Chapter 4
Biotechnology in Medicine: Advances-II

Sudeepa Rajan, Aadil Hussain Bhat, Sudipa Maity, and Omika Thakur

Abstract Therapeutically important proteins have been obtained from relevant organisms using biotechnological methods. The main caveat in procuring these proteins in such a way is their insufficient quantity in the natural sources. This difficulty has been circumvented by the recombinant protein technology, which entails the abundant expression and purification of the protein of interest in a heterologous system. Certain drawbacks in the *E. coli* (bacterial) expression system, the most widely used and economical expression system, have prompted the development of alternative expression systems like insect and mammalian ones for therapeutic and diagnostic products including human insulin, growth hormones, and antibody fragments. Major expression systems are described in this chapter along with a focus on a variety of protein purification tags. Brief information on expressed sequence tags is also provided. Later, the role of bioinformatics in handling huge amounts of genomics and proteomics data together with an application of various tools for protein structure analysis is explained. Finally, the significance of different protein detection arrays in biomarker discovery and diagnostics is reviewed followed by methods for data interpretation and analysis.

Keywords Protein expression system · Protein purification · Purification tags · Protein detection array · Protein domains · Homology modeling

S. Rajan
Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA, USA

A. H. Bhat (✉)
UCLA School of Dentistry, Los Angeles, CA, USA

S. Maity · O. Thakur
Indian Institute of Technology Roorkee, Haridwar, Uttarakhand, India

M. Anwar et al. (eds.), *Fundamentals and Advances in Medical Biotechnology*,
https://doi.org/10.1007/978-3-030-98554-7_4

4.1 Protein Expression: Introduction to *E. coli* Expression System, Yeast Expression System, Insect Expression System. Higher-Eukaryotic Expression Systems

4.1.1 Protein Expression

The main objective of recombinant protein expression is to produce a target protein in commercial quantity in a cost-effective manner without compromising its biochemical and biological activity. Several protein expression systems have been used to produce large-scale proteins for therapeutic (biopharmaceuticals) and diagnostic purpose. Expression and purification of full-length protein is sometimes unnecessary, as the desired activity or property of the protein of interest can be achieved by a specific domain(s). Once the desired protein or domain(s) is finalized, next critical step is choosing the suitable expression system.

An expression system is defined as genetic constructs (expression vector with the desired gene sequence) specifically designed to produce protein of interest at high level, inside a host cell. The fundamental criteria for expression system selection include anticipated application of desired protein; resources availability; expenditure cost; and time. The four most well established expression systems in use for pharmaceutical purposes are bacteria, yeast, baculovirus, and mammalian, and will be discussed in details.

4.1.1.1 Bacterial Expression System

Bacterial expression system is most popular and first choice for rapid and economical production of pharmaceutically important recombinant proteins [1] *Escherichia coli* has emerged as most commonly used industrial microorganism, as about one-third of all the pharmaceutical relevant proteins are purified from it [2]. Other bacterial strains used for pharmaceutical purpose are *Lactococcus lactis* and *Bacillus subtilis*. The major advantages of using this system are: (1) Simple and scalable, with no sophisticated equipment requirement, (2) Quickest, (doubling time of *E. coli* is 20 min), (3) Easily manipulated due to availability of its genetics knowledge, and (4) Economical, as one can produce tones of recombinant protein in no time [3, 4].

An expression vector is required to insert the target gene in the bacterial cells. Numerous articles are available online that extensively discuss the desired properties of an expression vector. Inexpensive usage, a little or no leaky expression with a reliable and tunable induction promoter are some of the desirable qualities of an expression vector for large-scale protein production [5, 6]. Commonly used and commercially available vectors are the pET series, pQE series (Qiagen), pGEX, etc. for single protein expression and pACYC, pBAD, and pSC101 series or single plasmid system like the Duet vectors (Novagen) for more than one proteins expression at the same time.

