Proteomics and Protein-Protein Interactions

Biology, Chemistry, Bioinformatics, and Drug Design

PROTEIN REVIEWS

Recent Volumes in this Series

VIRAL MEMBRANE PROTEINS: STRUCTURE, FUNCTION, AND DRUG DESIGN
Edited by Wolfgang B. Fischer

THE p53 TUMOR SUPPRESSOR PATHWAY AND CANCER
Edited by Gerard P. Zambetti

PROTEOMICS AND PROTEIN-PROTEIN INTERACTIONS: BIOLOGY, CHEMISTRY, BIOINFORMATICS, AND DRUG DESIGN
Edited by Gabriel Waksman

Proteomics and Protein-Protein Interactions

Biology, Chemistry, Bioinformatics, and Drug Design

Protein Reviews Volume 3

Edited by

Gabriel Waksman
Institute of Structural Molecular Biology
Birkbeck and University College London
United Kingdom

Springer

Library of Congress Cataloging-in-Publication Data

Proteomics and protein-protein interactions : biology, chemistry, bioinformatics, and drug design / Gabriel Waksman, editor.
p. cm. — (Protein reviews)
Includes bibliographical references and index.
ISBN 0-387-24531-6
1. Protein-protein interactions. 2. Proteomics. I. Waksman, Gabriel. II. Series.

QP551.5.P765 2005
612′.01575—dc22 2005040224

ISBN-10: 0-387-24531-6 e-ISBN 0-387-24532-4 Printed on acid-free paper.
ISBN-13: 978-0387-24531-7

Printed in Singapore. (TB/KYO)

9 8 7 6 5 4 3 2 1

springeronline.com

Preface

Gabriel Waksman

Institute of Structural Molecular Biology, Birkbeck and University College London, Malet Street, London WC1E 7HX, United Kingdom

Address for correspondence:

Professor Gabriel Waksman
Institute of Structural Molecular Biology
Birkbeck and University College London
Malet Street
London WC1E 7H
United Kingdom
Email: g.waksman@bbk.ac.uk and g.waksman@ucl.ac.uk
Phone: (+44) (0) 207 631 6833
Fax: (+44) (0) 207 631 6833
URL: http://people.cryst.bbk.ac.uk/∼ubcg54a

Gabriel Waksman is Professor of Structural Molecular Biology at the Institute of Structural Molecular Biology at UCL/Birkbeck, of which he is also the director. Before joining the faculty of UCL and Birkbeck, he was the Roy and Diana Vagelos Professor of Biochemistry and Molecular Biophysics at the Washington University School of Medicine in St Louis (USA).

The rapidly evolving field of protein science has now come to realize the ubiquity and importance of protein–protein interactions. It had been known for some time that proteins may interact with each other to form functional complexes, but it was thought to be the property of only a handful of key proteins. However, with the advent of high-throughput proteomics to monitor protein–protein interactions at an organism level, we can now safely state that protein–protein interactions are the norm and not the exception. Thus, protein function must be understood in the larger context of the various binding complexes that each protein may form with interacting partners at a

given time in the life cycle of a cell. Proteins are now seen as forming sophisticated interaction networks subject to remarkable regulation. The study of these interaction networks and regulatory mechanism, which I would like to term "systems proteomics," is one of the thriving fields of proteomics.

The birds-eye view that systems proteomics offers should not, however, mask the fact that proteins are each characterized by a unique set of physical and chemical properties. In other words, no protein looks and behaves like another. This complicates enormously the design of high-throughput proteomics methods. Unlike genes, which, by and large, display similar physicochemical behaviors and thus can be easily used in a high-throughput mode, proteins are not easily amenable to the same treatment. It is thus important to remind researchers active in the proteomics field of the fundamental basis of protein chemistry. This book attempts to bridge the two extreme ends of protein science: on one end, systems proteomics, which describes, at a system level, the intricate connection network that proteins form in a cell, and on the other end, protein chemistry and biophysics, which describe the molecular properties of individual proteins and the structural and thermodynamic basis of their interactions within the network.

Bridging the two ends of the spectrum is bioinformatics and computational chemistry. Large datasets created by systems proteomics need to be mined for meaningful information, methods need to be designed and implemented to improve experimental designs, extract signal over noise, and reject artifacts, and predictive methods need to be worked out and put to the test. Computational chemistry faces similar challenges. The prediction of binding thermodynamics of protein–protein interaction is still in its infancy. Proteins are large objects, and simplifying assumptions and shortcuts still need to be applied to make simulations manageable, and this despite exponential progress in computer technology.

Finally, the study of proteins impacts directly on human health. It is an obvious statement to say that, for decades, enzymes, receptors, and key regulator proteins have been targeted for drug discovery. However, a recent and exciting development is the exploitation of our knowledge of protein–protein interaction for the design of new pharmaceuticals. This presents particular challenges because protein–protein interfaces are generally shallow and interactions are weak. However, progress is clearly being made and the book seeks to provide examples of successes in this area.

I would like to thank all contributors for their participation to this book, which, I believe, is timely and provides a good overview of the field. It is their hard work that has made this book what it is, a fascinating foray into the complex world of protein science. Proteomics and Protein-Protein Interactions: Biology. Chemistry. Bioinformatics. And Drug Design.

Gabriel Waksman

Contents

1

Introduction: Proteomics and Protein–Protein Interactions: Biology, Chemistry, Bioinformatics, and Drug Design

Gabriel Waksman and Clare Sansom

ABSTRACT

In this chapter, a general introduction to the book is provided. We explain the organization of the book and how the chapters are interconnected to each other. We also provide some illustrations and highlights that will complement the various chapters.

1. INTRODUCTION

The formidable advances in protein sciences in recent years have highlighted the importance of protein–protein interactions in biology. Before the proteomics revolution, we knew that proteins were capable of interacting with each other and that protein function was regulated by interacting partners. However, the extent and degree of the protein–protein interaction network was not realized. It is now believed that not

GABRIEL WAKSMAN AND CLARE SANSOM • Institute of Structural Molecular Biology, Birkbeck and University College London, Malet Street, London WC1E 7HX, UK.

Proteomics and Protein–Protein Interactions: Biology, Chemistry, Bioinformatics, and Drug Design, edited by Waksman. Springer, New York, 2005.

only are a majority of proteins in a eukaryotic cell involved in complex formation at some point in the life of the cell, but also that each protein may have on average six to eight interacting partners (Tong et al., 2004). We start this book with two chapters that describe the successes and limitations of two of the most productive methods used to study protein–protein complexes on a large "proteome" scale: the yeast two-hybrid method (Auerbach and Stagljar, Chapter 2) and mass spectrometry (Wang, Yazdi, and Qin, Chapter 3). A brief description of the principle of the methods is provided in Section 2 of this chapter.

The study of protein–protein interactions predates the proteomics revolution. The pioneering work on antigen–antibody and protease–protease inhibitor complexes has provided insight into protein–protein interfaces and their properties (Ruhlmann et al., 1972; Amit et al., 1986). However, more recently, the structure of larger complexes that function as molecular machines has been determined, shedding light into important cellular functions such as transcription (the structure of the RNA polymerase II core complex [Cramer et al., 2001]), translation (the structure of the ribosome [Ban et al., 2000; Wimberly et al., 2000]), replication (the structure of the γ-complex in bacteria [Jeruzalmi et al., 2001]), or the cytoskeleton (the structure of the Arp2/3 complex [Robinson et al., 2001]), to cite only a few. Section 3 of this chapter provides highlights for some of these structures and describes some of the most striking advances that these structures have contributed. It is not the purpose of this book to make an exhaustive list of all protein–protein complexes, the structure of which has been determined to date, but instead to provide general concepts on the common and distinctive features of protein-protein interfaces and their roles.

Protein–protein interactions can be classified into approximately three subtypes, depending on their stability and the mode of interactions (see Walker-Taylor and Jones, Chapter 5). Although any one protein may be involved in interactions with many others, they may form stable interactions only with a few. These form core complexes, which are stable, can be purified, and are amenable to structural studies. A number of core complex structures have been determined and these structures have been instrumental in understanding how these core complexes carry out their functions (see list above). A second category of interactions are transient, and the proteins involved in transient associations are often regulatory proteins, the role of which may be to confer short-lived, physiologically regulated properties to other proteins or to core complexes. The complexes that proteins form transiently may be unstable and difficult to purify. Successes in determining the structure of transient complexes have been dependent on the affinity of the various constituents participating in complex formation. One historical breakthrough in this regard has been the determination of the structure of the first antigen–antibody complex (Amit et al., 1986), which defined the architecture and chemistry of protein–protein interactions in this versatile structural framework. Sundberg and Mariuzza (Chapter 4) in this book provide an exhaustive review of the antigen–antibody and MHC–TCR complexes. Finally, protein–protein interactions in core complexes or transient ones may be mediated by specialized, small, domains dedicated to protein–protein recognition (Pawson and Scott, 1997).

