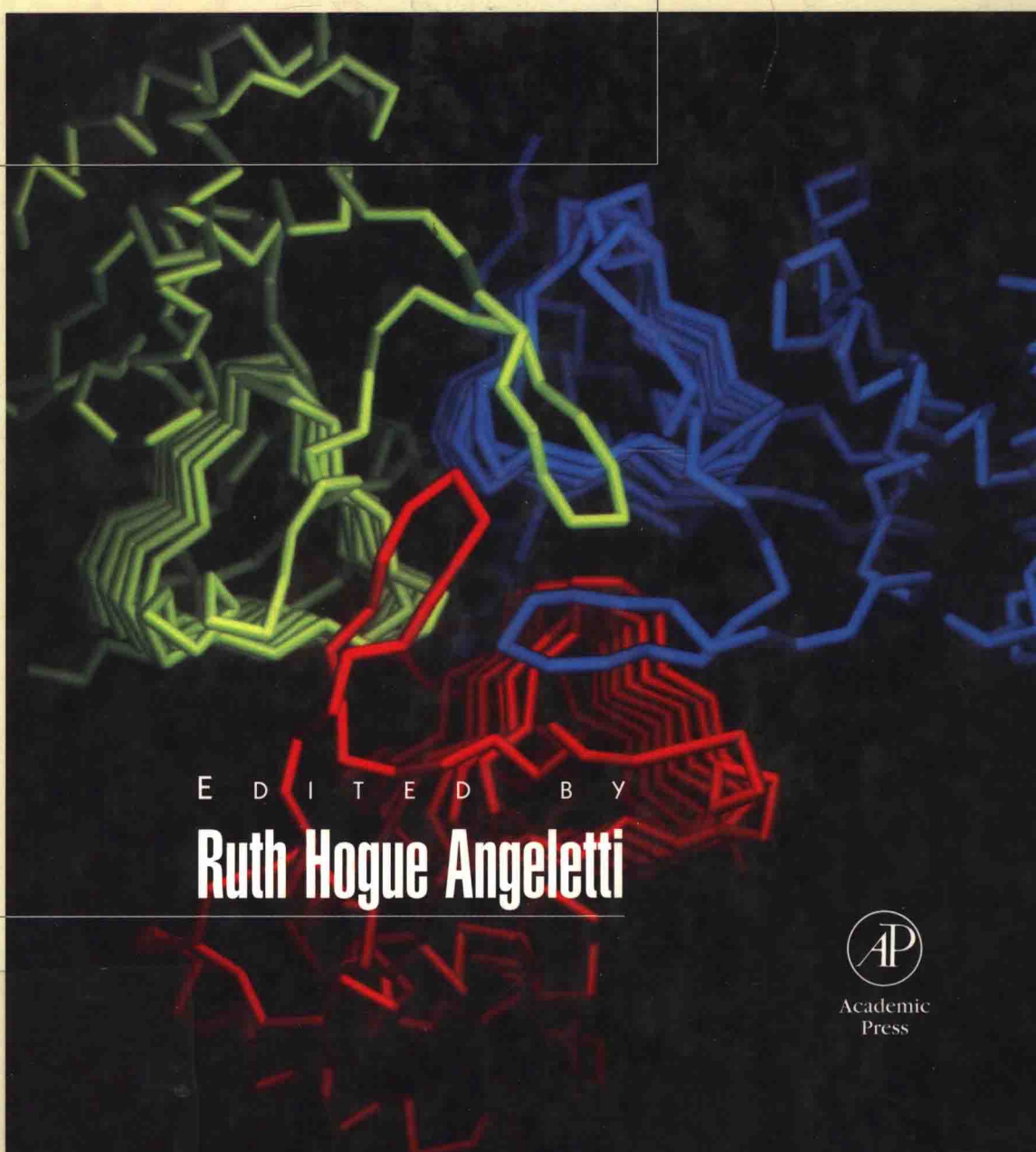
Proteins

Analysis and Design



EDITED BY

Ruth Hogue Angeletti

Academic Press

Proteins
Analysis and Design

Edited by

Ruth Hogue Angeletti

Albert Einstein College of Medicine of Yeshiva University
Bronx, New York

Academic Press

San Diego London Boston New York Sydney Tokyo Toronto

Front cover photograph: Alpha carbon trace of trimeric tetrahydrodipicolinate *N*-succinyltransferase. The triangular domains are the left-handed parallel beta helices formed by tandem-repeated copies of hexapeptide amino acid sequence motif. Courtesy of Todd W. Beaman, David A. Binder, John S. Blanchard, and Steven L. Roderick, Albert Einstein College of Medicine, Bronx, New York.