Despite being the most efficient and cost-effective expression system, it has several shortcomings especially when expressing mammalian proteins. These problems include stability of mRNA and protein inside the host, codon bias [7], inclusion bodies formation [8], and protein folding and solubility [9], and absence of key post-translational modifications (glycosylation, carboxylation, and amidation). Increasing demand of the pharmaceutically important proteins in the last 5 years has led to the development of new codon optimized and genetically modified bacterial strains capable of performing desired post-translational modification, and improved expression vectors to overcome above mentioned problems [10–12]. A list of the commercially available modified strains with their key features is given in Table 4.1. The use of other strategies like growing cells at lower temperature to prevent protein aggregation 15–23 ºC [15, 16], co-expression of molecular chaperones or the post-translational modifying enzyme [17–19], and translocating protein to periplasmic space has helped in increasing protein yield [20–22]. The protein expression in periplasmic space emerged as most successful strategy to produce several pharmaceutical proteins such as scFc, growth hormone, etc. owing to several advantages over others [23]. Protein isolation from inclusion bodies is also a viable and cost-effective option for pharmaceutical companies, as the purity of the protein in IBs is ~95% [24–26]. It is often better to change expression system to higher hierarchy as they are better equipped in folding and complex post-translational modifications to maintain the bioactivity of desired protein.

4.1.1.2 Yeast Expression System

Yeast being a eukaryotic organism and acts as a connecting link between *E. coli* and mammalian systems. It has properties similar to both the systems thus serves as an excellent expression system for pharmaceutically relevant protein production. The major advantages it has over other eukaryotic expression systems are: (1) Ability to grow in high densities in limited time, (2) Simple and cheap media requirements, (3) Easy genetic manipulation, and (4) Safe pathogen-free production [27]. Unlike bacterial expression system, yeast expression system has the ability to accomplish proper posttranslational modifications and extracellular expressions [28, 29]. *Saccharomyces cerevisiae*, *Pichia pastoris*, *Kluyveromyces lactis*, and *Schizosaccharomyces pombe* [30] are few commonly used yeasts. *S. cerevisiae* and *P. pastoris* are used extensively for the production of several therapeutic important recombinant proteins such as human insulin, human serum albumin, alpha 2b, trypsin, collagen, hepatitis B vaccines, and human papillomavirus (HPV) vaccines [31, 32]. The expression vector is designed so that genomic integration of expression cassette could take place leading to generation of stable expression clones with multiple copies of target gene [33]. *P. pastoris* based vectors also have inducible promoters (P_{AOX1} (most widely used), GAP, FLD1, PEX8, and YPT1) and antibiotic selection markers, for the selection of multi-copy transformants [34]. Some commercially available plasmids have these features incorporated (such as the pYEDIS, pPIC9K, pPICZα vector).

Table 4.1 List of modified bacterial strains [10]

Strain	Key feature	Application	Company
Promoter inducible strains			
BL21(DE3)	• IPTG-inducible T7 RNAP(DE3) promoter	General protein expression	Multiple companies
	• Protease deficiency		
BL21 star (DE3)	• IPTG-inducible T7 RNAP(DE3) promoter	General protein expression from low-copy plasmid	Invitrogen™ (ThermoFisher)
	• Mutation in RNaseE gene, resulting in longer mRNA half-life		
BL21(DE3) pLysS	• Contains pLysS plasmid which express T7 lysozyme to suppress leaky expression	Toxic protein expression	Multiple companies
BL21-AI	• T7RNAP gene under control of the araBAD promoter	Toxic protein expression	Invitrogen™ (ThermoFisher)
	• Tight regulation of protein expression		
BLR (DE3)	• RecA-deficient, improves plasmid monomer yield	Expression of unstable proteins that might cause loss of DE prophage	Novagen (Merck)
	• Stabilizes plasmids with repetitive sequences		
Tuner(DE3) and derivatives	• Mutation in lac permease (*lacY*) allows uniform entry of IPTG	Expression of difficult protein: membrane proteins, toxic proteins, and proteins prone to insoluble expression	Novagen (Merck)
	• High regulation of IPTG induced protein expression		
Lemo21(DE3)	• Modulated levels of T7 lysozyme (inhibitor of T7RNAP) by L-rhamnose addition	Expression of difficult protein: membrane proteins, toxic proteins, and proteins prone to insoluble expression	NEB
	• Tunable expression of protein		
RiboTite	• Integration of orthogonal riboswitches upstream of the T7RNAP gene thereby permitting fine tuning of protein expression	For secretion of recombinant proteins in periplasm	Dixon laboratory
OverExpress™ C41(DE3) and C43(DE3)	• Derived from standard BL 21 (DE) strains	Expression of membrane toxic proteins from all classes of organism (yeast, plant, virus, and mammals)	Lucigen and Sigma [13]
	• Contain genetic mutation in *t7rnap*, lowering the T7 RNAP accumulation		
	• It phenotypically selected for conferring toxicity tolerance.		

