrazone ligation.

Next are the applications of chemical protein synthesis to solve various biochemical, biophysical, and biomedical problems. In Chapter 13, Vladimir Torbeev describes how to use chemical protein synthesis to decipher the mechanism of protein folding. Many interesting methods have been developed including modification of protein backbone amides, insertion of β-turn mimetics, inversion of chiral centers in protein backbone and side chains, modulating cis–trans proline isomerization, covalent tethering to facilitate folding of designed proteins, and foldamers and foldamer–peptide hybrids. These methods facilitate biophysical studies on intrinsically disordered proteins (IDPs) that perform important functions such as gene transcription and chromatin remodeling. They are also helpful to studies on improper protein folding related to many diseases such as Alzheimer's and Parkinson's disorders.

One of the most exciting applications of chemical protein synthesis is the development of chemical tools to study the biological functions and mechanisms of proteins bearing posttranslational modifications. In Chapter 14, Ashraf Brik et al. described their contributions in developing innovative methods to construct a variety of ubiquitin conjugates of high purity, in sufficient quantity to facilitate detailed biophysical and biochemical analysis. As a "game changer" to overcome the inherit limitations of the enzymatic approaches, chemical protein synthesis has enabled readily construction of versatile ubiquitin conjugates to study and target ubiquitin-processing enzymes. In Chapter 15 Yasuhiro Kajihara et al. describe the cutting-edge examples of glycoprotein synthesis employing various synthetic methodologies and ligation strategies. The combination of both chemical and enzymatic methodologies gives insight into how glycans function in their biological environment and may eventually lead to homogeneous glycoprotein pharmaceuticals.

In addition to protein posttranslational modification, chemical protein synthesis also plays important roles in the biochemical studies on some unique protein families. In Chapter 16, Christian Becker et al. delineate the chemical synthesis of membrane proteins, a unique protein family comprising about 30% of the proteins encoded by the human genome. Access to membrane proteins by chemical synthesis has helped to decipher important aspects of membrane protein functions due to the unique opportunities that chemical synthesis provides in terms of manipulations of the backbone and side chains of proteins. In Chapter 17, Norman Metanis et al. describe chemical synthesis of selenoproteins, which contain the rare coded amino acid selenocysteine. Studies on synthetic selenoproteins facilitate elucidation of the role of selenium in both natural and unnatural processes, providing increasing insights into this previously mysterious micronutrient's role in human health and disease. Furthermore, in Chapter 18, Champak Chatterjee describe the chemical

synthesis of histones bearing various modifications. Due in large part to the addition of semisynthetic and fully synthetic site-specifically modified histones to the repertoire of tools available to investigate the histone code hypothesis, our understanding of the many mechanistic roles for histones has rapidly expanded in the past 10 years.

In the last chapters, W. Seth Horne et al. describe in Chapter 19 how to apply chemical synthesis to engineer protein backbone connectivity. Studies in this area enable exploration of the structural and functional consequences of protein backbone alteration and open the door to new bio-inspired entities with myriad potential applications. In Chapter 20, Christian Hackenberger et al. describe how to synthesize unusually phosphorylated peptides and proteins for proteomic research. Details are provided on mimics of endogenous phosphate esters as well as rare, naturally occurring phosphorylations of functional amino acids such as cysteine, lysine, histidine, and arginine . Finally, in Chapter 21, David J. Craik et al. describe how to synthesize cyclic peptides via ligation methods. Target molecules include orbitides, paws-derived peptides, cyclic conotoxins, θ-defensin, and cyclotides. These cyclic peptides have been a topic of considerable interest in recent years due to their unique structural and pharmacological features that make them excellent starting points in drug design.

We believe that the 21 chapters exemplifies the multidisciplinary nature of research in the field of chemical protein synthesis in the twenty-first century. The readers of the book will be exposed to the state-of-the-art chemistry that the field has been developing through the last few decades and the remaining challenges remained to be tackled. The book will also be very useful source for students and scientists as well to learn about the various synthetic aspects of the peptide fragments synthesis, peptide ligation based on different strategies.