The proteomics revolution (and before it, the genomics revolution) has in many ways overwhelmed our ability to keep up with the sheer volume of data that it has generated. In this regard, the rapid development of computer science in general and bioinformatics in particular is crucial for us to be able to make sense of the data. In two chapters, by Walker-Taylor and Jones (Chapter 5) and by Marshall and Vasker (Chapter 6), the authors describe the bioinformatics tools that have been deployed to understand the general principles of protein–protein interactions (Walker-Taylor and Jones, Chapter 5) and how this knowledge has been exploited to design predictive docking algorithms to model protein–protein interactions from structure (Marshall and Vasker, Chapter 6).

Unfortunately, prediction of protein–protein interaction is a rather difficult endeavour, not least because proteins may undergo vast conformational changes on association with other proteins. One particularly relevant example of such a case is the interaction of the human immunodeficiency virus (HIV) envelope glycoprotein gp120 with its human receptor CD4. Structural rearrangement in gp120 serves not only as an allosteric trigger for cellular invasion, but also as a mechanism for evading the host immune system. Doyle and Hensley in Chapter 7 show how large conformational changes in proteins can be probed thermodynamically (see also Section 3 of this chapter for the structure of the gp120/CD4/anti-gp120 antibody ternary complex [Kwong et al., 1998]). Another example of the use of thermodynamic investigation of conformational changes, this time in the unbound state, is provided in Section 4 of this chapter.

One outcome of such studies is the observation that protein–protein interfaces possess a high degree of versatility and plasticity (Jones and Thornton, 1996; Lo Conte et al., 1999). This is in part due to the fact that protein–protein interactions encompass a wide range of affinities. However, even within a particular range of affinities, the structural features underpinning binding may vary. For example, protease inhibitors appear to use main-chain–main-chain interactions, whereas antigen-antibody interaction is mediated by side-chain side-chain interactions (Jackson, 1999). Side-chain–side-chain interactions may be more likely to determine specificity. In contrast, serine protease inhibitors must bind tightly to their target proteases. This may be best achieved using constrained "main-chain–main-chain" conformation, and means that the inhibitor will be highly committed to the enzyme. Similar observations corresponding to similar requirements have been observed in the interaction of pilus subunits with bacterial chaperones (see Section 5 of this chapter and [Choudhury et al., 1999; Sauer et al., 1999]). Proteins may be able to use the same template for interactions with different proteins or with different parts of the same protein. For example, in the growth hormone–growth hormone receptor complex (GH–GHR), two receptor molecules bind to different parts of the same ligand (see Section 6 of this chapter and de Vos et al., 1992). This complex structure and the subsequent site-directed mutagenesis studies have defined an important concept in protein–protein interaction, that of "hot spots" (Section 6 and Clackson and Wells, 1995). In the protein–protein interface observed in the GH–GHR complex, although a myriad of hydrogen bonds, van der Waals contacts, and electrostatic interactions is observed, only a limited few of these

interactions ("hot spots") have been shown to play an important role in binding. Similar observations have been made in other systems (Dall'Acqua et al., 1996; Bradshaw et al., 1999). One important property of protein–protein interfaces is the shape complementarity between the two regions coming together in the interactions. Remarkably, such complementarity is very often mediated by water molecules, judiciously placed at the interface to fill in holes and increase contacts (Bhat et al., 1994; Lubman and Waksman, 2003). The role of water in both the structural and thermodynamic basis of protein–protein interactions is essential and yet very poorly understood.

One remarkable feature of protein–protein interactions is that they are often mediated by small domains that specifically bind small sequence motifs on proteins. The Src-homology 2 (SH2) domain was the first such domain to be recognized (Sadowski et al., 1986). SH2 domains are involved in the building up of large complexes at and around signaling receptors. SH2 domains bind specifically sequences containing phosphorylated tyrosines and are able to discriminate between tyrosine-phosphorylated sites by exercising some preferences for residues located C-terminally relative to the phosphotyrosine (Bradshaw and Waksman, 2002). Since the discovery of SH2 domains, a large number of protein domains with specialized roles in protein–protein interactions have been found (see http://www.mshri.on.ca/pawson/domains.html for an exhaustive list of such domains). SH3 and WW domains specifically recognize and bind sequence motifs containing prolines (Musacchio, 2002). Proline-rich motifs are among the most common motifs identified, and thus SH3 and WW domains play major roles in protein–protein interactions. PDZ domains are essential for integrity of the postsynaptic density, a large protein complex formed around glutamate and *N*-methyl-D-aspartate (NMDA) receptors in the nervous system (Sattler and Tymianski, 2001). Finally, bromodomains are similar to SH2 domains in that both bind (and thus induce recruitment of proteins that contain them) to sites of protein modifications. Bromodomains, unlike SH2 domains that bind to tyrosine-phosphorylated sites, bind specifically acetylated lysines and thus play important roles in chromatin remodeling during transcription and replication (Dhalluin et al., 1999). Three chapters of this book are dedicated to the review of the protein–protein interaction domains listed above. Ladbury in Chapter 8 provides an account of the structural and thermodynamics work that has enhanced our understanding of SH2 and phosphotyrosine binding (PTB) domain recognition of tyrosine-phosphorylated sites during transduction of cellular signals. Bedford and Sudol (Chapter 9) describe the roles and functions of SH3 and WW domains. Finally, Yan and Zhou (Chapter 10) provide a detailed account of the discovery of bromodomains and also of their struture–function relationship.

There is no unifying theme among the structures of the protein–protein interaction domains listed above. However, as their structures have been characterized, intense efforts have been devoted to designing specific binding inhibitors capable of disrupting protein–protein interaction mediated by these modules. For example, the SH2 domain of the Src kinase has been targeted for molecular design and binding competitors able to inhibit osteoclast function have been found (Sawyer et al., 2002). As the major phenotype in Src knockout mice is a thickening of the bones, it is hoped

that a Src SH2 domain binding inhibitor could be used to combat osteoporosis, a devastating disorder in elderly women. Sawyer et al. in Chapter 11 review the field of SH2 domain binding inhibitors and also provide a fascinating account on the remarkable progress made in designing new tyrosine kinase inhibitors.

Using peptides or peptide mimics to disrupt protein–protein interfaces is not a novel idea, but this approach has benefited from structural information. Notably, the molecular design of rigid peptidomimetics is believed to enhance greatly the potential of peptides as therapeutics by not only locking the peptide in a defined binding conformation but also by preventing or slowing degradation (Patani and LaVoie, 1996). However, peptides or peptide-based compounds do not readily cross the cell membrane barriers and thus are not as effective as hoped. Recently, a peptide derived from the NR2 chain of the NMDA receptor known to interact with the second PDZ domain of PSD95, an essential component of the postsynaptic density, was made effective in reducing cerebral infarction in rats subjected to transient focal cerebral ischemia by fusing it to the HIV1-Tat translocator peptide (Aarts et al., 2002). Thus, the use of translocator peptides may be a promising avenue of research for the delivery of therapeutic peptides or proteins (Becker-Hapak et al., 2001). In this book, Aarts and Tymianski (Chapter 12) provide a detailed account of this work and place it in the general context of NMDA receptor signaling and its more general role in the response to the devastating damages to the human nervous system caused by stroke, epilepsy, and head and spinal injury.

As our knowledge of protein–protein interactions increases, such interactions will be more frequently targeted for drug design. Chapter 13 by Freire provides a guide and a general strategy to improve the hit-to-lead route that is so often paved with multiple insurmountable obstacles. However, we should not ignore the vast amount of work that is being achieved in the design of novel forward or reverse genetics methods or target-guided self assembly methods (reviewed in Alaimo et al., 2001) and how these can be exploited to probe the living cell in general and protein–protein interaction networks in particular.

With this book, we have attempted to provide a multifaceted account of the research taking place in protein science. We have attempted to cover the field in a most exhaustive way, choosing highlights from leaders in the field in such a way that our choice should reflect the enormous diversity and complexity of the principles underpinning protein–protein interactions. We hope we have achieved this goal.

2. PRINCIPLES OF MASS SPECTROMETRY (MS) AND OF THE YEAST TWO-HYBRID METHOD

MS and the yeast two-hybrid method are two key proteomics technologies that have been developed in recent years and that can be applied to the study of protein interactions and protein–protein complexes.