(continued)

Table 4.1 (continued)

Strain	Key feature	Application	Company
Marionette	• Protein coexpression independent control of expression using 12 different inducers	Protein complexes	Addgene
(KRX) Competent Cells	• Stringent control provided by the rhamnose-driven T7 RNA polymerase	Expression of proteins	Promega
Codon biased strains			
BL21 (DE3) CodonPlus-RIL/ RP	• Contains pRI(P)L plasmid provides extra copies of rare tRNAs genes • Codon bias correction	Enable efficient high-level expression of heterologous proteins in *E. coli*	Stratagene
Rosetta or Rosetta (DE3)	• Contains pRARE plasmid provides extra copies of rare tRNAs genes • Good for "Universal" translation	Enable efficient high-level expression of heterologous proteins in *E. coli*	Novagen(Merck)
Others			
Origami and derivatives	• Has mutation in trxB and gor reductase resulting in more oxidant conditions in the cytoplasm	Disulfide-bonded protein production	Novagen (Merck)
SHuffle® T7	• Contains the deletions of the genes for glutaredoxin reductase and thioredoxin reductase (Agor ΔtrxB) • Constitutively expresses a chromosomal copy of the disulfide bond isomerase DsbC, which promotes the correction of mis-oxidized proteins • DsbC is also a chaperone that can assist in the folding of proteins that do not require disulfide bonds	Expression potentially toxic protein	NEB
ArcticExpress (DE3)	• Production of aggregation-prone proteins/constitutive expression of chaperones Cpn10 and Cpn60 from the psychrophilic bacterium Oleispira antarctica, which show high refolding activities at 4–12 °C		Agilent
TatExpress BL21	• Strong inducible promoter, *ptac* upstream of *tatABCD* operon for increased levels of Tat secretion pathway	Enhanced production of industrially challenging proteins	Robinson laboratory [14]

In recent times, *P. pastoris* being a obligate aerobic yeast has gained more prestige as it can use methanol as a carbon source which lead to the development of an expression system based on the utilization of the inducible AOX1 promoter [35]. Other benefits include correct protein folding and secretion (by Kex2, PHO1, or α-MF as signal peptidase) outside the cell [36].

Despite being an efficient expression system, yeast has drawbacks like limited secretory expression of heterologous proteins resulting in low protein yield and limits its commercial use [37]. To address this issue, strategies involving optimization of cultivation parameters (induction temperature and time, pH, oxygen, and nutrient supply), protein-based host selection, gene copy number, co-expression of secretory proteins such as chaperones, engineering of secretory pathways, have been employed to improve the expression of proteins [38, 39]. The other major drawback is hyper *N*- and *O*-linked glycosylation of proteins unlike mammalian system, which affects the protein immunogenicity [40]. To overcome this shortcoming humanization of yeast by expressing mammalian specific glucose transferase and omitting yeast specific genes involved in glycosylation has been tried and found its application in producing humanized IgGs in yeast [41–45]. Latest technologies such as CRISPR/Cas9 [46] and GlycoSwitch [47] are now used for yeast genome engineering for this purpose. Many improvements are still required in the yeast expression system before it can be used for therapeutic commercial purposes; therefore, companies are shifting their focus to other more complex expression systems.

