Proteins in a DNA world: expression systems for their study

Jorge Mauricio Reyes-Ruiz, † Hugo Alberto Barrera-Saldaña*

* ULIEG-Facultad de Medicina de la UANL.

ABSTRACT

Every day, new proteins are discovered and the need to understand its function arises. Human proteins have a special interest, because to know its role in the cell may lead to the design of a cure for a disease. In order to obtain such information, we need enough protein with a high degree of purity, and in the case of the human proteins, it is almost impossible to achieve this by working on human tissues. For that reason, the use of expression systems is needed. Bacteria, yeast, animals and plants have been genetically modified to produce proteins from different species. Even “cell-free” systems have been developed for that purpose. Here, we briefly review the options with their advantages and drawback, and the purification systems and analysis that can be done to gain understanding on the function and structure of the protein of interest.

Key words. Expression systems. Functional proteomics. Protein purification. Bioreactors.

INTRODUCTION

DNA has been the center of attention for the past half century. The completion of the sequencing of the human genome has been the most important achievement of the human race in modern times and its implications will cause controversy for years to come for the ethical, social and even religious consequences that this discovery has brought. Now and then we read in newspapers or hear on the radio that a gene has been linked to a particular disease or affliction, even that some genes could be linked to criminal behavior, sexual preferences or drug-addictions. Consequently, we are gaining an understanding of the human genome, and will be able to attack these problems targeting the accused genes. But rarely in a headline is a protein the center of attention, when the proteins are actually the functional part of the great majority of the genes. Besides Mad Cow Disease, the importance of proteins is almost null in the everyday world. Fortunately, among scientists the story is different: with an actual estimate of about 30,000 genes and with as much as 90,000 – 120,000 pro-
**Figure 1.** Selection of the expression system. In order to make the right selection, one must consider the complexity of the protein to be expressed, the cost and the easiness of handling. Despite the low cost and easy manipulation of cell-free systems, bacteria and yeast, complex proteins will be hard to produce successfully; cell culture, on the other hand, will be able to produce complex proteins, but the cost and the need of specialized facilities make it less appealing, though necessary frequently. Animals and plants as bioreactors will produce great amounts of biologically active proteins, but the cost implied in the maintenance and other factors, such as the time needed to harvest and the cost of purification from complex tissues add up significantly.

**Figure 2.** Basic strategy to follow in protein studies. Choosing the right expression system can take some time and makes the difference between success and failure. Once the protein is expressed, a purification scheme can then be planned. The use of a TAG should not be overlooked. The ultimate goal will be to gain a major understanding of the function of the protein by performing bioassays and by exploring possible targets for drugs and structural composition.
Proteins coded by the human genome (estimates say that because of alternative splicing, a gene on average can code for 3 or 4 proteins) and the function of 50% of these proteins remaining unknown, more studies on proteins rather than DNA are required.

Proteins being the targets of the top-selling prescription drugs, proteomics is expected to make a profit of over $5 billion in sales by the end of 2005, according to the Proteome Society.2 Pharmaceuticals on the market target fewer than 500 human gene products3 suggesting that there is an enormous untapped pool of human gene-based targets for therapeutic intervention.

Efforts to understand the relationship between protein structure and biological function have intensified, and the huge number of candidate proteins rendered by functional genomics has generated interest in all aspects of protein expression and purification. Proper expression and purification techniques are essential for the large-scale production of pure proteins, which can be used in subsequent analyses, i.e. high-throughput screening and 3-D structure determination. In addition, expression and purification systems that rapidly yield high levels of pure recombinant proteins are fundamental for the identification of target molecules for drug development. The keys to efficient, high production expression systems are good host strains, vectors, and growth conditions. Bacterial, yeast, insect and mammalian cells are the most common types of expression systems used by researchers today.4

Once the protein is over expressed, detection methodologies as well as functional tests are used to follow up the protein during another critical step of study: purification.

We will briefly review the most common expression systems (Figure 1) and discuss alternatives for detection, biological assays and purification of recombinant proteins (Figure 2).