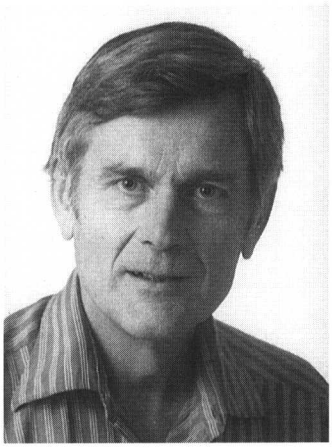
This book is printed on acid-free paper. ♾

Academic Press
a division of Harcourt Brace & Company
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA
http://www.academicpress.com

Academic Press
24–28 Oval Road, London NW1 7DX, UK
http://www.hbuk.co.uk/ap/

Proteins : analysis and design / edited by Ruth Hogue Angeletti.
p. cm.
Includes bibliographical references and index.
ISBN 0-12-058785-8 (hc : alk. paper)
1. Proteins. 2. Proteins--Analysis. 3. Protein engineering.
I. Angeletti, Ruth Hogue.
QP551.P697774 1998
572'.6--dc21 98-4746
CIP

Printed in the United States of America
98 99 00 01 02 MM 9 8 7 6 5 4 3 2 1



Finn Wold

1928–1997

This volume is dedicated to the memory of Finn Wold, who devoted his career to the study of protein biology and to the nurturing of protein science and protein scientists.

Preface

It is now possible to perform our dream experiments with proteins: to analyze details of their structure–function relationships and their life cycles within the cell. Many of us need to embrace these technologies to achieve full understanding of the systems we have chosen to study. Although the experimental approaches are within the reach of all of us, we have not all been trained in their implementation. This book was designed to provide guides, both philosophical and practical, from experts in protein analysis and design.

Ruth Hogue Angeletti

Contents

Contributors

Numbers in parentheses indicate the pages on which the authors' contributions begin.

Ruedi Aebersold (3) Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195

Ruth Hogue Angeletti (1) Albert Einstein College of Medicine of Yeshiva University, Bronx, New York 10461

Gary S. Coombs (259) Howard Hughes Medical Institute, Department of Pharmacology, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75235

David R. Corey (259) Howard Hughes Medical Institute, Department of Pharmacology, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75235

Gregg B. Fields (207) Department of Chemistry and Biochemistry, Center for Molecular Biology and Biotechnology, Florida Atlantic University, Boca Raton, Florida 33431

Michael W. Klemba (313) Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544

Radha Gudepu Krishna (121) Diagnostic Systems Laboratories, Inc., Webster, Texas 77598

Janelle L. Lauer (207) Department of Chemistry and Biochemistry, Center for Molecular Biology and Biotechnology, Florida Atlantic University, Boca Raton, Florida 33431

Mary Munson (313) Department of Microbiology, GBF-National Research Center for Biotechnology, D-38124 Braunschweig, Germany

Lynne Regan (313) Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520

Scott D. Patterson (3) Amgen Inc., Protein Structure, Amgen Center, Thousand Oaks, California 91320

Finn Wold* (121) Department of Biochemistry and Molecular Biology, University of Texas Medical School, Houston, Texas 77225

*Deceased.

Introduction

Ruth Hogue Angeletti

Protein biology is the study of protein structure revealing details of the function and life cycle of individual polypeptides. Completion of the sequences of animal and plant genomes will not be the consummation of modern biology, but a new beginning. The experimental and philosophical challenge is not just to identify polypeptide gene products for all coding genes, but to provide a rationale for how they work. As part of a cellular response to external stimuli, proteins may be altered by chemical modification or change subcellular localization, processing, degradation, or concentration. How do proteins work? How do they work together? How do they work over time and space?