The development of MS as a precise technique for the identification of proteins and peptides from their molecular mass and charge has led to an explosion in

high-throughput protein analysis methodologies. The basic principle of MS (shown in Fig. 1.1, Ia) involves the production of a stream of ions—in proteomics experiments these will typically be ionized peptide fragments of a protein or complex—from the sample. The ion stream then passes through an analyzer where the ions are sorted by mass/charge ratio, and then enters the detector where the mass/charge spectrum is recorded. The most widely used type of analyzer is the time-of-flight (ToF) analyzer. In this, the ions are accelerated so ions of like charge have the same kinetic energy: it can be seen from basic physical principles that there is an inverse relationship between the time taken for an ion to travel between the source and the detector (measured in microseconds) and its mass/charge ratio.

Currently, the two most commonly used ionization techniques are MALDI (Matrix Assisted Laser Desorption Ionization), shown schematically in Figure 1.1, Ib, and ESI (ElectroSpray Ionization), shown in Figure 1.1, Ic. Of these, the gentleness of the ESI method makes it particularly suitable for the study of protein–protein interactions, as noncovalent interactions may sometimes be retained during this ionization process. In MALDI the sample of peptides is embedded in a matrix. When an ultraviolet laser beam ($\sim$337 nm) is shone onto this matrix the laser energy is transferred into the peptides, releasing a pulsed stream of ions. In contrast, in ESI the peptide sample, dissolved in a suitable solvent, is sprayed out of a needle with a thin tip at a high voltage into an inert, drying gas. This produces charged droplets containing the sample ions, and the droplets then evaporate leaving the ions in the gas phase. These are swept through a sampling cone toward the (usually ToF) analyzer and detector systems.

The yeast two-hybrid method, shown in Figure 1.1, II, is now a well-established methodology for detecting proteins or domains that are capable of interacting. One important advantage of this method is that it results in the cloned genes for the interacting proteins becoming immediately available. Plasmids are constructed containing, respectively, the DNA-binding domain of the yeast transcriptional activator GAL4 fused to the known protein for which interacting partners are sought (top diagram) and the GAL4 activation domain bound to a library of proteins or fragments (middle diagram; n fragments). Interaction between a protein from the library (say protein i) and the known protein (bottom diagram) causes the GAL4 activation domain to bind to DNA, leading to transcriptional activation of a reporter gene that contains a binding site for GAL4. β-Galactosidase is typically used as the reporter gene.

3. EXAMPLES OF LARGE PROTEIN COMPLEX STRUCTURES

Recent determination of the structures of high molecular weight protein complexes has shed light on the mechanism of several important cellular processes, including transcription, replication, cell motion, and viral adhesion. The mechanism of these complexes is often driven by conformational changes both within and between subunits.

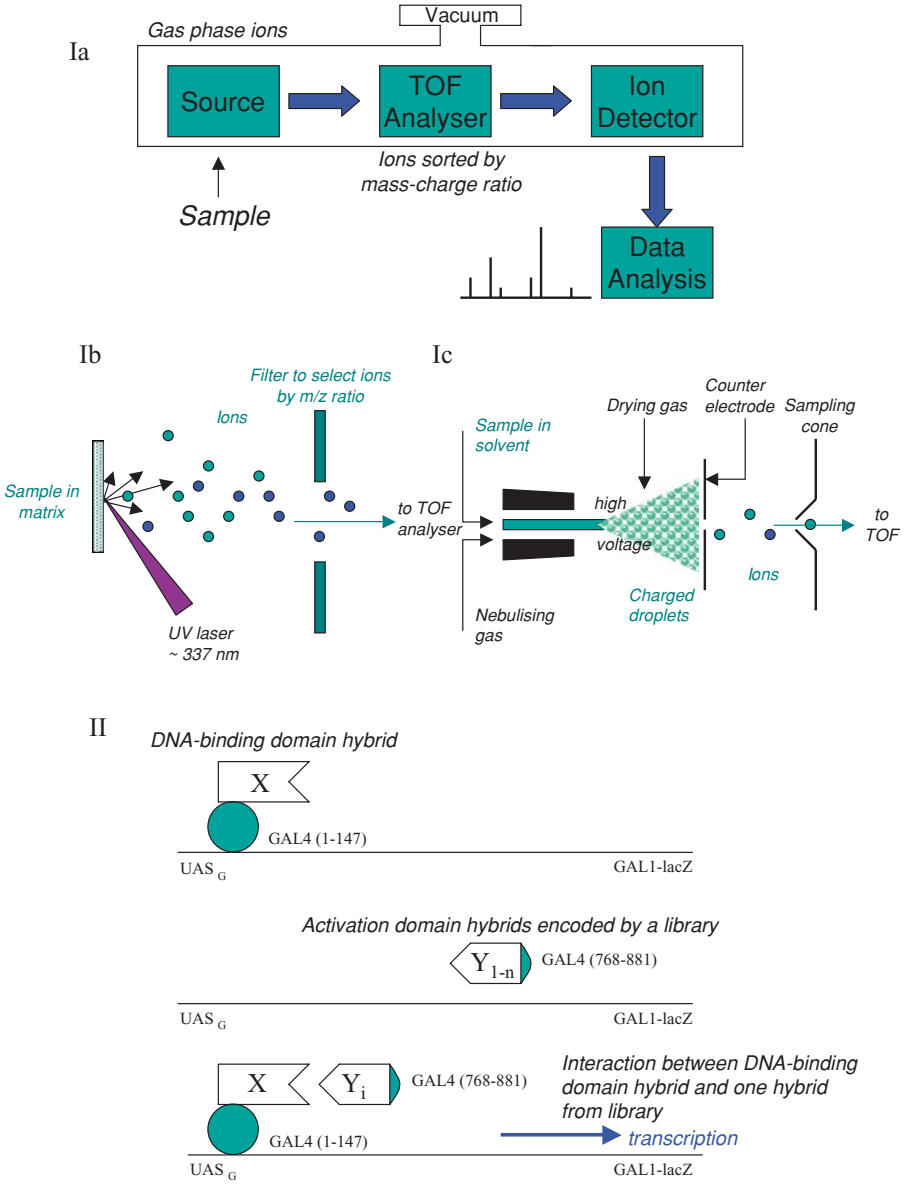


Figure 1.1. Description of the techniques of mass spectrometry and of the yeast two-hybrid method to study macromolecular complexes.

RNA polymerase II is responsible for all mRNA synthesis in eukaryotes. The structure of 10 of the 12 subunits of this enzyme from yeast, at 2.8A resolution, is shown in Figure 1.2, I (Cramer et al., 2001). The authors propose that the DNA double helix could enter the cleft of this "open" structure and be held in place for transcription by a massive protein "clamp" consisting of parts of subunits 1 (shown in *red* in this figure) and 2 (shown in *yellow*). The DNA cleft runs from the bottom right to the top left of the molecule as shown here. A second structure, solved at lower resolution, shows the clamp swung round toward the active centre of the molecule in a "closed" conformation (Gnatt et al., 2001). The models suggest that the most likely route for RNA exit would be via a groove at the base of the clamp.

In bacteria, an ATPase known as the γ complex plays an important part in DNA replication. It is the part of the larger DNA polymerase III complex that loads the pol III β-subunit (the sliding β-clamp) onto the DNA; it is therefore a homolog of eukaryotic replication factor C. Once attached to the loaded clamp, the catalytic α-subunit will move along the DNA and catalyze replication. The *Escherichia coli* γ complex, shown in Fig. 1.2 II, is a pentamer of five subunits, each with the same fold (an N-terminal recA-like domain followed by two helical domains) but with different interdomain orientations in each subunit (Jeruzalmi et al., 2001). The nucleotide binding sites of the subunits are arranged to face the inner surface of the complex. The three "middle" subunits, which are the most similar, are termed γ and the two outer ones δ (the wrench) and δ' (the stator); it is the wrench that binds to the β-clamp. The diagram shows an "open" form of the enzyme in which the wrench is free to bind to the clamp; this structure suggests a mechanism for replication in which the complex switches between this and a closed form where the wrench is occluded.

The Arp2/3 complex is an assembly of seven proteins that initiates actin polymerisation in eukaryotic cells. This process generates the network of branched actin filaments that pushes forward the leading edge of motile cells. This assembly, shown in Figure 1.2, III, consists of two central subunits with the same fold as actin, Arp2 (cyan) and Arp3 (*blue*), and five peripheral subunits (p40 to p16) with different folds. Part of the Arp2 subunit was not observed in the crystal structure. The complex is inactive until it is activated by nucleation-promoting factors such as members of the WASp/Scar protein family (WASp, the first member of this family to be discovered, is the Wiskott–Aldrich syndrome protein). The structure as shown suggests that activation could take place by motion of the ARPC subunits inducing a conformation change bringing the Arp2 and Arp3 subunits in contact with each other to form a nucleation site.