EXPRESSION SYSTEMS

Bacteria

The first choice for the heterologous production of a human protein is Escherichia coli, because of the low cost of culture media, the availability of a great variety of strains and expression vectors, their easy handling and acceptable production yields of recombinant proteins in a short time.5 The drawbacks of this system is that as a prokaryote organism, this bacterium is unable to complete some post-translational modifications, such as glycosylation and formation of disulfide bonds among others, required for many eukaryotic proteins to function properly. Likewise, its codon-usage differs from that of human cells,6,7,8 which leads to low yields9 or sometimes a truncated protein, a non-functional protein or no protein at all.10 But now there are some improved commercially available strains to overcome some of these limitations. For example, if the protein contains disulfide bonds, proper folding is stimulated with a more oxidizing cytoplasmic environment and Novagen11 has the AD4A4 and Origami strains mutated in the thioredoxin reductase and glutathione reductase genes to aid for this purpose. When the protein contains a high number of rare E. coli codons, one can try to express it using a strain that co-expresses the tRNAs for these rare codons, for example: CodonPlus-RIL, CodonPlus-RP and Rosetta are strains available from Stratagene12 and Novagen that co-express some rare tRNAs. The regulation of expression can be constitutive or triggered by the addition of an inducer such as IPTG or L-arabinose, or just by a shift in temperature14, to turn on the promoter (depending on the vector used). The time to harvest the expressed protein varies between 3 to 12 hrs, depending on the expression levels achieved, or if the protein is toxic for the host. Other bacteria have been used as expression hosts (for example, B. subtilis,15 Lactococcus, Lactobacillus, Leuconostoc, Pediococcus and Streptococcus spp16) but E. coli remains the most widely used.

Yeast

Next in the list of the most commonly used expression hosts is an expanding group of yeasts. Yeasts are attractive hosts for production of mammalian proteins requiring posttranslational processing, because they offer advantages of both eukaryotic biosynthetic mechanisms and bacteria-like growth and handling.17 Thus for the past years yeasts have been used to overcome the shortcomings of bacterial expression systems. They can provide intracellular as well as extracellular expression by applying short signaling sequences.18 Saccharomyces cerevisiae is usually the first yeast of choice and since it tends to hyper-glycosylate secreted proteins, it is primarily used for intracellular production.19 Other yeast strains (i.e., Pichia pastoris, Hansenula polymorpha, Schizosaccharomyces pombe, Klyveromyces lactis, and Yarrowia lipolytica) are more suited for secretion.19,20 Among the best-developed systems for large-scale protein production are those based on P. pastoris.21 Notable advantages of P. pastoris compared with S. ce-
revisiae are that the former may produce more abundant heterologous protein 10-100-fold and it does not hyperglycosylate them as the latter does upon secretion. Both yeasts have a majority of N-linked glycosylation of the high-mannose type, but in Pichia the average length of the oligosaccharide chain is 8-14 mannose residues per side chain, whereas in Saccharomyces it is between 50-150 mannosides. In addition, core oligosaccharides in S. cerevisiae have terminal α1,3 glycan linkages (believed to be responsible for the antigenic nature of proteins produced by secretion from S. cerevisiae), while P. pastoris has none.22

One of the most important drawbacks of yeast is the cell wall if the protein is not to be secreted. Breaking the cell wall and recovering the interior of the cell intact can be a major headache. The typical way to do it is by using glass beads, but this has limitations on the size of the sample. For a larger sample, a French Press capable of reaching 24,000 psi or a Bead Mill are better options.

**Mammalian cultured cells**

Mammalian cells in culture are the next choice. By using a cell line from higher eukaryotes one can usually avoid most of the problems encountered in the bacteria or yeast systems, but at a steep price: the cost is sky-high when compared to bacteria or yeast cultures, handling requires trained personnel and appropriate facilities, there are safety risks and the yield is not necessarily high. Nevertheless, mammalian host cell lines, although non-human, are the protein production vehicles of choice as they are capable of producing proteins with almost human-like glycosylation. Glycosylation is of particular importance for the recognition of “foreign” proteins by the immune system of higher mammals. Pharmacokinetic analysis has shown that insufficient or inconsistent glycosylation results in accelerated clearance of these proteins by carbohydrate specific receptors in the human liver preventing them from fulfilling their therapeutic purpose.24