The development of technologies and experimental approaches that were required to answer the questions of protein biology accelerated during the early period of genomic analysis. These technologies are now essential tools of experimental biology. All of the strategies aim toward the highest sensitivity analysis possible. Chromatographic and electrophoretic separation and detection, Edman sequence analysis, amino acid analysis, and mass spectrometry are now all performed routinely with a few picomoles of protein or peptide. Subpicomole and low femtomole levels of analysis are reported in the literature more frequently, and experiments detecting attomoles of protein or peptide have been described. Pushing these barriers of sensitivity is of intense interest because of the desire and need to understand changes that take place in a few molecules within a few cells or a single cell, to study not the population of molecules, but those few relevant to the physiological events at hand.

Covalent modifications of proteins are an essential part of the language of intracellular and intercellular communication. They may be reversible or irreversible. They may be required for biological activity, or simply modulate it, functioning as "molecular switches." They are important for signaling and molecular and cellular recognition. These modifications may impart structural stability or unique structural features, facilitate protein folding, or promote intermolecular interactions. They may anchor a protein in a membrane or determine intracellular or extracellular position. Modifications can alter the biological lifetime of a protein, and define the process by which it is degraded. Whereas some modifications are present on an entire population of a polypeptide, others are present on only a small subset of the protein at any one moment. Quantification of the extent and range of modifications present in this popula-

tion is not routinely analyzed, but could have important biological implications. Thus, knowledge of the possible modifications of a particular protein, as well as modifications induced by external stimuli, are of fundamental importance to research in structural biology and cell biology alike.

The core technology for analyzing protein covalent structure is mass spectrometry. Once the province of specialists, mass spectrometry must now be in the repertoire of all biologists. The ability to measure accurately the molecular weights of both large and small biological molecules means that mass measurements can be used to unambiguously identify a polypeptide or gene product, to establish the presence of a covalent modification and localize its position within the polypeptide chain, and to sequence or verify structures by patterns of molecular bond breakage. The strategies of mass spectrometry provide insight into biological processes and are a key element in the discovery process. Establishing lines of communication between analytical protein chemists and cell biologists will be needed to release the full power of detailed genomic and cellular information.

Protein biology also includes the design of new proteins to test biological and physiological hypotheses. One can now design and prepare novel polypeptides with new functions that cannot be isolated from nature. How can one manipulate proteins to understand their original function or to make them perform new ones? On what principles can these designs be based? Protein analysis and protein design are linked disciplines. In order to understand how a protein functions, one must understand its complete covalent structure, including all cross-linking, processing, and modification events. To design a novel polypeptide, one must know what those modifications are in order to incorporate them into the structure, or to know when they can be ignored. The most creatively designed protein model cannot be used to prove the starting hypothesis if the final structure has not been rigorously analyzed and verified.

Both molecular biological and chemical synthetic procedures have made it possible to alter polypeptide structure to probe details of function. Strategies can be developed to alter a single amino acid, to substitute a nonnatural residue, or to mix and match functional units to create new biological entities. Pragmatic concerns, bioassays, and considerations of secondary structural elements can all be used to build on knowledge and function of known molecules. More challenging still is *de novo* design of new protein activities from first principles, a less developed field. Progress is being made in this direction, and should proceed rapidly as the related fields of protein design mature.

We can no longer be practitioners of a single art or technology, but must take a problem solving, or systems approach, to the biological questions that we ask. These protein strategies are only a part of modern experimental biology, yet are a core mechanism to identify the most important questions and find their solutions.

Chapter 1

Current Problems and Technical Solutions in Protein Biochemistry

Ruedi Aebersold and Scott D. Patterson

I. Introduction

The vast majority of biological processes and pathways are tightly controlled. This applies equally to well-understood, relatively simple processes such as oxygen transport and storage and to as yet molecularly poorly understood phenomena of extreme complexity such as development, cell differentiation, and signal transduction pathways that serve to elicit the appropriate intracellular responses to extracellular stimuli.