One key target for anti-HIV drug design is the binding of the HIV virus envelope glycoprotein (gp120) to the CD4 receptor of the host T cells, prior to viral entry into the cells. The structure of the gp120–CD4 complex has been solved, in complex with a neutralizing human antibody (Fig.1.2, IV). The presence of a hydrophobic cavity at the CD4/gp120 interface in this structure is a clear indication that the formation of the complex must induce a conformational change in gp120. The structure also shows a conserved chemokine receptor-binding site, and illustrates possible points for intervention by potential anti-HIV entry drugs. This section of Figure 1.2 also serves as a figure for Chapter 7 by Doyle and Hensley.

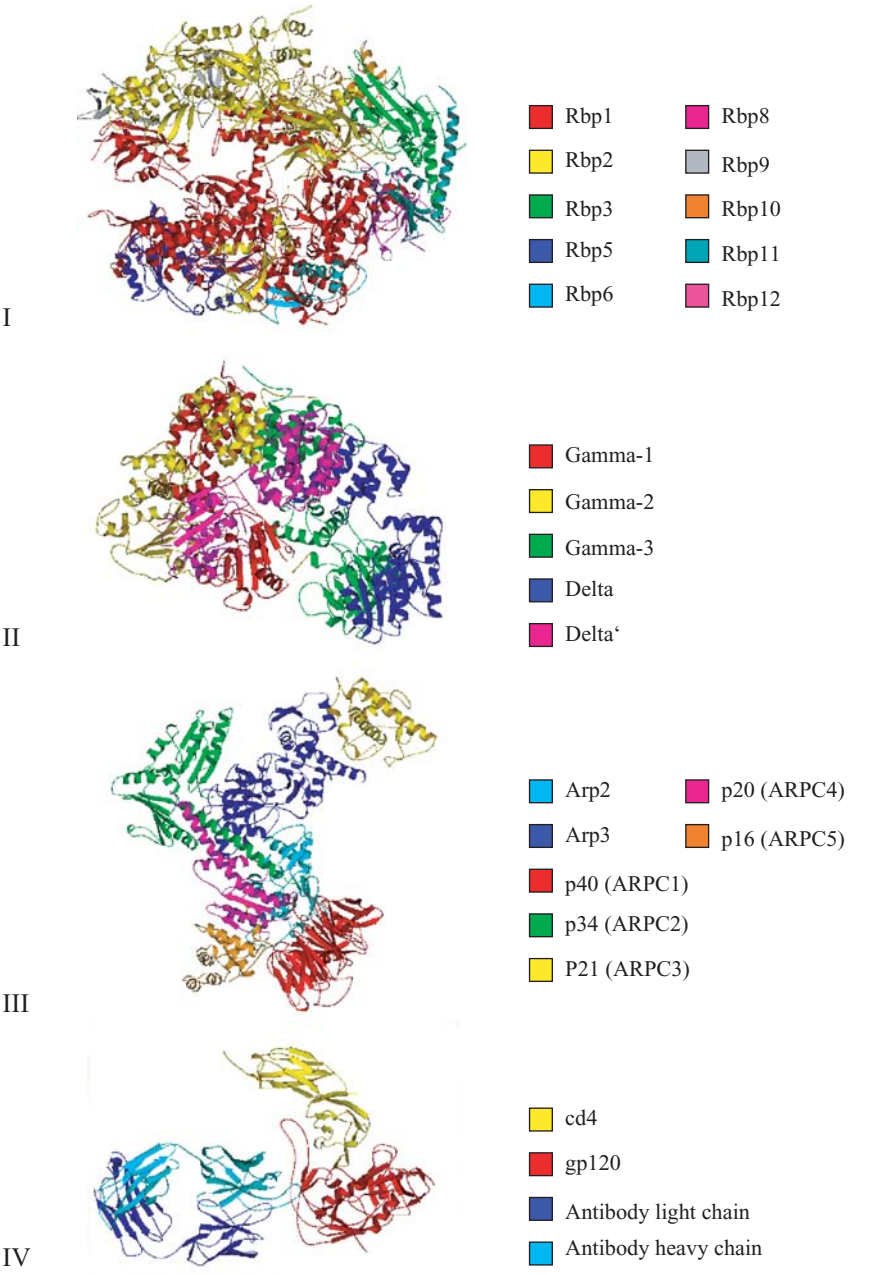


Figure 1.2. Macromolecules complexes formed during transcription, replication, actin polymerization, and HIV host recognition.

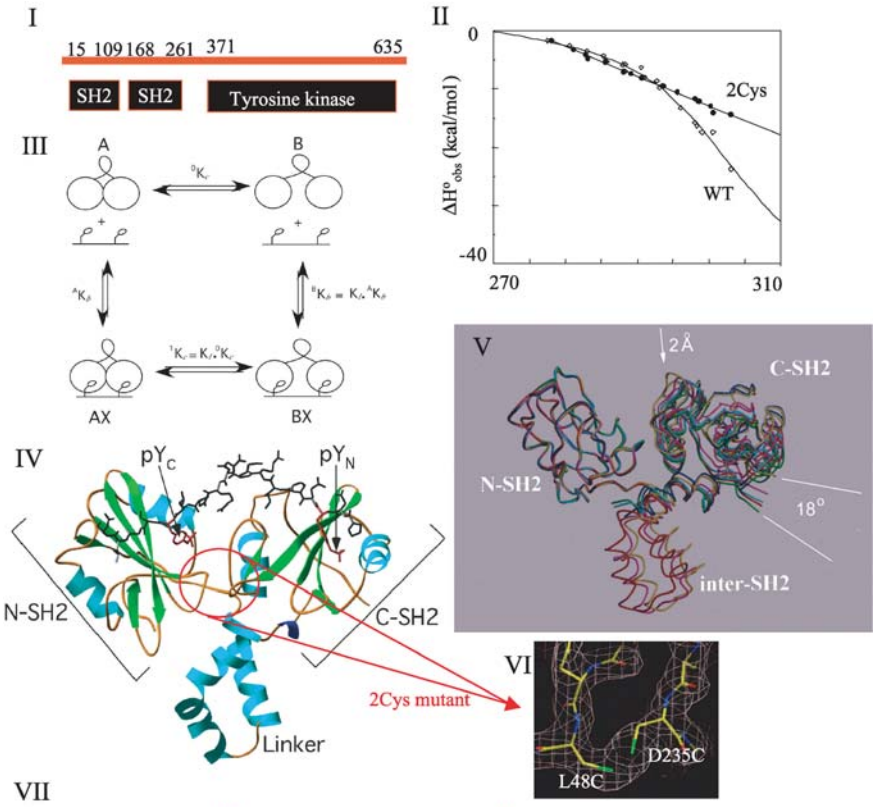
4. PROBING CONFORMATIONAL CHANGES IN THE UNBOUND STATE: BINDING THERMODYNAMICS OF THE TANDEM SH2 DOMAIN OF THE SYK KINASE

As illustrated in Chapter 7 by Doyle and Hensley, "induced fit" conformational changes occurring on binding can be probed using calorimetry. Indeed, such conformational changes have distinct thermodynamic signatures, that is, they are generally characterized by a large heat capacity change. However, the heat capacity change can also be used to characterize other types of conformational transitions, and notably as demonstrated recently by Kumaran, Grucza, and Waksman, to characterize conformational transitions occurring in the unbound state (Kumaran et al., 2003).

The macromolecular system studied by Kumaran et al. (2003) is the tandem SH2 domains of the Syk kinase. The Syk kinase is involved in signal transduction pathways mediated by immune receptor. It contains two SH2 domain located in tandem (termed Syk-tSH2; I in Fig. 1.3). Syk-tSH2 allows recruitment of Syk to tyrosine-phosphorylated sites termed ITAMs (*I*mmuno-receptor *T*yrosine-based *A*ctivation *M*otifs) on the immune receptor. The sequence of three such receptor ITAMs is shown in VII (Y* indicates a phosphotyrosine). Syk-tSH2 is remarkable because it can bind with high affinity to doubly phosphorylated ITAMs (dpITAMs) that have widely different sequences and also very different lengths of sequence (indicated as "spacer region" in VII) between the two phosphotyrosines (compare dpITAM binding of FcR-γ and FcRIIA in VIII for wild-type Syk-tSH2). How can Syk-tSH2 do this?