1

Characterization of Protein Molecules Prepared by Total Chemical Synthesis

Stephen B. H. Kent

Department of Chemistry, The University of Chicago, Chicago, IL 60637, USA

> *"Nevertheless, the chemical enigma of Life will not be solved until organic chemistry has mastered another, even more difficult subject, the proteins, in the same way as it has mastered the carbohydrates."*
>
> Source: Emil Fischer (Nobel Lecture 1902)

1.1 Introduction

Proteins are the "natural prodicts" of the twenty-first century. Protein molecules are ubiquitous in the biological world, with numerous diverse functions that range from acting as structural materials such as the keratins, to integral membrane proteins that serve as ion channels or as active molecular transporters in cells, to proteins that act as hormonal messengers in higher animal species, and to proteins that regulate gene expression [1]. The most important function of protein molecules is as enzymes, the potent and specific catalysts of the chemical reactions of biological metabolism, without which life would be impossible [2]. Thanks to modern DNA sequencing methods applied to genome [3] and metagenome [4] sequencing, vast numbers of proteins are being discovered at the nucleic acid level as open reading frames that code for a protein's polypeptide chain.

The central dogma of protein science is that the amino acid sequence of the polypeptide chain encodes the folded structure of the protein molecule in its natural environment, and that it is the folded structure of the protein molecule that gives rise to its biological function(s) [5]. Proteins range in mass from less than 5 kDa to more than 100 kDa. The median size of globular protein molecules is ~35–45 kDa, comprising a polypeptide chain of ~300–400 amino acid residues. Proteins are typically made up of two or more domains, autonomous units of folding, each of ~120–160 amino acid residues [6].

Protein molecules are not simply really big peptides. In its native environment, each globular protein molecule has a defined folded structure that gives rise to the

Total Chemical Synthesis of Proteins, First Edition. Edited by Ashraf Brik, Philip Dawson, and Lei Liu.

functional properties of that protein, including biochemical and biological activities. Synthetic proteins are organic molecules of high molecular mass, comprised of linear polypeptide chains (typically of 50–300 or more amino acid residues) that fold to form complex, dynamic structures. The large size of proteins together with the intricacy of their covalent and folded structures creates special challenges in the characterization of these synthetic molecules. For this reason, it is important for researchers to rigorously characterize protein molecules prepared by total chemical synthesis to meet standards similar to those used in synthetic organic chemistry. Analytical methods and criteria for the rigorous characterization of synthetic proteins have recently been enunciated [7].