Such biological processes differ significantly in their complexity as well as in the level and mechanism of control. They have in common that they involve multiple components, that the activities of at least some components are regulated, and that multiple components structurally and functionally interact. Frequently, key components are proteins and regulatory control is mediated by a variety of mechanisms, including reversible protein modification, most notably protein phosphorylation, formation of protein : protein complexes, proteolytic protein processing, *de novo* synthesis, and intracellular translocation. For the experimental biologist this means that a comprehensive analysis of such a process requires the identification of each system component, the characterization of the identified species with respect to their chemical composition and functional control, and the investigation of the interaction between system components. Finally, the results obtained at these levels should be sufficient to establish a model of the process studied that satisfies all the experimental observations.

The traditional approach to investigating complex systems has been reductionist in nature. Proteins representing a single step in a process were purified and analyzed in great detail. This step was repeated until seemingly all the elements of the pathway were identified, isolated, and studied. Reconstitution of the system from the isolated elements was used as a criterion to indicate that the system had been analyzed in its full complexity or at least to a sufficient degree to support a model of the process. Contraction of striated and smooth muscle are illustrative examples for the success of this approach. Insights from light microscopic, electron microscopic, physiologic, and biochemical studies culminated in the formulation of the sliding filament model for contraction of striated muscle (1). Further biochemical analyses resulted in the identification and structural and functional analysis of the components involved, as well as in the reconstitution of contractile elements from purified proteins *in vitro* (2,3).

Significantly, more than 30 years after formulation of the sliding filament model, substantial structural and regulatory features of muscle contraction remain to be discovered [e.g., (4)]. This suggests that the reductionist approach, although very successful in describing essential events in a complex system, is less suited for a comprehensive systems analysis.

Dramatic advances in several independent research areas indicate that an integrated biotechnology capable of successfully analyzing biological systems in their whole complexity is rapidly emerging. The most dramatic developments include the following:

1. *Initiation of systematic, genome analysis programs.* Large-scale DNA mapping and sequencing efforts, aimed at deciphering the complete genomic sequences of a number of species (5–8) or the expressed sequences represented by cDNAs of differentiated tissues and cells (9), are the most widely publicized and discussed global programs in biology. It is projected that the genome sequence of yeast (*Saccharomyces cerevisiae*) will be completed before this chapter appears in print (6) and that the sequence of the complete human genome may be determined within the next 10 years (8). Although these programs promise to generate a unique resource for research into biology and medicine, the frequent implication that knowledge of genome sequences will be sufficient to achieve understanding of biological systems is simplistic. An array of expressed gene sequences does not describe a biological system, mainly because proteins, the most important biological effector molecules, are not simply the linear translation of gene sequences. Proteins should be considered the products of mRNA translation *and* of diverse posttranslational processing and modification events.

2. *Development of powerful new analytical technology.* In particular, the introduction of mass spectrometric techniques and instruments compatible with the routine analysis of proteins and peptides at high sensitivity now makes investigation of the dynamics of biological systems, including induced protein modifications, reversible protein : protein associations, protein translocations, targeted destruction and *de novo* synthesis experimentally accessible.

3. *Rapid advances in data analysis, data storage, and data distribution.* Centralized, continuously updated, and interconnected databases that are remotely accessible are a necessary prerequisite for integrated systems analysis, and new algorithms for the analysis of data derived from large-scale programs are being developed.

4. *The capability for targeted genetic manipulation of cells and organisms.* In a number of experimental organisms, transgene expression, including the expression of dominant negative mutants, targeted gene knockout, and gene replacement, are now relatively standard procedures. These techniques are indispensable for testing hypotheses related to the structure and function of genes and proteins.