The answer to this question was provided by a calorimetric experiment in which the binding enthalpy was measured as a function of temperature (results in Fig. 1.3, II, open diamonds.) Note that these results were obtained using the CD3-ε peptide shown in Fig.1.3, VII). As can be seen, the binding ΔH has a nonlinear dependence on temperature. This is odd, as most binding reactions involving proteins display a linear dependence. These results were interpreted in light of a two-conformer model, which is shown in Figure 1.3, III. In this model, the unbound form of Syk-tSH2 exists in a temperature-dependent conformational equilibrium involving two forms, A and B. The A form predominates at low temperature, and is the preferential binding form for the CD3-ε dpITAM peptide. At low temperature, the observed binding enthalpy is that of binding alone. As the temperature is raised, the B state becomes populated, and the binding enthalpy is now composed of two terms, the enthalpy arising from the B to A transition and the intrinsic binding enthalpy (A to AX). This model fits the data very well (see line through experimental data points). But is it right?

The model was proven to be correct by a second series of experiments. It was hypothesized that the A and B forms of Syk-tSH2 corresponded to a closed and an open form of the protein, respectively. The closed form corresponds to the two SH2 domains being close together. The open form corresponds to the two SH2 domains



VIII

dpITAM	Wild type tSH2 domain	2Cys tSH2 domain	Wild type tSH2 domain + 15 mM DTT	2Cys tSH2 domain + 15 mM DTT
CD3-ε	35.0 ± 5.0 nM	33.6 ± 5.1 nM	40.5 ± 6.0 nM	35.0 ± 8.0
FcR-γ	1.7 ± 0.8 nM[a]	29.2 ± 6.6 nM	30.0 ± 9.0 nM	32.3 ± 5.0
FcRIIA	1.6 ± 0.8 nM[a]	100.0 ± 9.1 nM	26.0 ± 6.0 nM	15.0 ± 4.0

Figure 1.3. Thermodynamics of conformational changes: the case of the tandem SH2 domain of the Syk kinase.

being far apart. This hypothesis was suggested by experimental X-ray crystallographic evidence that shows that in the bound state (i.e., bound to the CD3-ε dpITAM peptide of Fig.1.3, VII), the relative orientation of the two SH2 domains is subject to significant variations (see structure of the Syk-tSH2 domain bound to the CD3-ε dpITAM peptide in Fig.1.3, IV; this structure contained six independent views of the complex; when superimposed (see Fig.1.3, V), there is evidence of variability in the relative orientation of the SH2 domains with the two extreme conformations differing by an angle of 18° and a translation of about 2 Å). Thus, it was hypothesized that in the unbound state, the relative orientation of the two SH2 domains could vary even more, perhaps explaining how Syk-tSH2 can bind with equal affinity to dpITAMs with widely different spacer length between the two phosphotyrosines. To test these hypotheses, a mutant (termed 2Cys) containing two judiciously-located Cys residues was engineered (see *red circle* in Fig.1.3, IV). These Cys residues, when oxidized, form a disulfide bond (see Fig.1.3, VI showing the electron density for the bond), which locks the two SH2 domains in the "closed" conformation, that is, when the two SH2 domains are close to each other. If the above hypotheses are correct, such a mutant, being unable to transition to the open form, should no longer display a nonlinear dependence of the binding enthalpy on temperature (it should be linear), and moreover, this mutant should see its affinity for "long" dpITAMs (such as FcRIIA in VII) considerably reduced. This is indeed the case. In II, closed circles, the 2Cys mutant is shown to display a linear dependence of its binding enthalpy on temperature, and as shown in VIII, its binding affinity is considerably reduced for the FcRIIA dpITAM peptide but not for the CD3-ε dpITAM peptide.

5. PROTEIN–PROTEIN INTERACTIONS IN THE CHAPERONE-USHER PATHWAY OF PILUS BIOGENESIS

Bacterial pili (shown in Fig.1.4, I) are important organelles used by pathogenic bacteria to sense their environment and attach to host tissues in a specific manner. Thus, pili play important roles in initiating infections. Of all pili, the P pilus has been the most thoroughly investigated. P pili are found on the surface of uropathogenic *Esherichia coli* (UPEC) and are known to initiate attachment of UPECs to the kidney epithelium. This specific recognition event is due to the presence of an adhesion protein called PapG at the tip of the pilus (see Fig.1.4, I), which binds to a kidney receptor called globoside (GbO4). The P pilus is composed of a thick part at its base that is formed by the subunit PapA and of a thin and flexible part that is formed by the subunit PapE (see Fig.1.4, I). Two subunits act as adaptor subunits: PapK inserts between the thick and thin part of the pilus (i.e., between PapA and PapE), and PapF inserts between the terminal PapG subunit (or adhesin) and PapE. All subunits are assembled into a pilus by a conserved mechanism involving a chaperone, PapD, and an outer-membrane protein, the PapC usher (see Fig.1.4, I). After translation, each subunit is translocated to the periplasm by the Sec pathway where each

is taken up by the chaperone PapD, forming a binary chaperone–subunit complex. Each complex is then targeted to the usher PapC where it is assembled within the pilus. The first subunit to be processed is PapG, hence its location at the tip of the pilus.

The molecular basis of pilus assembly has been elucidated (Choudhury et al., 1999; Sauer et al., 1999; Barnhart et al., 2000; Dodson et al., 2001). Each subunit has an Ig fold where the last strand (strand G) is missing. The result of this missing secondary structure is a large groove on the subunit surface (see Fig.1.4, II, where the

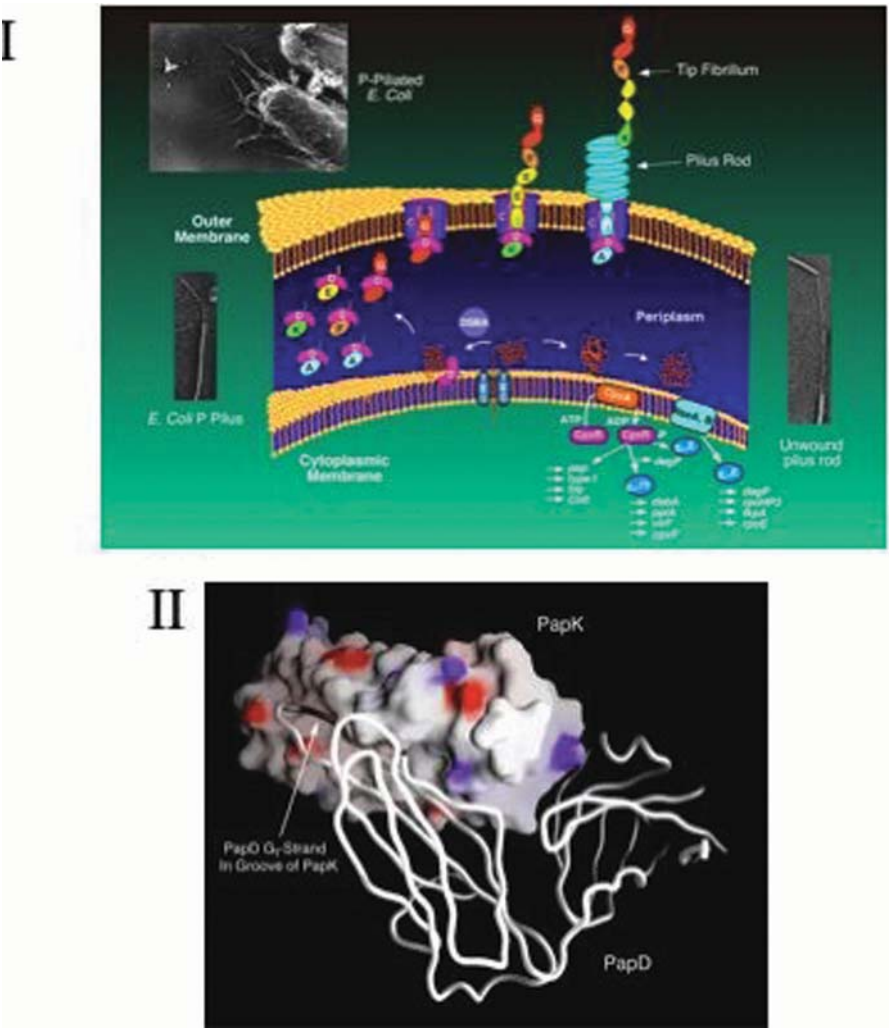


Figure 1.4. Protein-protein interactions in the chaperone-usher pathway of pilus biogenesis.