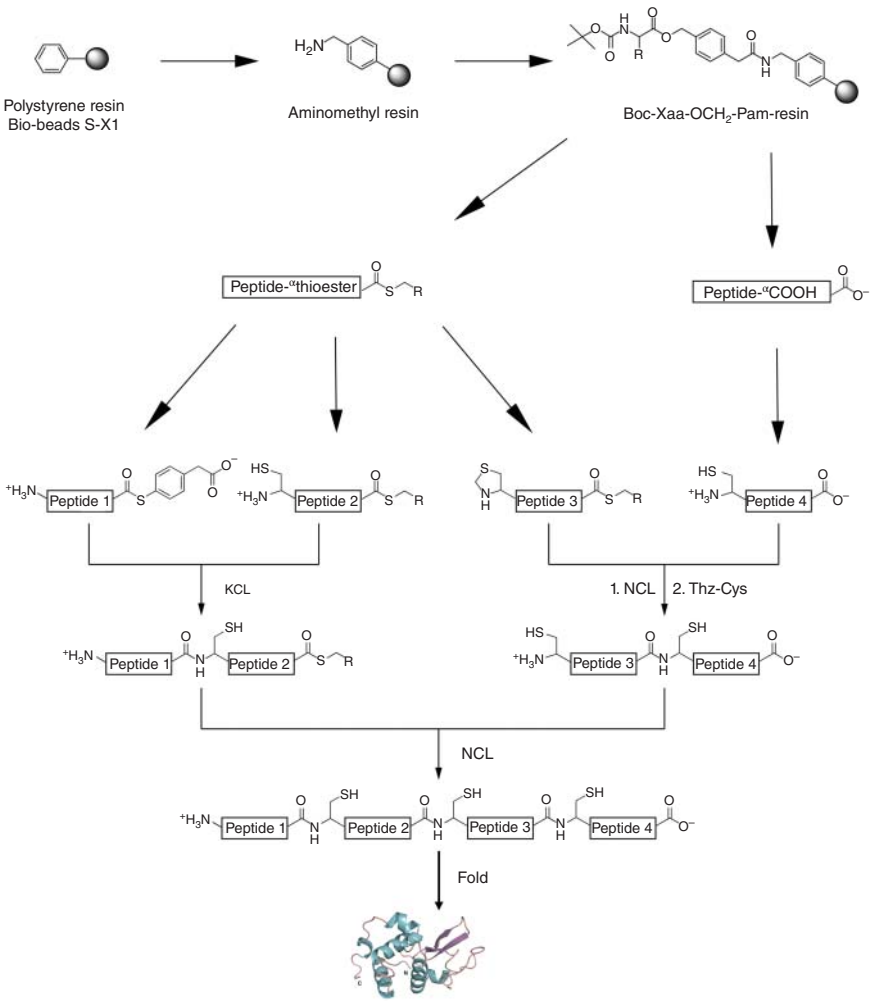
1.2 Chemical Protein Synthesis

> *"The chemical ligation approach .... breaks the conceptual shackles imposed by the peptide bond, frees us from the linear paradigm of the genetic code, and opens the world of proteins to the entire repertoire of chemistry."*
>
> Stephen Kent (23rd European Peptide Symposium 1994)

Chemistry, enabled by total synthesis, has an essential role to play in developing a fundamental understanding of the principles that give rise to the structures and biological functions of protein molecules. Thanks to modern chemical ligation methods [8–10], the total synthesis of protein molecules in the research laboratory is both practical and increasingly robust.

Characterization of a protein prepared by total chemical synthesis should begin with the synthetic process itself. Analytical control of the bond-forming steps, purification and analysis of each synthetic intermediate, and folding of the full-length synthetic polypeptide chain to form a defined tertiary structure are integral to verifying the structure of the final synthetic protein product. The steps involved in a typical total chemical synthesis of a protein molecule are listed below. Key aspects pertinent to **characterization** of the synthetic protein are in **bold**:

- **Establish a verified amino acid sequence of the polypeptide chain of the target protein molecule.**
  - Use a database such as UniProt (www.uniprot.org); include posttranslational processing where that is known to give the mature protein molecule.
  - Resolving data base/literature ambiguities in the reported amino acid sequence of a protein can be problematic. It is easy to make a mistake and to end up making an incorrect target sequence.
- Design a (convergent) synthesis of the protein's polypeptide chain, starting from peptide segments containing fewer than 40–50 amino acid residues (Scheme 1.1).
- Stepwise solid-phase synthesis (SPPS) of peptide segments equipped with suitable functionalities for chemical ligation.
- Boc chemistry [11] or Fmoc chemistry SPPS [12], with **documentation of the amino acid sequence actually made.**



Scheme 1.1 Convergent chemical synthesis of a protein molecule from four synthetic peptide segments prepared by stepwise SPPS. Key: NCL, native chemical ligation; KCL, kinetically controlled ligation; Thz-Cys, conversion of N-terminal thiazolidine-CO– to Cys–. Source: Adapted from Durek et al. [11].

- Purification by preparative high-performance liquid chromatography (HPLC), which should be performed under *displacement mode* [13].
- Combined fractions should be checked by a high resolution technique based on a ***separation principle distinct from that used in the purification***, such as capillary isoelectric focusing (CIEF) [14]. Because HPLC is used for the purification of the synthetic peptide segment, it is NOT sufficient to use analytical HPLC or LCMS run under similar chromatographic conditions to verify its homogeneity.
- *Direct infusio*n electrospray ionization mass spectrometry (ESI-MS) for checking purified peptide segments for impurities that have different masses.