In this chapter we describe recently developed protein analytical techniques and concepts and attempt to demonstrate how these tools converge toward a comprehensive technology for the analysis of complex biological systems. We focus on the analysis of proteins because proteins are the most important effector molecules in biology. To highlight the stunning technological progress during the past few years in this area, we will begin each section with a

case study, describing the experimental approach and the techniques used to successfully analyze a complex, regulated system. As the case study we chose the signal transduction pathway induced by the virus-fighting protein interferon from the cell membrane to the nucleus, which represents one of the best-understood regulatory pathways in cells. Following the case study we will discuss novel technology and approaches and we will illustrate the impact of these new tools on biological research. The sections will, therefore, predominantly focus on discussing challenges and promising approaches to protein analytical problems, rather than on detailed treatment of individual techniques. Each builds on chapters elsewhere in this volume describing mass spectrometry of proteins and peptides, separations technologies, and posttranslational modifications, respectively.

II. Early Events in Signaling by Interferons: From Membrane Receptor to Induced Genes

It is well established that binding of polypeptides such as lymphokines, hormones, neurotransmitters or growth factors to specific cell surface receptors causes changes in the transcription rate of specific sets of genes by transmitting a signal from the cell membrane to the nucleus. At least the early events (transcriptional activation within minutes of membrane receptor stimulation) are based on the activation of preexisting transcription factors by a variety of mechanisms, including protein phosphorylation, protein : protein complex formation, intracellular translocation, and proteolysis. The involvement of multiple protein components and the tight regulation of function define such signal transduction pathways as complex biological systems.

The interferons are a family of secreted polypeptides (cytokines) that act as mediators in the host defense against viruses and parasites. As with other cytokines, the interaction of interferons with their specific membrane receptor induces rapid transcription of a set of immediate-early genes in the absence of protein synthesis (10). Over the past 10 years, studies from a number of groups converged on establishing the complete sequence of events required to transduce the signal from the membrane to the nucleus, making this the most completely understood signal transduction pathway in mammalian cells (11).

The work began with the identification of several genes that were transcriptionally upregulated in response to interferon stimulation of cells (12,13). Detailed analysis of the 5′ upstream genomic sequence of these genes identified the interferon-stimulated response element (ISRE) (14–18) as the binding site occupied after interferon α (IFNα) stimulation, presumably by a transcriptional activator (17). One of the factors binding to the regulatory regions was termed interferon-stimulated gene factor 3 (ISGF3). This factor was the only one to show several prerequisite characteristics as an interferon α activated transcription factor, including absolute dependence on interferon treatment (13,19), rapid increase in activity without protein synthesis (16,17) and specific binding to the nucleotide sequence required for interferon α dependent gene activation (18,20). Subsequent work led to the isolation of the ISGF3 complex (21), sequence analysis of the respective proteins and coding genes (22–24), and elucidation of the sequence of events leading to the interferon-induced activation of ISGF3

(25,26), and finally, to the establishment of a molecular and functional connection between interferon receptor stimulation and transcriptional activation of interferon-induced genes (27).

The model now accepted suggests that binding of interferon to its cognate receptor activates a nonreceptor tyrosine kinase of the family of Janus kinases (JAKs), which is noncovalently associated with the cytoplasmic tail of the receptor (see Fig. 1). The tyrosine kinase, activated through ligand–receptor interaction, phosphorylates on tyrosine residues one or more of three large polypeptides of 113 kDa, 91 kDa, and 84 kDa that preexist in the cytoplasm, causing them to aggregate and translocate to the nucleus, where they interact with a smaller protein of 48 kDa. The now complete ISGF3 protein complex forms the transcription factor that activates immediate-early gene transcription. By convention, the latent cytoplasmic transcription factors have been termed STATs (signal transducers and activators of transcription), whereby the 91- and 84-kDa proteins are named Stat 1α and Stat 1β, respectively, and the 113-kDa polypeptide is named Stat 2. In this chapter we have chosen this pathway (shown in Fig. 1), which has been described in more detail in several excellent reviews (27–29), as a prototypical example for a complex biological system for the following reasons: (i) Transduction of the interferon signal is currently the most completely understood signal transduction pathway. There is a direct and complete link from the membrane receptor to the nuclear factors transactivating immediate-early genes. (ii) The type of regulatory events controlling this pathway, including protein phosphorylation, translocation, and protein : protein complex formation are common to the regulation of many, if not most, complex biological systems. (iii) Most of the elements involved in this pathway, as well as the mechanisms controlling the activity, were determined using protein

