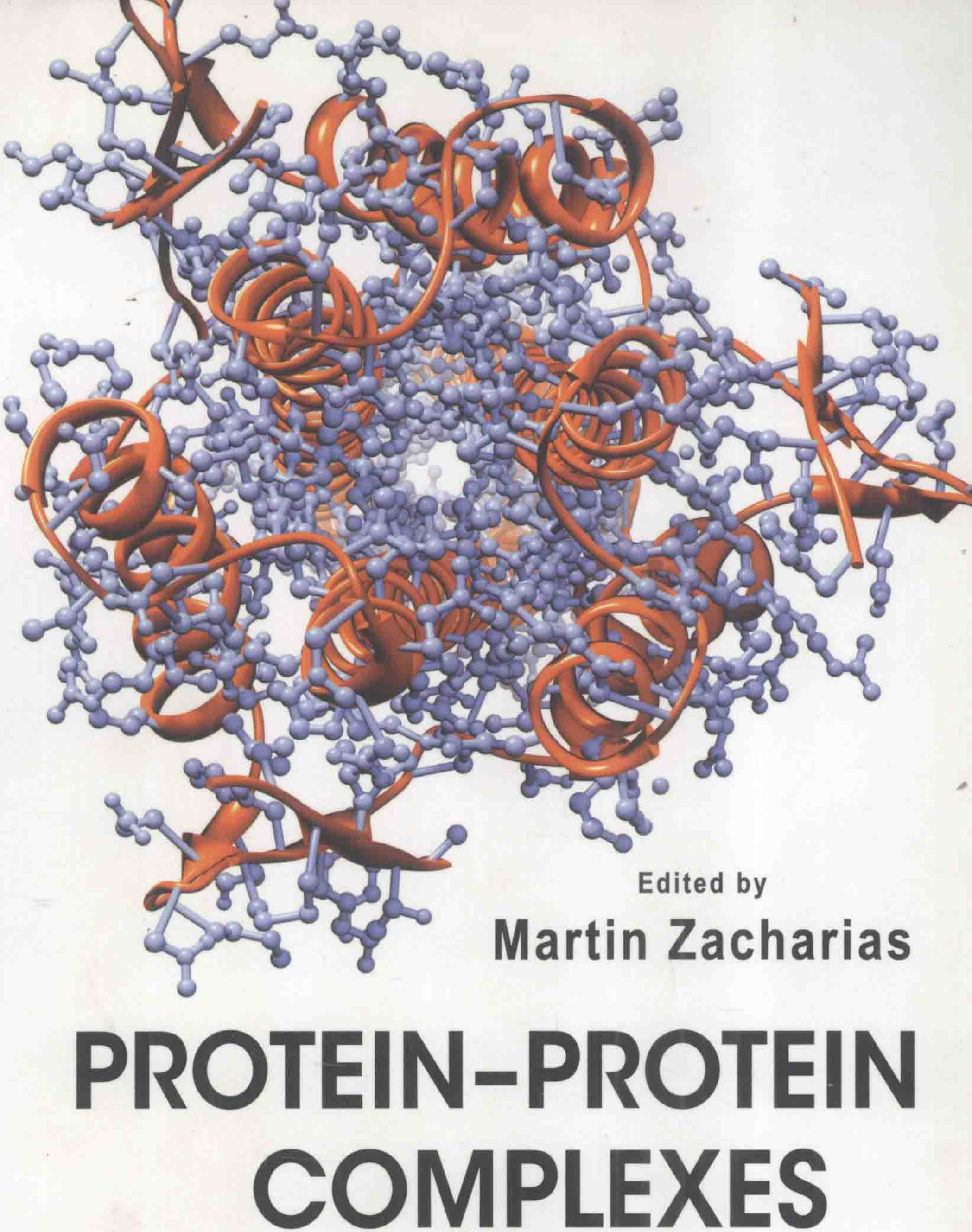


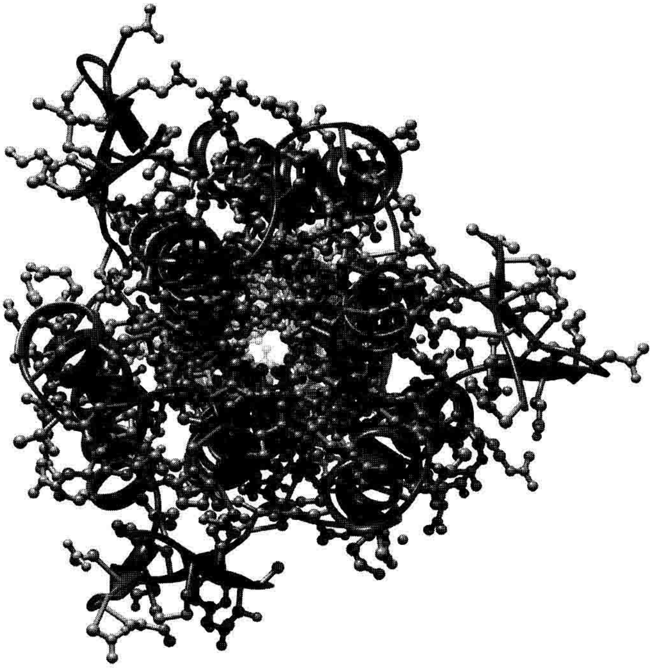
Edited by

Martin Zacharias

PROTEIN–PROTEIN COMPLEXES

Analysis, Modeling and Drug Design

Imperial College Press



PROTEIN–PROTEIN COMPLEXES

Analysis, Modeling and Drug Design

Edited by

Martin Zacharias
Technische Universität München, Germany

ICP
Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

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

PROTEIN-PROTEIN COMPLEXES
Analysis, Modeling and Drug Design

ISBN-13 978-1-84816-338-6
ISBN-10 1-84816-338-X
ISBN-13 978-1-84816-339-3 (pbk)
ISBN-10 1-84816-339-8 (pbk)

Printed by FuIsland Offset Printing (S) Pte Ltd. Singapore

PROTEIN–PROTEIN COMPLEXES

Analysis, Modeling and Drug Design

PREFACE

Basically all processes in a cell involve proteins and the great majority of biological functions are mediated not only by isolated proteins but also by the interaction of proteins. Powerful experimental techniques are available to systematically investigate the network of protein–protein interactions in cellular systems. However, for a full understanding of protein–protein interactions, knowledge of the three-dimensional structure of complexes formed between interacting proteins is essential. Immense progress has been achieved in recent years to elucidate protein–protein complex structures and to better understand the physical principles of complex formation. What are the driving forces for protein–protein association? What can we learn about specific recognition from studying protein–protein interfaces? How can this knowledge be used to predict protein–protein interactions and is it possible to influence protein–protein interactions by small drug molecules? These and many other questions will be tackled in the 13 chapter contributions in this volume.

Although the book covers the state-of-the-art research in the area of protein–protein complex analysis and modelling, it is not primarily directed at specialists in the field. The book is also meant to be a useful guide for students and researchers in the area of Chemistry, Biochemistry and Biophysics with an interest in proteins and protein–protein interactions. Most chapters contain significant introductory information in addition to the most recent progress in the field. Readers will gain insight into the recognition principles of proteins; how to determine, analyse and predict protein–protein interactions and complex structures, as well as learn about possibilities of interference with protein–protein interactions.

Leading researchers in the field have been selected to contribute chapters to the book. Authors were free to select the exact scope of their contribution and express their own view on the field. Possible overlapping between chapters can be profitable for the reader since key information is provided from different perspectives by leading scientists.