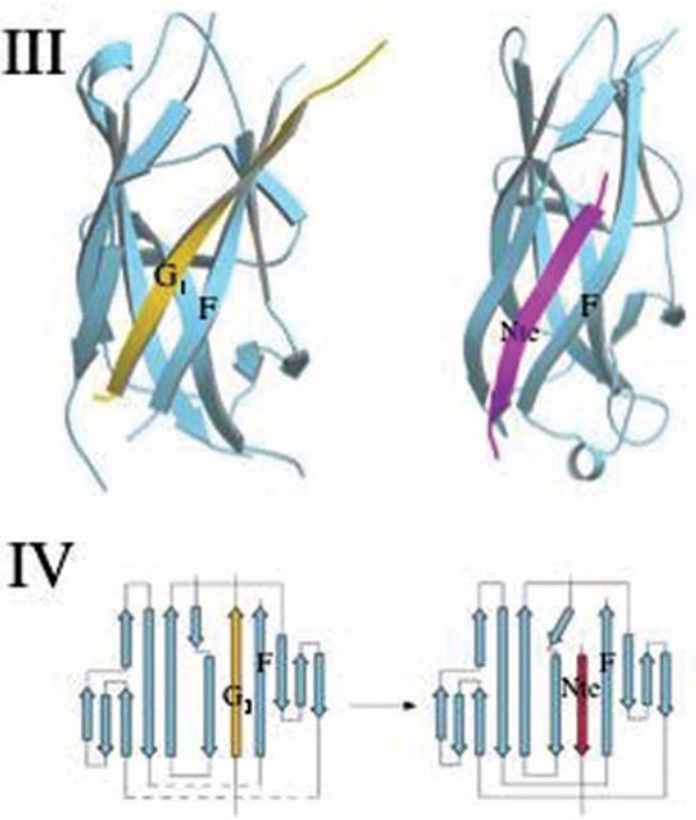


Figure 1.4. (Continued)

subunit is represented in a surface diagram while the chaperone is shown in a ribbon diagram). The missing secondary structure is provided in *trans* by the chaperone, which "donates" its G1 strand (indicated in Fig.1.4, II). The mechanism by which the chaperone provides the missing secondary structure is called "donor-strand complementation." During pilus assembly, the chaperone–subunit complex is targeted to the usher where the complex dissociates. The chaperone is released in the periplasm while the subunit binds to the subunit that was assembled in the previous round of assembly. This process occurs through a mechanism termed "donor-strand exchange" which is explained in Fig.1.4, III and IV. In Fig.1.4, III at left, the complex of the PapE subunit (in cyan) with the G1 strand of the chaperone (in yellow) is shown. In Fig.1.4, IV at left, the same complex is shown but in a topological representation. Both show that the G1 strand of the chaperone is complementing the fold of the subunit PapE. However, note that the G1 strand is running parallel to strand F and thus the reconstituted Ig fold is atypical. During donor-strand exchange, the G1 strand dissociates and is replaced by the N-terminal sequence of the subunit which comes next in the assembly

line. Indeed, all subunits contain in their N-termini a sequence containing alternating hydrophobic residues that can insert in the groove of the subunit assembled previously. This is shown in Fig.1.4, III, right, in a ribbon representation and in Fig.1.4, IV, right, in topological representation. Note that, this time, the N-terminal sequence peptide (termed "N-terminal extension" and labelled Nte (red) in Fig. 1.4, III and IV at right) runs anti-parallel to strand F and thus a typical Ig fold is reconstituted. It is this transition from an atypical fold to a typical fold that provides the energy for assembly.

6. STRUCTURE OF THE COMPLEX BETWEEN GROWTH HORMONE AND ITS RECEPTOR

The structure of the complex between human growth hormone (hGH) and the extracellular domain of its receptor (hGHR) is a clear illustration of the principle that proteins are able to use the same template for different interactions. The complex has a 1:2 stoichiometry, with one molecule of the hormone bound to two molecules of the hormone, as shown in Figure 1.5, I (de Vos et al., 1992). The hormone (*red*) folds into a four-helix bundle with an unusual topology; the receptor molecules (*blue* and *green*) are all-beta structures, each containing two separate immunoglobulin-like domains. Although the interactions the receptor molecules make with the hormone are completely different—and the surface area buried by interactions with the "right-hand" (as shown here; *green*) receptor molecule is much greater than that buried by interactions with the "left-hand" receptor molecule, the residues that each receptor contributes to hormone binding are equivalent. The complex is also stabilized by interactions between the C-terminal domains of the receptor molecules.

Mutation of a single residue of the hormone, glycine 120, to arginine, is sufficient to turn the hormone into an antagonist that can only bind one receptor molecule. This 1:1 complex binds its single receptor molecule in a conformation that is extremely similar to that of the tight-binding receptor molecule in the active 1:2 complex. Detailed analysis of the crystal structure of this 1:1 complex has, with alanine scanning, revealed the molecular basis for hormone-receptor affinity (Clackson and Wells, 1995).

The hormone-binding site of the receptor is centered on a hydrophobic patch at the junction of the two immunoglobulin-like domains, and which interacts with hydrophobic residues on the receptor surface. This hydrophobic patch, shown as *blue/green* in the 1:2 complex in Fig. 1.5. II, is centered on two tryptophan residues, Trp 104 and Trp 169, and surrounded by more hydrophilic residues. Alanine scanning showed that the two central tryptophans contribute most of the binding energy to the hormone–receptor complex. The receptor surface area associated with these two residues is shown in *blue* in Fig. 1.5. II; surface area associated with other residues in the "hot spot," including the crucially important Arginine 43, is shown in *green*. Kinetic analysis has shown that electrostatic interactions of this positively charged residue with the receptor play a peripheral role in binding: this residue's importance

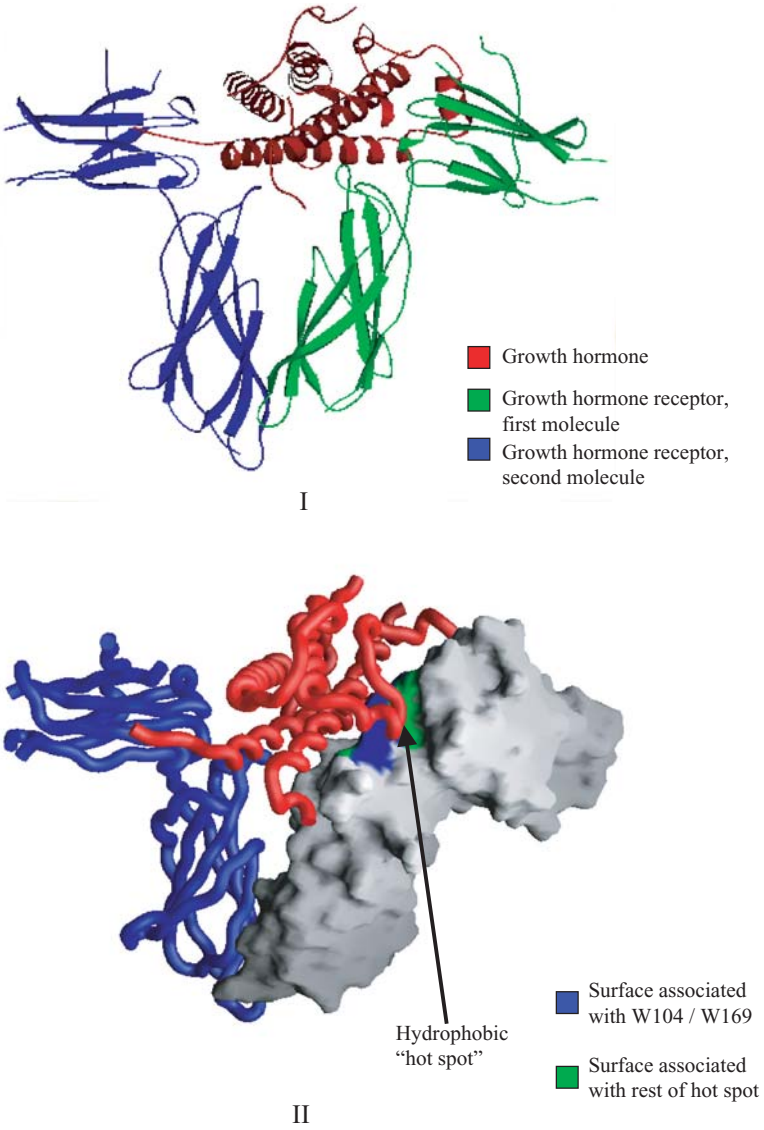


Figure 1.5. Structure of the complex between growth hormone and its receptor.

derives from the fact that its aliphatic side chain packs against Trp 169, wedging it into its required binding conformation. Several other hydrophobic residues in the patch, including proline 106, also serve to "lock" the crucial tryptophan residues into position.

Analyses of many protein–protein interaction surfaces have shown that, although an interaction surface typically involves many residues and buries many

hundreds, if not thousands, of Å^2 of surface area, "hot spots" of a few residues that contribute most of the binding energy to the interaction may be the rule, rather than the exception. This study of the growth hormone–receptor complex highlights the important role of peripheral residues in orienting the residues of the hot spot in an optimum position for binding.