**Insect cells**

Insect cells are valued for their ability to express some recombinant proteins at a higher level than most mammalian expression systems.25 The baculovirus-based system is a eukaryotic expression system and is therefore able to produce overexpressed recombinant proteins with proper folding, disulfide bond formation and oligomerization.26 This system is also capable of performing many post-translational modifications including O-linked and N-linked glycosylation, phosphorylation, acylation, amidation, carboxymethylation and cleavage of certain proteins to their active forms.27 The most frequently used insect cells that are susceptible to baculovirus infections are SF9 and SF21 cell lines. Both of these lines are originally established from ovarian tissues of Spodoptera frugiperda larvae. These cell lines may be grown in suspension and can therefore be used in a bioreactor.28 High Five cells, derived from Trichoplusia ni egg cell homogenates, have been shown to be capable of expressing significantly higher levels of secreted recombinant proteins compared to other insect cells. They can be transformed by baculovirus or by using liposomes.

In contrast to the baculovirus system, the Drosophila system relies on stable cell lines that express the foreign protein.29 Since the Drosophila cells will spontaneously incorporate hundreds of copies of the transfected genetic information, the expression levels should be very high, and no extensive testing of transfected clones for expression is required. Stable cell lines can be established after co-transfection with a plasmid that confers resistance to the antibiotic hygromycin B.30 The expression of the foreign protein is under the control of the metallothionein promoter, and the system can be induced by the addition of non-toxic concentrations of metal ions such as copper.30

**OTHER INTERESTING ALTERNATIVES**

**Cell-free systems**

Cell-free protein expression removes the need for cell culture: protein is expressed in vitro using cellu-
lar extracts that contain all the machinery and biochemical constituents required for transcription and translation. The cellular extract may be obtained either from prokaryotic or eukaryotic cells.31 The use of in vitro translation systems can have advantages over in vivo gene expression when the over-expressed product is toxic to the host cell, when the product is insoluble or forms inclusion bodies, or when the protein undergoes rapid proteolytic degradation by intracellular proteases. In principle, it should be possible to prepare a cell-free extract for in vitro translation of mRNAs from any type of cell. In practice, only a few cell-free systems have been developed for in vitro protein synthesis. The most frequently used cell-free translation systems consist of extracts from rabbit reticulocytes, wheat germ and
Escherichia coli. All are prepared as crude extracts containing all the macromolecular components (70S or 80S ribosomes, tRNAs, aminoacyl-tRNA synthetases, initiation, elongation and termination factors, etc.) required for translation of exogenous RNA. To ensure efficient translation, each extract must be supplemented with amino acids, energy sources (ATP, GTP), energy regenerating systems (creatine phosphate and creatine phosphokinase for eukaryotic systems, and phosphoenol pyruvate and pyruvate kinase for the E. coli lysate), and other co-factors (Mg$^{2+}$, K$^+$, etc.). There are two approaches to in vitro protein synthesis based on the starting genetic material: RNA or DNA. Standard translation systems, such as reticulocyte lysates and wheat germ extracts, use mRNA as a template; whereas “coupled” and “linked” systems start with cDNA templates. Escherichia coli-coupled transcription-translation system yields as much as 6 mg of protein per milliliter of reaction volume.

**Plants**

The use of plants as expression systems for recombinant proteins is currently receiving a lot of attention all over the world. Plant expression systems have advantages over other in vitro expression systems in terms of low production costs. In addition, contamination of endotoxin and animal viruses has been associated with bacterial and eukaryotic cultured cell expression systems for recombinant proteins. Although transgenic plants may contain plant-specific viruses or bacteria, they are not known to be harmful to humans and animals. Thus, using the plant expression systems, several biologically active proteins and peptides for potential pharmaceutical applications, such as vaccine antigens and immunomodulators have been expressed in plants. However, the production of recombinant proteins in higher plants has drawbacks, which can make this system less attractive from an economic standpoint. The first drawback is the substantial length of time required from the initial transformation event to small-scale evaluation and production: often 2 years or more. This length is primarily due to the relatively slow growth rates of terrestrial plants relative to other organisms used for protein expression. Hence, the initial generation of transformants, their propagation to flowering, genetic crosses, and production of seed stocks are handicapped by this intrinsically slow rate of growth. Higher plant expression systems are also associated with complex processing issues, because recombinant proteins are produced and deposited in specific organs such as leaves, fruits, and seeds. These proteins must be purified to homogeneity out of a complex mixture of tissues and cell types, a requirement that can add significantly to the costs of purification.