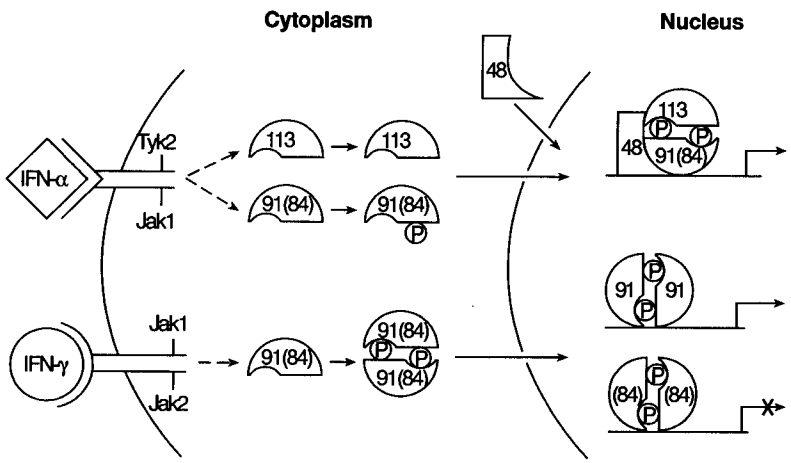


Fig. 1 Overview of the proteins identified in interferon α- and γ-dependent signal transduction and gene activation. The Jak kinases are phosphorylated on tyrosine in response to ligand, but the sites and the requirement for such modification are not yet established. The circled P's on the Stat proteins are tyrosine phosphates, and the indentations symbolize SH2 domains. Reprinted with permission from Darnell, J. E., Jr., I. M. Kerr, and G. R. Stark. 1994. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* 264:1415–1421. (27). Copyright 1994 American Association for the Advancement of Science.

analytical technology, thus providing a framework to discuss novel technology in comparison with accepted, standard techniques. Finally, (iv) it has become apparent that at least some of the elements involved in the interferon signal transduction pathway and proteins structurally and functionally related to the Jak and Stat proteins constitute a network of intracellular signaling molecules that are collectively involved in transmitting the signal from at least 25 different membrane receptors (29). This suggests that signal transduction pathways initiated by different receptors (in the same cell) are not independently controlled, but are interwoven to an extent that is only slowly becoming apparent. This illustrates the premise of this chapter that comprehensive investigation of complex biological systems requires new analytical technologies, which are better suited to achieving a global perspective.

III. Detection of Systems Components

A. Overview

Depending on the complexity of the organism, a cell contains thousands to tens of thousands of different types of proteins. Collectively, these proteins constitute the structure and perform and control numerous cellular functions. Establishment of a complete inventory of all those proteins that are involved in a particular biological process is complicated by the fact that usually neither the number, nor the identities, nor the activities of the elements of the process are known. A credible model for a process cannot be formulated, however, as long as the components and their functions are not known. In this section we describe experimental approaches for the detection of those proteins that are part of a particular process. The discussion is limited to approaches for the detection of proteins with known or unknown function that are part of a process. DNA-based approaches such as subtractive hybridization and cDNA profiling using ordered DNA arrays are not further discussed (30,31).

B. ISGF3 Case Study: Detection of Proteins Involved in Interferon Signal Transduction

Detection of the components of the ISGF3 complex was essentially based on the observation that an interferon induced transcription factor was binding to specific DNA sequences in the promoter regions of interferon-regulated genes (Fig. 2A). This provided a bioassay for the detection and purification of the factor. Comparative gel-shift assays from interferon stimulated and unstimulated control cells showed that two of the three observed protein : oligonucleotide complexes were specifically induced by interferon stimulation of the cells (17). The more rapidly induced of the two complexes, which did not require protein synthesis, was furthermore shown to constitute, or at least to contain, the transactivating factor ISGF3 (16,17). The gel-shift assay (or electrophoretic mobility shift assay, EMSA), like many other bioassays, conclusively demonstrates the presence of a specific activity in a sample. It does not, however, yield any molecular information on the activity. To further investigate the activity, ISGF3 factor was purified by column chromatography, incubated with the [^{32}P]phosphate-labeled oligonucleotides containing the specific recognition