The first part of the volume introduces the analysis of experimentally determined structures of protein–protein complexes. Experimental protein structures contain rich information on the principles of interaction. The systematic analysis of the interface region of protein–protein complexes and the comparison with other surface regions of a protein reveal the physical characteristics of protein binding sites. A deeper understanding of the driving forces of protein–protein complex formation also requires an analysis of the thermodynamics of protein–protein association. The first part of the book includes an overview of experimental methods to investigate the thermodynamics of protein–protein binding, and also discusses theoretical methods to calculate energetic and entropic contributions. The study of the kinetics of association and dissociation of protein–protein interactions is of central importance to understanding the mechanism of protein complex formation. How the kinetics of protein–protein binding can be studied experimentally and theoretically is at the focus of a separate chapter. Proteins bind to specific sites on the surface of proteins with high affinity. The physico-chemical character of binding sites can differ from the properties of other surface regions. In addition, often the amino acids at protein binding sites are evolutionarily more conserved then the rest of the protein surface. The properties and conservation of protein functional sites and how they can be used to identify relevant amino acid residues for protein–protein recognition are discussed in the fifth chapter.

Due to the large number of putative protein–protein interactions and the transient nature of many protein–protein complexes, only a fraction of possible protein–protein complex structures can be determined experimentally. A variety of computational docking prediction methods have been developed in recent years to tackle the problem of providing at least structural models of important protein–protein complexes. A general overview of docking methods is provided, followed by chapters on how to best include experimental data or information from bioinformatics resources to high-resolution docking methodologies. Typically, modelling protein–protein complex structures is not a one-step procedure but instead distinguishes an initial exhaustive search followed by a refinement and rescoring phase. The options of refining and identifying the most realistic predicted complex structure are also introduced.

The last five chapters of the volume shift the focus from three-dimensional modelling of protein–protein interactions towards approaches that influence or interfere with protein–protein interactions. A significant fraction of protein–protein interactions – particularly in higher organisms – are mediated by reoccurring motifs or interaction patterns. Chapter 10 gives an overview of several examples of biological and medical importance. The chapter also includes a discussion of the involvement of motif-mediated interactions in diseases. Mutations in proteins may perturb interactions with other partners. However, site-directed mutagenesis can also be used to redesign protein binding regions to create new or altered protein–protein interactions. Methods to estimate changes in protein–protein affinity, due to residue substitutions at the interface, are described and the possibility to directly and specifically interfere with protein–protein interactions is at the focus of two separate chapters. The concepts are introduced and discussed on examples that are of relevance to several human diseases. Proteins can undergo conformational changes upon association. In addition, the binding process can also influence the flexibility of binding partners which may even mediate long-range allosteric communication. The analysis of such dynamical recognition processes and the possibility to influence them by drug-like molecules is the subject of the last chapter.

It is my great pleasure to thank all authors for the time and efforts they devoted to the demanding work of contributing book chapters to this volume. I am grateful to the editors of Imperial College Press for their cooperation and also to my co-workers and family for their patience and support.

Munich, July 2009

Martin Zacharias

CONTENTS

CHAPTER 1

X-ray Study of Protein–Protein Complexes and Analysis of Interfaces

Joel Janin
Yeast Structural Genomics, IBBMC UMR 8612 CNRS,
Université Paris-Sud, 91405 Orsay, France
E-mail: joel.janin@u-psud.fr

Highly efficient procedures to express genes and prepare individual proteins for structural analysis, developed during the first round of the Structural Genomics initiatives world wide, are now being extended to protein complexes and multi-subunit assemblies. These structures are still few in the Protein Data Bank, but one can exploit the abundant information on binary protein–protein complexes and oligomeric proteins to set up appropriate methods of analysis, and derive rules on protein–protein interaction, which will be applicable to larger assemblies when their structures become available.