REFERENCES

Aarts, M., Liu, L., Besshoh, S., Arundine, M., Gurd, J.W., Want, Y.T., Salter, M.W., and Tymianski, M. (2002). Treatment of ischemic brain damage by perturbing NMDA receptor-PSD-95 protein interactions. *Science* 298:846–850.

Alaimo, P.J., Shogren-Knaak, M.A., and Shokat, K.M. (2001). Chemical genetic approaches for the elucidation of signaling pathways. *Curr. Opin. Chem. Biol.* 5:360–367.

Amit, A.G., Mariuzza, R.A., Phillips, S.E., and Poljak, R.J. (1986). Three-dimensional structure of an antigen-antibody complex at 2.8 A resolution. *Science* 233:747–753.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* 289:905–920.

Barnhart, M.M., Pinkner, J.S., Soto, G.E., Sauer, F.G., Langermann, S., Waksman, G., Frieden, C., and Hultgren, S.J. (2000). PapD-like chaperones provide the missing information for folding of pilin proteins. *Proc. Natl. Acad. Sci. USA* 97:7709–7714.

Becker-Hapak, M., McAllister, S.S., and Dowdy, S.F. (2001). TAT-mediated protein transduction into mammalian cells. *Methods* 24:247–256.

Bhat, T.N., Bentley, G.A., Boulot, G., Greene, M.I., Tello, D., Dall'Acqua, W., Souchon, H., Schwarz, F.P., Mariuzza, R.A., and Poljak, R.J. (1994). Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc. Natl. Acad. Sci. USA* 91:1089–1093.

Bradshaw, J.M., and Waksman, G. (2002). Molecular recognition by SH2 domains. *Adv. Protein Chem.* 61:161–210.

Bradshaw, J.M., Mitaxov, V., and Waksman, G. (1999). Investigation of phosphotyrosine recognition by the SH2 domain of the Src kinase. *J. Mol. Biol.* 293:971–985.

Choudhury, D., Thompson, A., Stojanoff, V., Langermann, S., Pinkner, J., Hultgren, S.J., and Knight, S.D. (1999). X-ray structure of the FimC-FimH chaperone-adhesin complex from uropathogenic *Escherichia coli*. *Science* 285:1061–1066.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science* 267:383–386.

Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* 292:1863–1876. Epub 2001 Apr 1819.

Dall'Acqua, W., Goldman, E.R., Eisenstein, E., and Mariuzza, R.A. (1996). A mutational analysis of the binding of two different proteins to the same antibody. *Biochemistry* 35:9667–9676.

de Vos, A.M., Ultsch, M., and Kossiakoff, A.A. (1992). Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science* 255:306–312.

Dhalluin, C., Carlson, J.E., Zeng, L., He, C., Aggarwal, A.K., and Zhou, M.-M. (1999). Structure and ligand of a histone acetyltransferase bromodomain. *Nature* 399:491–496.

Dodson, K.W., Pinkner, J.S., Rose, T., Magnusson, G., Hultgren, S.J., and Waksman, G. (2001). Structural basis of the interaction of the pyelonephritic *E.coli* adhesin to its human kidney receptor. *Cell* 105: 73–743.

Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science* 292:1876–1882.

Jackson, R.M. (1999). Comparison of protein–protein interactions in serine-protease and antibody-antigen complexes: implications for the protein docking problem. *Protein Sci.* 8: 603–613.

Jeruzalmi, D., O'Donnell, M., and Kuriyan, J. (2001). Crystal structure of the processivity clamp loader gamma (gamma) complex of *E. coli* DNA polymerase III. *Cell* 106:429–441.

Jones, S., and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc Natl. Acad. Sci. USA* 93:13–20.

Kumaran, S., Grucza, R.A., and Waksman, G. (2003). The tandem Src homology 2 domain of the Sky kinase: a molecular device that adapts to interphosphotyrosine distances. *Proc. Natl. Acad. Sci.* USA 100:14828–14833.

Kwong, P.D., Wyatt, R., Robinson, J., Sweet, R.W., Sodroski, J., and Hendrickson, W. A. (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* 393:648–659.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* 285:2177–2198.

Lubman, O.Y., and Waksman, G. (2003). Structural and thermodynamic basis for the interaction of the Src SH2 domain with the activated form of the PDGF beta-receptor. *J. Mol. Biol.* 328:655–668.

Musacchio, A. (2002). How SH3 domains recognize proline. *Adv. Protein. Chem.* 61:211–268.

Patani, G.A., and LaVoie, E.J. (1996). Bioisosterism: A rational approach in drug design. *Chem. Rev.* 96:3147–3176.

Pawson, T., and Scott, J.D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. *Science* 278:2075–2080.

Robinson, R.C., Turbedsky, K., Kaiser, D.A., Marchand, J.B., Higgs, H.N., Choe, S., and Pollard, T.D. (2001). Crystal structure of Arp2/3 complex. *Science* 294:1679–1684.

Ruhlmann, A., Schramm, H.J., Kukla, D., and Huber, R. (1972). Pancreatic trypsin inhibitor (Kunitz). II. Complexes with proteinases. *Cold Spring Harb. Symp. Quant. Biol.* 36:148–150.

Sadowski, I., Stone, J.C., and Pawson, T. (1986). A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of fujinami sarcoma virus p130$^{gag-fps}$. *Mol. Cell. Biol*. 6:4396–4408.

Sattler, R., and Tymianski, M. (2001). Molecular mechanisms of glutamate receptor-mediated excitotoxic neuronal cell death. *Mol. Neurobiol.* 24:107–129.

Sauer, F.G., Fütterer, K., Pinkner, J.S., Dobson, K.W., Hultgren, S.J., and Waksman, G. (1999). Structural basis of chaperone function and pilus biogenesis. *Science* 285:1058–1061.

Sawyer, T., Bohacek, R.S., Dalgarno, D., Eyermann, C.J., Kawahata, N., Metcalf C.A., III, Shakespeare, W., Sundaramoorthi, R., Wang, Y., and Yang, M.G. (2002). Src Homology 2 inhibitors: peptidomimetic and nonpeptide. *Mini Rev. Med. Chem.* 2:475–488.

Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., et al. (2004). Global mapping of the yeast genetic interaction network. *Science* 303:808–813.

Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T., and Ramakrishnan, V. (2000). Structure of the 30S ribosomal subunit. *Nature* 407:327–339.

2

Yeast Two-Hybrid Protein–Protein Interaction Networks

Daniel Auerbach and Igor Stagljar

ABSTRACT

The availability of complete genome sequences of numerous model organisms has initiated the development of new approaches in biological research to complement conventional biochemistry and genetics. Consequently, high-throughput methodologies also need to be applied in the emerging field of proteomics. Here, we discuss several methods that have been developed in the past years in order to characterize proteins and their functions on a large scale. We focus on the yeast two-hybrid system, which is the most widely used method to study protein–protein interactions and which has been used several times now to sucessfully map entire interaction networks on a large scale. We discuss small-scale pilot projects and how they have been upscaled to genome-wide screens, such as for the budding yeast *Saccharomyces cerevisiae*. We then compare the yeast two-hybrid system with several other screening methods that have been developed to investigate interactions between proteins in a high-throughput format, such as affinity purification methods coupled to mass spectrometry. Efficient adaptation of such methods to a high-throughput format, coupled with the increasing use of databases to compare interaction maps generated with different methods, will help in elucidating protein–protein interactions on a scale that would have been unthinkable just a few years ago.

DANIEL AUERBACH • Dualsystems Biotech Inc., Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.
IGOR STAGLJAR • Institute of Veterinary Biochemistry and Molecular Biology, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.

Proteomics and Protein–Protein Interactions: Biology, Chemistry, Bioinformatics, and Drug Design, edited by Waksman. Springer, New York, 2005.

1. INTRODUCTION

The availability of complete genome sequences of numerous model organisms has initiated the development of new approaches in biological research to complement conventional biochemistry and genetics. For example, only one third of all 6200 predicted yeast genes had been functionally characterized when the complete sequence of the yeast genome first became available (Goffeau et al., 1996). At present, only 3800 yeast genes have been characterized by genetic or biochemical means and there still remain approximately 1800 genes encoding proteins of unknown function (Kumar and Snyder, 2001). The same observation holds true for the human genome: approximately 80% of all predicted human genes have not been characterized to date (Aach et al., 2001). To answer this challenge, researchers have developed different high-throughput strategies to characterize unknown genes on a large scale.