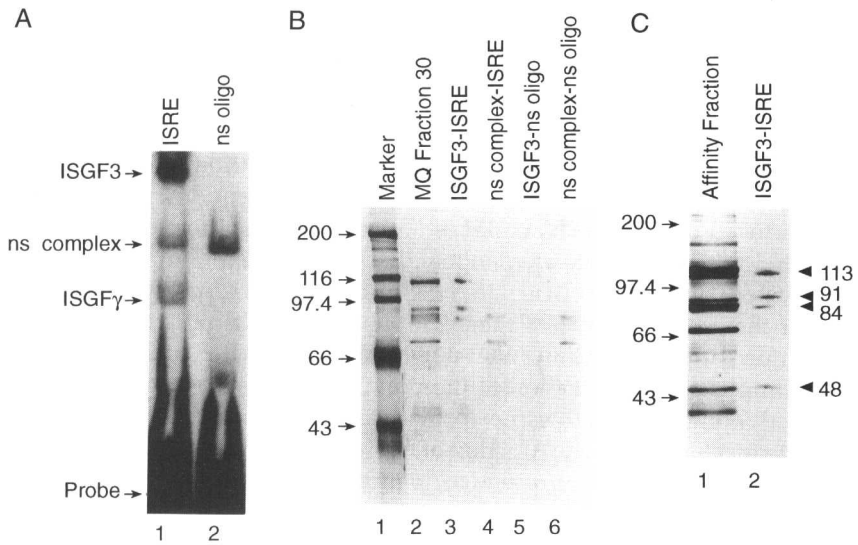


Fig. 2 ISGF3 is composed of four distinct polypeptides. (A) Partially purified ISGF3 was incubated with a radiolabeled ISRE (lane 1) or unrelated oligonucleotide [nonspecific (ns) oligo; lane 2], and protein : DNA complexes were fractionated by EMSA. ISGF3 and ISGF3γ complexes were formed with the ISRE, and a contaminant nonspecific (ns) complex was formed with both probes as indicated at left. After autoradiography, regions of the preparative gel corresponding to all three complexes were excised from both lanes. (B) Proteins present in the excised slices were electroeluted onto SDS–PAGE (8% gel) for electrophoretic separation and stained with silver. (C) An experiment similar to that described in A and B, but using a less pure affinity fraction of ISGF3 (lane 1). Four proteins (48, 84, 91, and 113 kDa; lane 2) were specifically recovered from the ISGF3–ISRE complex. Reproduced with permission from Fu, X. Y., D. S. Kessler, S. A. Veals, D. E. Levy, and J. E. Darnell, Jr. (1990). ISGF3, the transcriptional activator induced by interferon alpha, consists of multiple interacting polypeptide chains. *Proc. Natl. Acad. Sci. USA* **87**:8555–8559. Copyright 1990 American Association for the Advancement of Science.

sequence, and separated from the contaminating proteins by nondenaturing gel electrophoresis. The ISGF3 protein was excised from gels and subjected to denaturing sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE). These and complementing experiments established that ISGF3 consisted of at least four polypeptides of apparent molecular mass 48 kDa, 84 kDa (Stat 1β), 91 kDa (Stat 1α), and 113 kDa (Stat 2), respectively, of which the 48-kDa polypeptide and Stat 1 appeared to directly interact with DNA (21,32) (Fig. 2 B,C).

In the work described previously, the proteins were identified as components of the ISGF3 complex using a biochemical assay. The proteins were isolated and assayed based on their ability to bind to a specific nucleotide sequence. The example illustrates the success of the approach and also highlights some of the limitations of the biochemical "purify-and-assay" technique for the analysis of multicomponent systems. The limitations include the following: (i) Components participating in one activity, in the present case specific transcriptional activation of genes carrying the ISGF3 binding element, were successfully detected. Elucidation of the whole signaling pathway from the membrane to the nucleus required the combined use of genetics, biochemistry, and cell biology approaches (29). (ii) Analysis of the ISGF3 factor did not