1.1 Introduction

Following the completion of the first complete genome sequences at the turn of the century, the question was put to structural biologists: can crystallography and NMR provide three-dimensional structures for the products of all these genes? At that time, it was estimated that a set of 10,000 experimental structures, carefully chosen, would cover the space of existing folds; the remainder could be built by homology.[1] Structural Genomics (SG) initiatives were launched in the USA and Japan in the years 2000–2001, with that goal. With the end of 2009, they will have deposited more than 8,000 new structures in the Protein Data Bank (PDB, *http://www.rcsb.org/pdb/statistics/*), and the target of 10,000 will almost certainly be reached before 2010. But meanwhile, the landscape around has changed greatly. We now realise that the diversity of DNA sequences may be orders of magnitude greater than what was thought when only a few model genomes were known. Many of the new sequences are unrelated to what we have in the databases, and therefore, many protein folds have yet to be discovered. Moreover, it has become clear that most gene products do not exist and function as single entities. Genome-wide studies of protein–protein interaction have demonstrated that cells contain thousands of macromolecular assemblies of all sizes, from simple dimers to objects that comprise tens or hundreds of polypeptide and/or nucleic acid chains.[2,3] The examples of the ribosome and the nuclear pore show that the whole assembly, not the individual chains, carries the biological function. The structural analysis should, therefore, not be limited to the isolated components.

The number of solved macromolecular assembly structures is still small compared to that of isolated proteins.[4] In this review, attempts will be described to characterise macromolecular assemblies similar to the systematic studies that SG initiatives performed on single proteins. While these studies are ongoing, we may look at simpler systems for which the PDB offers more examples: protein–protein complexes and homodimeric proteins. Their atomic structures contain a wealth of information on the chemistry and physical chemistry of the non-covalent interactions that allow polypeptide chains recognising each other and self-assembling into a functional macromolecular entity.[5–9] The methods developed to extract this information, the observations and rules derived from its analysis, will undoubtedly help us to understand the more complex systems when their structure becomes available.

1.2 Preparing Proteins for Structural Studies

The first genome-wide studies of protein–protein interactions were completed at about the same time as the SG initiatives of the first generation. As a result of that coincidence, the second generation of SG initiatives that started in 2005–2006, included several programmes that are concerned with macromolecular assemblies.[10–11] Thus, the Yeast Structural Genomics, a small-scale pilot-project that we carried out in

Orsay in 2001–2004, is now part of two programmes funded by the European Union, SPINE2-Complexes and 3D-Repertoire (*http://www.spine2.eu*, *http://www.3drepertoire.org*). Both combine high-resolution X-ray/NMR and medium/low resolution cryo-electron microscopy studies (cryo-EM) in order to study multi-component systems; some of their targets, like RNA polymerase or the exosome that degrades mRNA, have a well-established status in biology. Others have just been identified in systematic tandem-affinity purification/mass spectrometry studies. These complexes have no known function, but with yeast, a wealth of genetic and biochemical tools are available to characterise them while the structural analysis is ongoing. Atomic resolution may not be reachable for some of the targets, but useful models can be obtained by docking into the electron density of cryo-EM images, the high-resolution models obtained by X-ray crystallography on some of the components.

All these studies integrate the expertise acquired by labs that were part of the first round of Structural Genomics initiatives to which they owe many of their tools and first of all, efficient methods to produce and analyse recombinant proteins.[12] Figure 1.1 describes the standard procedure that was set up to express and prepare proteins of *Saccharomyces cerevisiae* during the four years of the Yeast Structural Genomics pilot-project.[13] It comprises three major steps:

1. *Cloning:* We use the PCR reaction to amplify the target sequence in genomic DNA (mostly intron-free in *S. cerevisiae*); the two primer oligonucleotides contain appropriate restriction sites and the 3'-primer codes for a six-histidine tag placed just after the last codon. The PCR products are purified, digested with restriction enzymes and inserted into an expression vector. Their DNA sequence is checked. In *E. coli*, we use vectors derived from the pET plasmid, which place the target gene under control of the highly efficient phage T7 promoter.