- **Confirmation of the covalent structure of each synthetic peptide segment.**

- **Precise mass measurement**
  - from analytical LCMS, the mass spectrometric data *must* be collected across the *entire UV absorbing peak corresponding to the purified peptide*.
  - a correct mass is a necessary but not sufficient analytical criterion.
- **Amino acid sequence**
  - by MALDI-TOF MS 'ladder sequencing' [15] of terminated byproducts that are invariably present at low-levels in crude synthetic peptides made by SPPS.
  - by MS-MS of the purified peptide. This process is rendered more straightforward because the target sequence of the peptide segment is known.
- Convergent covalent condensation of the peptide segments
  - Chemical conversion of cryptic functional groups to the reactive form suitable for chemical ligation (e.g. Thz–peptide to Cys–peptide; [16] peptide–$CONHNH_2$ to peptide–thioester [17]).
  - Native chemical ligation (NCL); [18] pH 7.0; aqueous 6 M Gu.HCL as a near-universal solvent; ambient temperature; arylthiol catalyst [19].
- **Analytical control of synthetic steps**
  - "Hands-on" real-time analytical LCMS during ligation reactions is an essential tool for following the course of each bond-forming step.
- **Purification and characterization of each intermediate product**
  - Using the same protocols as those used for the synthetic peptide segments.
- **Purification to homogeneity and rigorous characterization of the full-length synthetic polypeptide chain.**
  - Purification by preparative HPLC, which should be performed under *displacement mode* [13].
  - Combined fractions should be checked by a high-resolution technique based on a ***separation principle distinct from that used in the purification***, such as CIEF [14]. Because HPLC has been used for the purification of the synthetic polypeptide, it is NOT sufficient to use analytical HPLC or LCMS run under similar conditions to verify its homogeneity.
  - *Direct infusio*n ESI-MS is also a good technique for checking the full-length synthetic polypeptide chain for impurities that have different masses.
- Folding the synthetic polypeptide chain to form the functional protein molecule
  - For disulfide-containing proteins, standard thiol redox couple conditions are used along with moderate amounts of solubility-enhancing agents ("denaturants") to keep misfolded polypeptide chains in solution so that they too are eventually able to fold correctly [20].
  - For proteins that do not contain disulfide bonds, it is frequently sufficient to simply dilute the polypeptide chain into native buffer or to slowly dialyze from 6 M Gu.HCl into native buffer conditions.
- Analytical control of protein folding
  - In the case of disulfide-containing proteins, the folded protein molecule will usually elute *earlier* than the unfolded polypeptide chain on analytical reverse phase HPLC, because hydrophobic side chains are less exposed in the folded, disulfide cross-linked protein molecule [21].

- For proteins that do not contain disulfide bonds, other techniques such as CD-ORD[1] or multidimensional nuclear magnetic resonance (NMR) must be used to monitor the progress of folding.

- **Purification of the folded synthetic protein to give a single defined molecular species.**
  - By one or more of the techniques of reverse-phase HPLC, ion exchange chromatography, size exclusion chromatography.
- **Homogeneity of the purified synthetic protein must be verified by a high-resolution technique based on a separation principle *distinct* from that used for purification.**
  - CIEF [14] is the preferred technique for rigorously establishing the homogeneity of the purified synthetic protein molecule.
- **Precise measurement of the mass of the synthetic protein molecule by *direct infusion* ESI-MS (direct infusion ESI-MS).**
  - *Direct infusion* ESI-MS provides data representing *all* the molecular species present in the final synthetic product. It thus provides both a precise experimental measurement of the mass of the synthetic protein and at the same time can reveal the presence of any other products that have a mass distinct from that of the target protein molecule. All measured mass data must include an experimental uncertainty (Figure 1.1).
  - HPLC with on-line ESMS can be used. However, impurity protein species present in the synthetic product may not be eluted from the reverse phase support. Mass spectrometric data from LCMS *must* **be acquired over the entire UV-absorbing peak** corresponding to the purified synthetic protein product.
    - ***It is NOT acceptable to report a mass determined at a single time point of the HPLC chromatogram, while ignoring molecular species with different masses that are present at other time points under the same UV absorbing peak.***
  - Wherever possible, the **monoisotopic mass** of a synthetic protein should be reported (together with experimental uncertainty), along with the calculated monoisiopic mass. An example is shown below (Figure 1.2).
- **Verification of the cysteine pairing for disulfide containing proteins.**
  - Proteolytic digestion followed by LCMS peptide mapping [24].
  - Cysteine pairing can also be verified by high resolution X-ray crystallography (see below).
- **Conformational homogeneity of the synthetic protein molecule.**
  - Multidimensional NMR "fingerprinting" can be used to establish that the synthetic protein has a single folded structure [11, 25]. Notes: (i) there are examples of natural protein molecules that form more than one defined folded

1 While a low-resolution technique such as CD-ORD is useful for monitoring the progress of folding, it must be supplemented with high-resolution characterization of the folded structure of the synthetic protein by techniques such as NMR and X-ray crystallography.