**Animals as bioreactors**

Expression of recombinant human proteins in the milk of transgenic dairy animals offers a source of clinically important proteins that cannot be produced as efficiently in adequate quantities by other methods. Milk expression has been reported for at least 17 different proteins in five livestock species, 11 of them at commercially feasible levels of >1 g/l. Concomitant advances have been made in purifying proteins from raw milk. Heterologous proteins have been expressed in the obvious dairy species: cows, sheep and goats. However, pigs and rabbits are also used for selected applications because of their large litters and shorter generation times. The size of the production herd required for a particular application will depend primarily upon total annual needed, the recombinant-protein expression level and the recovery efficiency.

Another approach is the use of the hen, which promises to be a low cost, high-yield bioreactor. More than half of the egg white protein content derives from the ovalbumin gene with four other proteins (lysozyme, ovomucoid, ovomucin and conalbumin) present at levels of 50 milligrams or greater. Taking advantage of the promoter of this gene to express a recombinant protein could bring yields of up to a gram or more. Since modern layers produce 300 eggs per year, three or four hens producing one gram per egg would yield a kilogram of raw product annually. The naturally sterile egg also allows a long shelf life of recombinant protein without loss in activity. In spite of these advantages, transgenic procedures for the bird have lagged far behind those of other organisms. Gene transfer into the avian genome has been achieved by the use of retroviruses, microinjection of DNA into cytoplasm of fertilized zygotes, and the use of Drosophila transposons. However, the production levels have been only 3 to 38 μg of protein per egg. Hence, development of more robust and efficient transgenic methods will be necessary before at-will modification of avian genetic material is feasible.

**DETECTION**

Once the protein is expressed, one must find a way to detect it throughout the procedure. Western
Purification

Once we have expressed the protein at satisfactory levels and are able to detect it from the raw sample, one must develop a purification strategy. This is true specially if the purpose of our research is to uncover physical characteristics related to function, or simply because we need it in a purer form (to be injected into experimental animals). It is a prerequisite to know as much as possible of its physical properties (such as isoelectric point). Then one can start to plan which of the available chromatography protocols to use. First, some differential centrifugations or precipitations may be needed to enrich a fraction and to put the sample in an appropriate state to interact with the solid phase (resins), i.e. resuspend the proteins in a suitable buffer. It is usually very important to work fast, using protease inhibitors and proceeding at 4 °C to avoid unwanted degradation of the sample. For protein separation, high performance liquid chromatography (HPLC) offers the best resolution, and may be very useful to separate proteins from a complex mixture. However, preparative HPLC is not the first choice, mainly because of limitations in the size of the sample that can be injected, not to mention the high price of the preparative columns. Moreover they are easily dogged (leaving them useless), and the pumps required for the preparative procedures are very expensive. For the above reasons, fast protein liquid chromatography (FPLC) remains the most used method in preparative chromatography, despite its lower resolution.