4.1.1.3 Insect Cell Expression System

Baculovirus-mediated insect cell expression systems are widely used these days to produce large quantities of proteins for pharmaceutical purpose. These proteins are difficult to express either in bacteria or yeast due to improper protein folding or posttranslational modifications. The major advantage of this system remains the presence of post-translational modifications similar to mammalian system, thus avoiding the problem of immune-reactivity [48]. The other advantages include: (1) Cheap protein production cost as compared to mammalian system, (2) High capacity of expressing multiple genes at the same time due to large and flexible viral genome (130 kb), (3) Safe, as they do not infect humans, (4) High protein yield driven by the strong promoters such as polyhedrin or p10, and (5) Easy downstream purification [49, 50].

Baculovirus vectors are used to insert desired gene and transfected into cultured insect cells. The most commonly used baculovirus system are Bac-to-Bac (Invitrogen), BacPAK (Takara), and BaculoGold (BD Biosciences), they are commercially available and has been widely used [51]. Recently, two rapid and simple baculovirus expression vector systems have been developed named as MultiBac [52] and Golden Gate-based system [53], which can be used to express multiple genes at the same time. The most common insect cells used for the production of protein are Sf9, Sf21 (*Spodoptera fugiperda),* High5 (*Trichopulsia ni*), and S2 (*Drosophila melanogaster* embryos) cell lines [54]. These cell lines have proven themselves for the application

of recombinant protein expression from a variety of expression platforms due to their ability to grow in suspension and serum-free medium [55]. Recent advancement in genetic engineering of baculovirus system, this system has been promoted from being used during pest control to production of several recombinant proteins (viral antigens) by biopharmaceutical companies for human use [56]. Several vaccines are now available for the commercial use, which include Cervarix™ (GSK, Rixensart, Belgium), FluBlok™ (Protein Sciences Corp., Connecticut, USA), Provenge (DendreonInc., Seattle, WA), and Chimigen (Virexx Medical Corp., Calgary, Canada), [56–58]. Several of the known subunit vaccines against Chandipura virus, hepatitis E virus, and West Nile virus are also synthesized in insect cells [59]. Since, baculovirus are known not to infect humans, they are being evaluated to be used as efficient delivery vehicles for gene and cell therapy [60].

Despite being a promising expression system, baculovirus has several drawbacks. For instance, time consuming cloning procedure to generate stable recombinant virus, expensive media requirements, and cell lysis by baculovirus infections resulting in suboptimal protein processing. The major drawback is differential glycosylation pattern in comparison to humans, thus limiting the therapeutics use of recombinant protein [50]. In recent years, efforts are directed to deal with these problems. The stable transformation of insect cell line with plasmid having early baculovirus constitutively active promoters (e.g., IE1) [61] or using pre-infected cells [baculovirus infected insect cells (BIIC)] to avoid making of stable system has significantly reduce the processing time along with improved protein production [62]. Engineering of insect cells to have mammalian glycosyltransferases enzymes is also tested [63]. Regardless of recent advancements there is a room for improvement to produce cost-effective, industrial scale, and therapeutics standard recombinant protein in insect cell lines [64]. Several good articles are available which covers the baculovirus system in more details [65–67].

4.1.1.4 Mammalian Expression System

Mammalian expression system is the ideal choice for production of therapeutically important proteins because they perform similar post-translational modifications along with proper protein folding, which are critical for bioactivity of the protein. Other advantages include secretion of proteins in the cell culture, preventing additional step of protein purification [68]. Some mammalian cell lines can grow in suspension culture and serum-free chemically defined media, enabling large-scale reproducible protein production [69]. Several different mammalian host cell lines, such as Chinese Hamster Ovarian (CHO) cells, baby hamster kidney (BHK21) cells, and *murine myeloma*-Sp2/0/NS0 cells, have been used for large-scale production of therapeutic proteins [70]. Recently, the use of human cell line (Human Embryonic Kidney (HEK 293) cells and fibro-sarcoma HT-1080 cells) has gained importance due to identical post-translational modifications of expressed proteins [71].