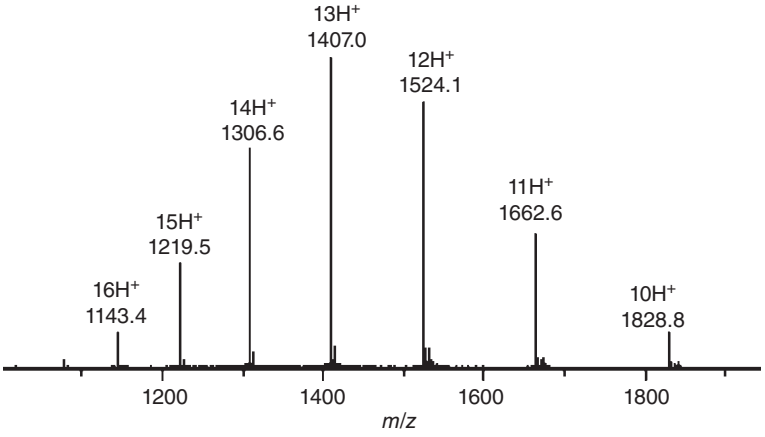


Figure 1.1 Direct infusion ESI-MS data for [Lys[24,38,83]]erythropoietin aglycone prepared by total chemical synthesis. Impurities of mass less than the target protein molecule are essentially absent, as shown by the lack of peaks on the low *m/z* side of each charge state. The peaks on the high *m/z* side of each charge state are Na and Ca ion adducts. Source: Adapted from Liu et al. [22].

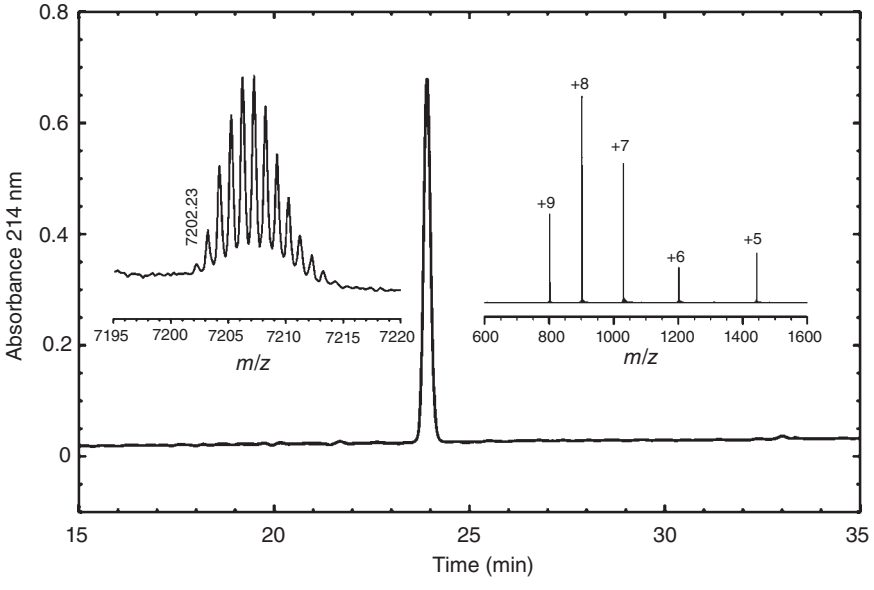


Figure 1.2 Analytical HPLC and MS characterization of a synthetic protein. With modern mass spectrometric instrumentation, the monoisotopic mass can be experimentally determined and compared with the theoretical mass. Source: Adapted from Durek et al. [23].

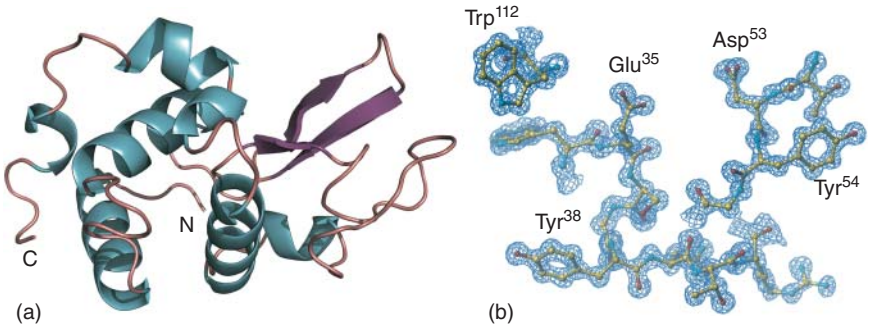


Figure 1.3 X-ray diffraction structure of crystalline human lysozyme prepared by total chemical synthesis. (a) Cartoon representation of the secondary structure of the folded human lysozyme synthetic protein molecule; (b) $2F_o - F_c$ electron density map , showing the quality of the data acquired at a resolution of 1.04 Å. Source: Durek et al. [11]. © 2007 National Academy of Sciences.

structure [26]; (ii) intrinsically disordered proteins only fold, if at all, in the presence of their target molecules [27].