2. *Protein Production:* The level of gene expression and the solubility of the target protein are evaluated in small-scale cultures of several *E. coli* strains, each grown at four different temperatures. The conditions that yield the most soluble protein are retained for large-scale production in 1 litre flasks.

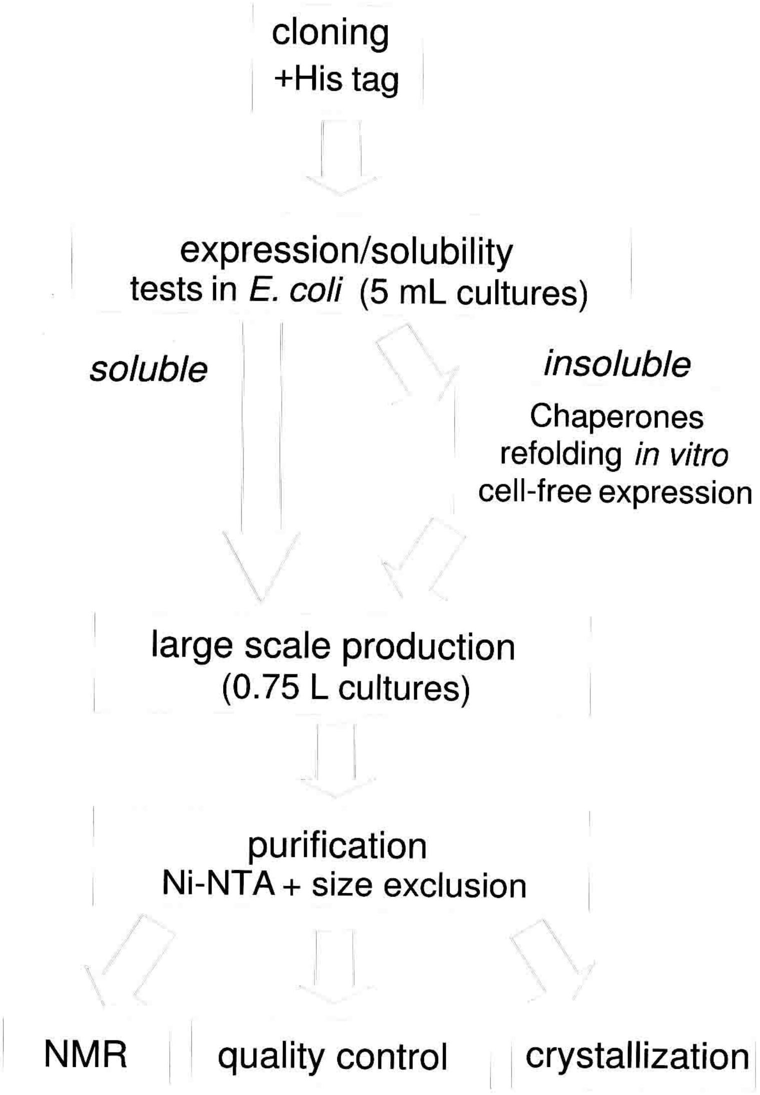


Fig. 1.1. Flowchart of the protein expression/purification procedure. During the Yeast Structural Genomics pilot-project, 250 *S. cerevisiae* genes were cloned and tagged in a standard protein preparation procedure. Expression in *E. coli* succeeded for 80% of the proteins with less than 350 residues. Soluble protein could be purified in two steps from the cell extract, and insoluble protein could be recovered in a number of cases from inclusion bodies (adapted from Ref. 13).

3. *Protein Purification and Quality Controls:* The His-tagged protein is purified on a Ni-NTA resin, concentrated and run on a size exclusion column. Its degree of purity (usually > 95%) is judged by electrophoresis on a SDS gel and its chemical integrity by mass spectrometry.