Like any other expression systems, plasmid based or viral based vectors (adenoviral vector, vaccinia vector, Semliki forest viral vectors) are used to transfect desired gene into cultured mammalian cells to form either transient or stable cell lines. However, efficient integration of transgenes into correct genomic loci still remains a major challenge during stable cell line production. Incorporation of chromosomal elements (nuclear scaffold/matrix attachment regions (S/MARs) and ubiquitous chromatin opening elements (UCOEs)) into plasmid vectors is found to have a positive effect on stable gene expression. Transposon based vectors and site-specific recombinase systems, such as Cre-Lox and Flp-FRT, are also found useful in targeted integration of the gene in host genome for the production of stable cell line expressing recombinant proteins [72].

In the year 1968, the production of FDA approved first recombinant glycoprotein, tissue plasminogen activator (tPA, Activase) in CHO made a revolution [69]. Subsequently, a number of vaccines like Herpes simplex virus (HSV) vaccine [73], Synagis® vaccine used against respiratory syncytial virus [74] etc. and several therapeutic proteins, Drotrecogin alfa (XIGRIS®; Eli Lilly Corporation, Indianapolis, IN), recombinant factor IX Fc and VIII Fc fusion protein (Biogen, Cambridge, MA), dulaglutide (TRULICITY®; Eli Lily, Indianapolis, IN) etc. [68] are produced in mammalian cell line. Some of them have already received FDA approval.

The major drawback of using mammalian expression system is high cost of protein production, due slow cell growth, expensive media, and culture conditions (continuous CO_2 supply, expensive transfection reagents). In recent years, mammalian cells have been further developed for the commercial production of broader range therapeutic proteins by selecting high protein-producing stable cell clones using methotrexate (MTX) amplification or glutamine synthetase (GS) system technology and high-throughput fluorescence-activated cell sorting (FACS)-based screening method [75–77]. Other advancements include genetic modification of mammalian cells by over-expression of anti-apoptotic proteins (bcl-2 family members and Bcl-x(L)) [78] or by inducing cell cycle arrest by adding anti-mitotic agents (such as hydroxyurea, nocodazole, colchicine, paclitaxel or vinblastine) [77] to increase cell viability along with high cell density which eventually lead to elevated protein productivity.

4.2 Protein Purification: Principle of Heterologous Protein Purification following Expression. Use of His Tag, GST-Tag, MBP-Tag, TAP-Tag, Myc-Tag

4.2.1 Protein Purification

Procuring pure and biologically active desired protein after expression is a daunting task. Separating desired protein from the rest of cellular protein pool is an essential prerequisite step for commercial production of therapeutic proteins. Several different chromatographic techniques like size exclusion, ion exchange, hydrophobic

interaction, affinity, and ammonium sulfate cut-off method, are widely used to isolate desired protein from other cellular impurity. These techniques rely on exploiting protein properties like, electric charge, solubility, size and hydrophobicity, therefore, optimization of the specific purification method is a time consuming and cumbersome task. Among the above mentioned techniques, affinity chromatography is the most time and cost-effective, and usually a single step purification method [79].