- **Determination of the atomic structure of the folded protein molecule.**
  - X-ray crystallography, including racemic protein crystallography, [28] can be used to determine the structure of the folded, homogeneous synthetic protein molecule [29, 30].
  - X-ray structural data will also confirm the connectivity of any disulfides that are present, and at sufficiently high resolution can reveal the presence of aberrant chemical modifications of the synthetic protein (Figure 1.3).
  - Note that because fractional crystallization from protein mixtures is common, determination of the crystal structure of a synthetic protein by X-ray diffraction does not in itself establish that all of the synthetic protein molecules have that structure and are undamaged. Homogeneity of the synthetic protein molecule must be established *before* determination of the crystal structure.
  - For small proteins, the folded structure can be determined using natural abundance protein NMR techniques [31].
- Biological/biochemical assays.
  - Quantitative assays, especially measurements of the catalytic activity of a synthetic enzyme molecule, can be very informative but are inadequate as proof of homogeneity or correct structure of the synthetic protein.

The total synthesis of the chemokine CCL2 (MCP-1) by Grygiel et al. [32] is a near-perfect example of a properly documented total chemical synthesis of a protein molecule, illustrating essentially all of the synthetic steps described above together with meticulous and complete characterization of the synthetic protein as a single molecular species of defined structure (Figure 1.4).

Selected further examples of well-characterized synthetic proteins can be found in Refs [24, 33–42].

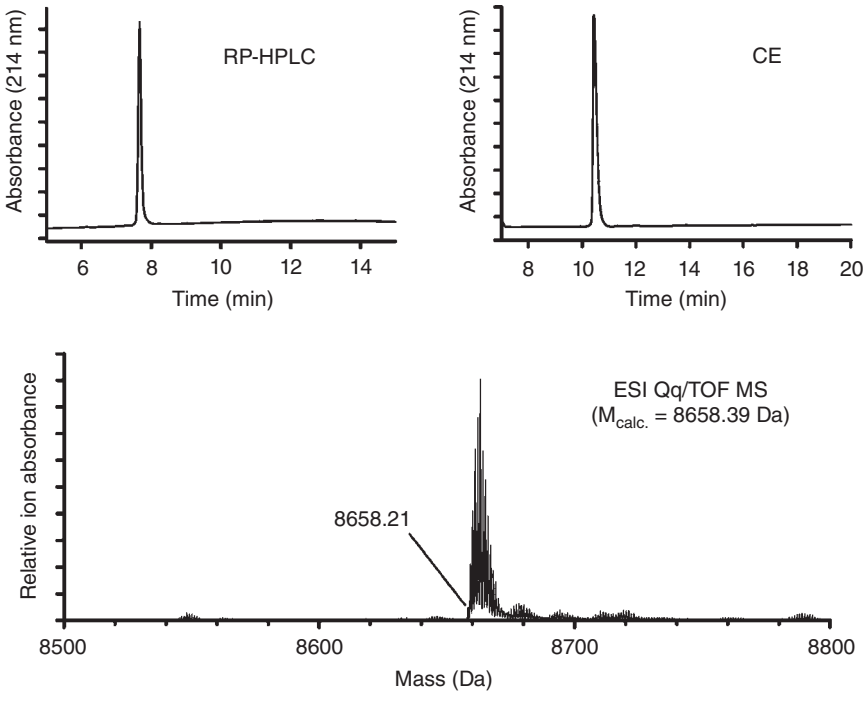


Figure 1.4 Analytical characterization of the protein CCL2 (MCP-1) prepared by total chemical synthesis. [32] Purity of the synthetic protein was verified by reverse phase HPLC and by capillary electrophoresis, two high resolution analytical methods based on different separation principles (hydrophobicity; charge). The monoisotopic mass was measured by LC-ESI Qq/TOF mass spectrometry. In addition to the data shown here, the disulfide bonds are verified by enzymatic digestion and LC-(MS-MS) of the resulting peptide fragments, and the crystal structure of the synthetic protein is determined at a resolution of 1.9 Å. Source: From Grygiel et al. [32].

1.3 Comments on Characterization of Synthetic Protein Molecules

> *"In the field of protein synthesis it is my confident hope that tomorrow's deeds will catch up with today's titles, and that we shall truly be able to obtain enzymatically active proteins as synthetic substances: as materials composed of a single molecular species."*
>
> Josef Rudinger (3rd American Peptide Symposium 1972)

1.3.1 Homogeneity

The single most important aspect of the characterization of a synthetic protein is rigorous verification of its homogeneity, i.e. that the synthetic product is a single