The cloning step was carried out on 250 *S. cerevisiae* target genes with a success rate above 90%. After optimization of the growth conditions, most of the cloned genes were highly expressed in *E. coli*; an overnight culture in a shaken flask yielded the target protein in milligram quantities. However, more than one-third of the constructions gave insoluble protein in inclusion bodies. About half of those could be recovered as soluble protein either by co-expressing bacterial chaperones, by solubilizing the inclusion bodies in 6 M guanidinium chloride and screening for refolding in a number of buffers,[14] or by using a cell-free expression system.[15]

Carrying out the whole procedure on all the targets was outside the scope of a pilot-project, and therefore, we focused our work on a subset of proteins of interest. Starting with 140 well-expressed yeast genes, we obtained 72 proteins purified to homogeneity in quantities of 0.5 to 10 mg that could be subjected to automated crystallization screens. A majority of the screens gave crystalline hits, not always of sufficient quality for structure determination, but some of these leads could be optimised as discussed below. Fourteen proteins had their X-ray structure determined to resolutions of 1.3 to 2.6 Å within the four-year course of the pilot-project[16] (*http://genomics.eu.org/spip/Overview*), and another ten during the two years after. Therefore, the goal of 20 new structures that we had initially fixed to the pilot-project had been reached by 2006, leaving the place for new projects mostly concerned with protein–protein complexes.

Other SG centres have had success rates similar to ours, often on a much larger scale.[17] The second generation programmes that opened in 2005 in the US and Japan, have built on that experience to set up high-throughput production chains for the structure determination of single gene products by both X-ray crystallography and NMR. Whereas most of the first-generation targets were from prokaryotes or yeast, more difficult targets from higher eukaryotes and including membrane proteins are now being addressed, albeit with a much lower throughput.[12,18]

1.3 Preparing Protein–Protein Complexes and Multi-component Assemblies

The preparation scheme of Fig. 1.1 has a success rate of 50% that may be considered as satisfactory on a target that is a single gene product. The same scheme can be used to produce multigenic protein assemblies by preparing each component separately. For a binary complex, the expected 25% yield makes it worth trying, but with more than two components the chance is poor that all the subunits can be prepared separately as soluble proteins that will self-assemble when mixed together. Nevertheless, the one-by-one approach has had some remarkable successes. For instance, the *Xenopus* genes coding for the four different histones that constitute the nucleosome core particle could be individually expressed in *E. coli*, and the core particle was reconstituted by mixing them together in appropriate proportions.[19] More frequently, some but not all of the components of a multi-component complex are obtained in soluble form. The complex itself cannot be reconstituted, but some of the soluble components form subcomplexes that can yield important information on the assembly, and they may be suitable for high-resolution structural studies complementing a cryo-EM analysis of the whole complex.

Figure 1.2 describes the strategy that we developed for preparing yeast protein–protein complexes. It offers several alternatives to the one-by-one gene expression approach (Pathway 3). One possible approach is to prepare the assembly directly from yeast extracts, either at its natural abundance (Pathway 1) or after over-expressing all its components (Pathway 2). Over-expression can also be attempted in *E. coli* (Pathway 4). Pathway 1 is the one that was used in the structural studies of bacterial ribosomes, and also of the yeast 20S proteasome.[20] The cells can be grown in large quantities, the ribosome and the proteasome are very abundant, and they can be purified by techniques that do not require affinity tags. In all other cases, the complexes must be over-expressed. A simple procedure would be to build an expression vector for each of the genes of interest, and introduce them into the same bacterial or yeast strain. However, it is difficult to maintain more than two plasmids in the same host, and even with a binary complex, the level of expression of two genes carried by different vectors is likely to be very unequal, compromising the formation of an assembly with a well-defined stochiometry. The approach that we and others favour is therefore to make operon-like genetic constructions, in which several genes of interest are placed next to each other.