In affinity chromatography, protein is expressed in fusion to an affinity tag which significantly help in protein purification. Some of the known affinity tags also increase protein solubility without affecting biological activity of the protein, an additional advantage [80]. Polyhistidine tag (his-tag) is the most commonly used affinity tag. The purification is based on IMAC (immobilized metal-ion affinity chromatography), where adsorption of protein occurs due to coordination between an immobilized metal-ion (Ni^{2+} or Cu^{2+} ion) and an electron donor groups from the protein surface (stretch of 6–10 histidine (tryptophan and cysteine)). The protein is purified using imidazole gradient [81]. The other commonly used tags are polyarginine-tag, FLAG-tag, c-Myc-, S-, and Strep-tag; they all are around the same ~10 amino acid. Due to their small size, fusing these tags either at N- or C-terminus of desired protein usually does not affect its structure or biological activity. FLAG-tag is a hydrophilic octapeptide (DYKDDDDK) recognized by the M1 mAb resin. Due to non-reusability of antibody resin, the effective cost is high thereby restricting its widespread use. Recently, the development of anti-FLAG molecularly imprinted polymers (MIPs) approach using tetrapeptide DYKD as template has provides a cost-effective alternative solution for purifying FLAG-derived recombinant proteins [82]. Strep-tag (WRHPQFGG) binds to biotin and therefore, recombinant protein can eluted using biotin as competitor in buffer. Use of desthiobiotin facilitates regeneration and repeated use of these resins. Recent development of Strep-tag®II and Strep-Tactin has enhanced the use of Strep-tag owing to their higher affinity for the biotin [83]. Next comes the large molecular weight affinity tags (more than 200 amino acids), glutathione-*S*-transferase (GST), maltose-binding protein (MBP), N-utilization substance protein A (NusA), thioredoxin (Trx), ubiquitin, and SUMO [84]. Some of these affinity tags not only aid in purification but also increase the solubility of protein. MBP and GST act as both solubility enhancer and affinity tag, while NusA, SUMO, and Trx only increase solubility [85, 86]. MBP is a periplasmic *E. coli* protein with high solubility. MBP when fused with desired protein increases fused protein solubility due to its intrinsic chaperone activity [87]. NusA is a 55 kDa, elongation factor which regulates transcription in *E. coli.* As a fusion partner it improves the solubility of the protein due to its intrinsic high solubility [88]. SUMO (small ubiquitin-related modifier) protein is a reversible post-translational modification at ε-NH_2-group of lysine residues of target protein. It is known to increase the solubility and expression of the protein. SUMO-specific proteases removes the SUMO tag from the target proteins thereby reducing the erroneous cleavage within the target protein [80]. Thioredoxin (Trx) is a small and highly soluble *E. coli* protein. Like NusA tag, Trx itself does not act as an affinity tag and thus, requires fusion partners during purification step. List of affinity tags are given in Table 4.2. However, now-a-days larger soluble tags are being replaced

Table 4.2 Affinity and solubility tags for recombinant proteins [modified from refs [89, 90]

Tag	Size (kDa)	Sequence	Characteristics	
			Binding	Elution
His-tag	0.84	H_6 or H_{10}	Metal-ion based adsorption	Imidazole 20–250 mM or low pH
Arg-tag	0.80	RRRRR	Cation exchange resin	NaCl linear gradient from 0 to 400 mM at alkaline pH > 8.0
FLAG	1.01	DYKDDDDK	Use anti-FLAG monoclonal antibody for purification	pH 3.0 or 2–5 mM EDTA
Step-tag II	1.06	WSHPQFEK	Strep-Tactin (modified streptavidin)	2.5 mM desthiobiotin
HA-tag	0.99	YPYDVPDYA	HA-tag-specific antibody	High concentration of the HA-tag peptide or by low pH buffer
c-myc	1.20	EQKLISEEDL	Monoclonal antibody	Low pH
T7 tag	1.21	MASMTGGQQMG	Monoclonal antibody	Low pH
S-tag	1.75	KETAAAKFERQHMDS	S-fragment of RNaseA	3 M guanidine thiocyanate, 0.2 M citrate pH 2, 3 M magnesium chloride
Calmodulin binding domains	2.96	KRRWKKNFIAVSAANRFKKISSSGAL	Calmodulin	EGTA or EGTA with 1 M NaCl