To date, most interaction maps have been created by genetic screening in yeast, namely by using the yeast two-hybrid system (Fields and Song, 1989). The reasons for the success of the yeast two-hybrid system in large-scale screening projects are manifold: as an in vivo genetic screening system, it is easily scalable, no purification steps or optimizations with regard to binding or washing conditions are involved, and automatization using robotic platforms is very easy. On the other hand, false positives and false negatives remain a problem of the yeast two-hybrid system; consequently, large-scale interaction maps derived by such methods require stringent selection criteria to yield useful information. Below, we first discuss the protein–protein interaction maps from various organisms that have been created using the yeast two-hybrid system and then discuss the advantages and disadvantages of this method. Finally, we briefly describe what has been done in analyzing those interaction maps to date.

2. THE YEAST TWO-HYBRID SYSTEM

The yeast two-hybrid system originally created by Fields and Song is a genetic system wherein the interaction between two proteins of interest is detected via the reconstitution of a transcription factor and the subsequent activation of reporter genes under the control of this transcription factor (Fields and Song, 1989). As depicted in Figure 2.1A, a protein X is expressed as a fusion to a DNA binding domain (DBD). The DBD–X fusion is commonly termed the "bait." Because of the affinity of the DBD for its operator sequences the bait is bound to a promoter element upstream of a reporter gene but does not activate it because it lacks an activation domain. A second protein Y is expressed as a fusion to an activation domain (AD) and is commonly termed the "prey." They prey is capable of activating transcription but usually does not do so because it has no affinity for the promoter elements upstream of the reporter gene (Fig. 2.1B). If bait and prey are coexpressed and the two proteins X and Y interact, then a functional transcription factor is reconstituted at the promoter site

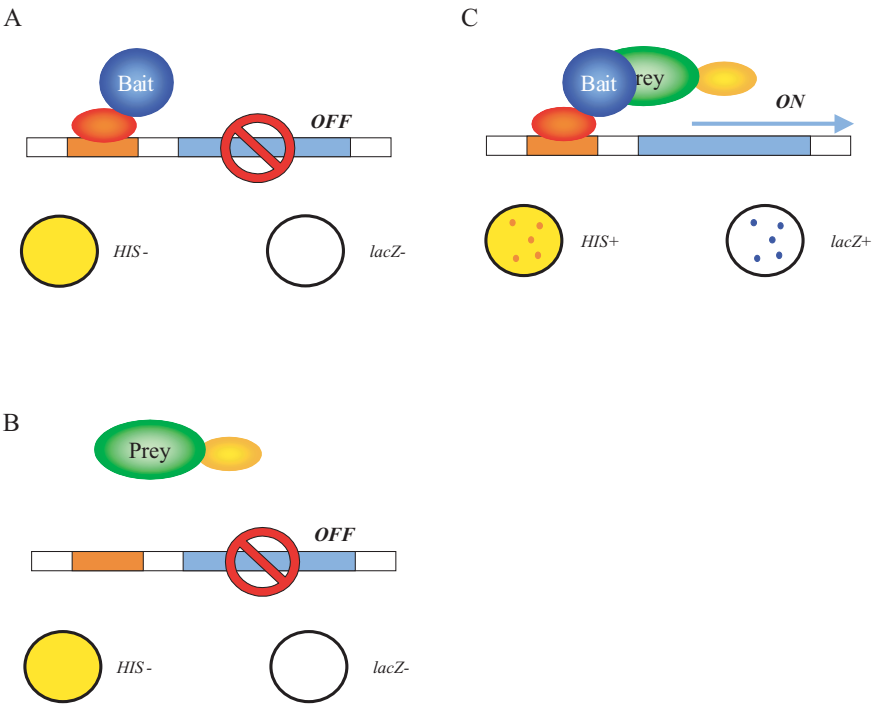


Figure 2.1. The yeast two-hybrid system. (**A**) A bait is expressed as a fusion to a DNA binding domain (DBD), for example, the *Saccharomyces cerevisiae* GAL4 transcription factor or the *Escherichia coli* LexA protein. The DBD–bait binds to the operator sequences present in the promoter region upstream of the reporter gene but does not activate its transcription since the DBD–bait does not contain an activation domain. (**B**) A prey is expressed as a fusion to an activation domain (AD), for example, from the GAL4 transcription factor or from the Herpes simplex virus protein VP16. The AD–prey fusion has the capability to activate transcription in yeast but because it is not actively targeted to the promoter it does not activate transcription of the reporter gene. (**C**) The interaction between bait and prey targets the AD–prey fusion protein to the promoter, thereby reconstituting an active transcription factor. The hybrid transcription factor is bound to the promoter upstream of the reporter gene and therefore activates transcription. The readout of the activated reporter gene is measured either as growth on selective medium (auxotrophic selection markers, such as *HIS3*, *URA3*, or *ADE2)* or in a color reaction *(lacZ)*. Yeast expressing only the DBD–bait or the AD–prey on its own do not grow on selective medium (*HIS*−) and do not display blue staining in a color assay (*lacZ*-), whereas yeast harboring an interacting DBD–bait and AD–prey display growth (*HIS*+) and blue color (*lacZ*+).

upstream of the reporter gene. Consequently, transcription of the reporter gene is activated. Thus, in a yeast two-hybrid assay a protein–protein interaction is measured through the activation of one or several reporter genes in response to the assembly of a transcription factor by the said protein–protein interaction (Fig. 2.1C). In common yeast two-hybrid screening schemes the prey is usually replaced by a collection of unknown preys expressed from a cDNA or genomic library. Screening of entire libraries against a defined bait may then lead to the discovery of novel interaction

partners. For large-scale screenings, two approaches are commonly used: the library screening approach, in which multiple baits are screened against a library, and the matrix approach, in which an array of defined preys is substituted for the library.

3. LARGE-SCALE SCREENS USING THE LIBRARY APPROACH

The library approach is schematically shown in Figure 2.2A. A particular bait is expressed in a yeast reporter strain of the mating type a, whereas a collection of preys (the library) is transformed into a yeast reporter strain of the mating type α. The bait-bearing strain is then mated with the mixture of library strains, and clones expressing an interaction pair are isolated on selective media. To determine the identity of the interacting prey, the library plasmid encoding it has to be isolated from the yeast strain and amplified in *Escherichia coli*. The region encoding the prey is then sequenced.

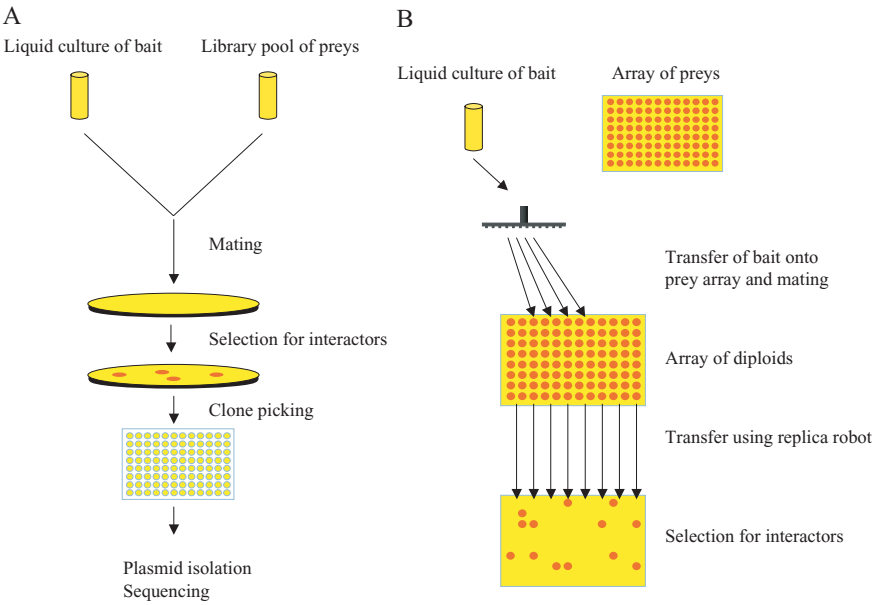


Figure 2.2. High-throughput approaches utilizing the yeast two-hybrid system. (**A**) The library screening approach. A yeast strain expressing a bait under investigation is mixed with a collection of yeast strains each expressing a random prey from a library. Incubation in rich medium allows the two strains to mate and diploids expressing bait and prey are selected. The diploids are then transferred to selective medium to isolate those clones containing interacting baits and preys (selection for interactors). Yeast clones that display growth on selective medium are picked up, transferred into multiwell plates, and processed for plasmid isolation and insert sequencing to identify the interacting prey. (**B**) The matrix or array approach. An array of preys is prepared by spotting yeast clones each expressing a known prey onto plates. The colonies on the array are then picked up by a robot and mated with a yeast strain expressing the bait under investigation. An exact replica of the array is transferred to a fresh plate to select for diploids expressing bait and prey and then to selective medium to select for interacting baits and preys. The identity of the prey in colonies that grow under selection is determined by its position within the array.