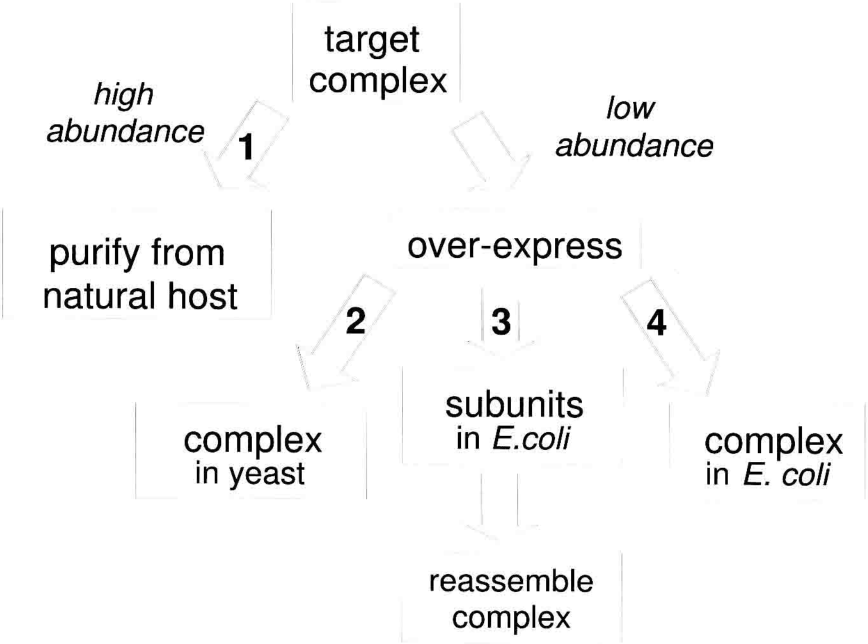


Fig. 1.2. Strategy for the purification of multi-subunit yeast complexes.

They form a single transcription unit under the control of the same promoter, and a ribosome binding site is placed between each stop and start codon.[21] In practice, five or more medium-size genes can be co-expressed in this way, one of them bearing an affinity purification tag. The construction can be facilitated by placing restriction sites at strategic locations, or dispensed of altogether by using synthetic DNA. The genes in the operon are transcribed into a single mRNA, they are translated at similar levels and their products are able to associate as they exit the ribosome. Thus, components that would be insoluble (or disordered and degraded) if expressed alone, can be rescued through their interaction with the partner chains. The procedure does not apply to systems such as the complexes of the respiratory chain, because their assembly requires specialized chaperones or cofactors. Still, the co-expression and self-assembly in *E. coli* of eukaryotic protein–protein complexes has had a remarkable success rate, and most of the methodological developments in progress follow Pathway 4.

1.4 Crystallization and X-ray Studies

Crystallization is a well-recognised bottleneck in structural studies. A number of new tools have been developed in recent years, mostly in SG labs. These techniques were designed primarily for single-gene products, but they work equally well for multi-component assemblies and play a key role in the present study. In spite of many attempts to make it rational, the crystallization of proteins, nucleic acids and their complexes still depends on testing hundreds of conditions that combine different precipitants, pHs and additives. One of the very first upshots of the SG initiatives, the one that spread the most quickly, was automatic crystallization. Unlike an attempt we had made[22] to use robotics in the early nineties, the devices and procedures that were developed ten years later in the framework of the SG centres immediately found industrial support and are now used routinely by the protein science community. Pipette robots and crystallization kits greatly facilitate the preparation of the precipitant solutions. Equally important, the amount of biological material required to do the tests has dropped by one or two orders of magnitude, thanks to liquid-dispensing robots that prepare arrays of nanodrops in 96-well plates.[23–24] A standard set of four plates can be prepared in a couple of hours with a minimum of human intervention, and it uses up only a milligram or two of pure protein material. Moreover, the success rate is remarkably high: in our hands, about half of proteins entering crystallization trials give crystals of some sort. As many are not suitable for diffraction experiments because of their size