Tag	Size (kDa)	Sequence	Characteristics	
			Binding	Elution
Chitin binding domain	5.99	TNPGVSAWQVNTAYTAGQLVTYNGKTYKCLQPHTSLAGWEPSNVPALWQLQ	Chitin	Fused with intein: 30–50 mM dithiothreitol, b-mercaptoethanol, or cysteine
Thioredoxin (Trx)	12	Protein	Does not have intrinsic affinity properties (require other fusion tag)	Depends other affinity tag
Glutathione *S*-	26.0	Protein	Glutathione	5–10 mM reduced

by small soluble tags like SET tag [91] and Fh8 tag [87] to overcome the problem faced by size of large tags.

Tandem affinity purification (TAP) has also gained popularity in recent times. In TAP, desired protein is fused with at least two different affinity tags [92] or one affinity and another solubility enhancer fusion tag [93], depending on need. This helps in purifying desired protein using both the tags sequentially resulting in considerable reduction of nonspecific proteins. Overall it helps in increasing the purity of the desired protein making it useful for therapeutic purpose [80]. Numerous combinations of solubility-enhancing and affinity tags have been exploited in order to enhance both protein solubility and yield of the desired protein. Some of the commercially available TAPs are S3S-TAP-tag (is a recently developed system suitable for purification of mammalian protein complexes), FF-ZZ TAP-tag, Strep/FLAG-TAP (SF-TAP), GS-tag, PTP-tag, etc. [80].

Even though affinity tags are routinely used in laboratory protein purifications, they have a limited use in commercially therapeutic applications. Several times these tags cause structural and activity changes or immunogenicity problems. Sequence specific protease or chemical cleavage methods are developed for the removal of these tags. The tobacco etch virus (TEV) protease, thrombin, Enterokinase, and factor Xa are some of the most commonly used protease to remove tags [89]. Most of the commercially available vectors have a protease cleavage site designed between fusion tags and desired proteins. The solubility of a desired protein after tag removal cannot be predicted and enzymatic cleavage might cause negative effects, such as product heterogeneity due to cleavage at multiple sites, precipitation or poor recovery [94]. Chemical cleavage with CNBr-based method has advantages over enzymatic cleavage, as it is easy to remove from the reaction mixture and is cheap. Their use is largely restricted due to their harsh reactive nature and unwanted protein modifications making purified protein unsuitable for therapeutic use [85]. The use of the poly-ionic peptide tags (addition of 3–5 charge amino acid sequence) has shown to enhance solubility of the desired protein, regardless of their position at N- or C-terminus of the protein. The tags enhance the solubility of protein by increasing repulsive electrostatic interactions between protein molecules due to additional of charge from the tags. Due to their small size, the presence of poly-ionic peptide tags do not affect the structure or biological activity of the protein, an add-on advantage [94].

4.3 Proteomics: Introduction, Protein Detection Array, Protein Informatics, Domain Analysis, and Structure Prediction

Proteomics is defined as the study of proteome or a set of proteins found in a cell, tissue, or a whole organism. The importance of proteomics lies in the fact that unlike genome, the proteome is not constant and changes from cell to cell over time,

making an individual unique or different. The proteome provides a snapshot of the cell in action and the proteomics aims at understanding the proteome status at a large scale under certain physiological or diseased conditions [95]. The term "protein" was initially introduced in 1938 by the Swedish chemist JönsJakob Berzelius, working in electrochemistry while trying to describe a class of macromolecules made up of linear chains of amino acids [96]. Although proteomics research began in 1975 with the introduction of 2-Dimensional Gel Electrophoresis by O'Farrell and Klose, it was not until the early 1990's, when the term "proteomics" was coined by Mark Wilkins, a Ph.D. student at the Macquarie University, Australia [97].