Again, the use of tags is a very powerful tool in the purification procedure. The main commercially available affinity resins have high quality standards and by simply modifying the expression vectors one can add a tag suitable for purification. The most common are: Histidine tags –6-10 histidines. GST – Glutathione affinity chromatography. MBP – Maltose affinity chromatography. Protein A -lgG affinity chromatography. Calmodulin binding peptide (CBP, 4kDa)–Calmodulin affinity chromatography. TAP – Combination protein A and calmodulin binding peptide. IMPACT –Chitin binding domain. FLAG - Synthetic FLAG peptide. Although the tags hardly ever interfere with the biological activity of the protein, if the protein is to be expressed for structural study purposes, it may be necessary to remove the tag after the purification procedures. Removal of the tag from a protein of interest can be accomplished with a site-specific protease, and cleavage should not reduce protein activity.47 It is important to be sure that the protein of interest lacks the recognition site of the protease selected. The most commonly used proteases are: enterokinase, tobacco etch virus (TEV), factor Xa, and thrombin. Enterokinase is often the choice for N-terminal fusions, since it specifically recognizes a five-amino-acid polypeptide (D-
electrophoresis being the most used ones. Mass spec-
in this purpose, with mass spectrometry and 2-D
with drugs. Several techniques are available to aid
correlate it with, for example, possible interactions
understanding of its structure, and being able to

assay is chosen, one must be aware that it should be
expressed in native conditions (after all, we did choose
heterologous expression system, right?), one must
be sure that the protein being expressed is biologi-
active and the tests are very different from one
protein to another. The protein assays are quantita-
tive methods to determine the amount of a given ac-
tivity. Some of them are: enzymatic reactions, when
one is able to quantify a specific product after a
reaction involving the protein of interest, the use of
a radioactive-labeled substrate that changes its phy-
properties after interacting with our protein, mea-
suring the ability to stimulate cell death, cell
proliferation or other effect on development
in vivo
or the retention in a DNA column, suggesting that
the protein interacts with nucleic acids. Whichever
assay is chosen, one must be aware that it should be
experimentally convenient, meaning that it is rea-
sonably sensitive and easy to perform, and the as-
say employed must be specific for the activity of in-

FUNCTIONAL STUDIES

Since the protein of our interest is not being ex-
pressed in native conditions (after all, we did choose
heterologous expression system, right?), one must
be sure that the protein being expressed is biologi-
ly active and the tests are very different from one
protein to another. The protein assays are quantita-
tive methods to determine the amount of a given ac-
tivity. Some of them are: enzymatic reactions, when
one is able to quantify a specific product after a
reaction involving the protein of interest, the use of
a radioactive-labeled substrate that changes its phy-
properties after interacting with our protein, mea-
suring the ability to stimulate cell death, cell
proliferation or other effect on development in vivo
or the retention in a DNA column, suggesting that
the protein interacts with nucleic acids. Whichever
assay is chosen, one must be aware that it should be
experimentally convenient, meaning that it is rea-
sonably sensitive and easy to perform, and the as-
say employed must be specific for the activity of in-

PERSPECTIVES

Being that proteins are the main target for drug
discovery, an important effort is dedicated to under-
standing their function. The field of proteomics has
advanced considerably and its common applications
are: target identification and validation (identifying
proteins whose expression levels or activities change
in disease states), identification of biomarkers from
biological fluids (used to assess whether target mod-
ulation has occurred), and research of mechanisms
of drug action or toxicity (screen compounds in pre-
clinical studies for target organ toxicities as well as
later on in development during clinical trials). But
the practice of proteomics ranges from the identifi-
cation of thousands of proteins in a particular model
system, to the detailed analysis of their 3D structu-
re, possible modifications/isoforms, and the function
of a single protein. Hence producing proteins in
high amounts and purity can aid in the study of
function and structure of novel proteins or help dis-
cover un-known properties of long-known ones that
ultimately will be fundamental in the development of
highly specific drugs.
ACKNOWLEDGMENTS

The authors thank Drs. William Dubinsky and Illarian V. Turko for their critical reading of the manuscript.

REFERENCES

2. The Proteome Society www.proteome.org
11. Novagen www.novagen.com
12. Stratagene www.stratagene.com

28. Invitrogen www.invitrogen.com
34. Roche www.roche-applied-science.com

Reimpresos:

Dr. Hugo A. Barrera Saldaña
Facultad de Medicina, Departamento de Bioquímica
Universidad Autónoma de Nuevo León
Av. Madero y Dr. E. Aguirre Pequeño s/n
Col. Mitras Centro, 64460 Monterrey, N.L.
Tel.: (81) 8329 41 73 y 74. Fax (81) 8333 77 47
Correo electrónico: hbarrera@fm.uanl.mx