Proteomics is a rapidly growing field with cutting-edge technologies used to investigate expression of proteins, post-translational modifications, and involvement of proteins in metabolic pathways and protein interactomics. The most commonly applied are mass spectrometry (MS)-based techniques such as Tandem-MS and gel-based techniques such as differential in-gel electrophoresis (DIGE). Another technique complementing to MS is protein microarray that has been widely applied as a promising proteomic technology with great potential for protein expression profiling, biomarker screen, drug discovery, drug target identification, and analysis of signaling pathways in health and disease [98]. The recording and analyses of the enormous amount of data generated by these high-throughput technologies are facilitated by the development of databases and online servers that are critical not only for recording and storing this data but also enable structure, function, and domain prediction of a protein [99]. For example, four major databases—UniProtKB, IntAct, Reactome, and PRIDE are responsible for storing all the up-to-date information generated for a protein [100–102]. In addition to that, several prediction software and servers such as Phyre2, FoldX, BisKit, etc. have facilitated protein structure prediction [103–105].

4.3.1 Protein Detection Array

Protein array analysis is a technique by which proteins spotted in defined locations on a solid support (a protein microarray, or protein chip) are probed for interactions with a probe molecule in a high-throughput, parallel manner [106–108]. Protein array analysis is used to screen protein function, drug discovery, biomarker discovery, expression profiling, and antibody analysis [109, 110]. Typically, a protein microarray is prepared by immobilizing proteins onto a microscope slide using a standard contact spotter or noncontact microarrayer [111]. The microscopic slide surface can be made of aldehyde and epoxy-derivatized glass that get attached to amines and nitrocellulose or the surface could be nickel-coated that relies on more specific affinity attachment of His6-tagged proteins which results in the generation of ten-fold better signals. After proteins are immobilized on the slides, they can be probed for a variety of functions/activities [112, 113]. Finally, the resulting signals are usually measured by detecting fluorescent or radioisotope labels. Protein microarrays are majorly categorized into two classes: analytical and functional [114,

115]. In addition, tissue or cell lysates can also be fractionated and spotted on a slide to form a reverse-phase protein microarray [116].

4.3.1.1 Analytical Microarrays

Analytical microarray is majorly represented by the antibody array which primarily employs the "analyte-labeled" assay format where an array is queried with: (1) a probe (labeled antibody or lig-and) or (2) an unknown biologic sample (e.g., cell lysate or serum sample) containing analytes of interest [117]. By tagging the query, molecules with a signal-generating moiety, a pattern of positive and negative spots is generated. For each spot, the intensity of the signal is proportional to the quantity of applied query molecules bound to the bait molecules. An image of the spot pattern is captured, analyzed, and interpreted [118, 119]. This format, successfully, found alterations in protein expression in cancer cell development, epithelial and stromal cells. However, one of the major limitations of the antibody array approach is the production of specific antibodies in a high-throughput manner. In addition, targeted protein labeling may lead to epitope destruction because of some chemical reactions [120].

This assay can also be explained as the original enzyme-linked immunosorbent assay (ELISA) in a multiplexed format, but can only detect dozens to hundreds of analytes simultaneously because cross-reactivity between antibodies can occur [120]. Recombinant antibodies have become a promising means of overcoming this problem; however, their fabrication issues such as cloning and protein expression add to complexities to their practical use [121]. To improvise on the sensitivity and specificity, analytical microarrays usually employ "sandwich" assay format [106]. This format employs two different antibodies to detect the targeted protein (1) the capture antibody that immobilizes the targeted protein on the solid phase and (2) the reporter or detection antibody that generates a signal for the detection system. This format was applied to successfully detect 75 cytokines with high specificity, femtomolar sensitivity, a 3-log quantitative range, and economy of sample consumption [106, 122, 123].

4.3.1.2 Functional Microarray

Functional protein microarrays are constructed using individually purified proteins that enable the study of various biochemical properties of proteins, such as binding activities, including protein–protein, protein–DNA, protein–lipid, protein–drug, and protein–peptide interactions, and enzyme–substrate relationships via various types of biochemical reactions [106, 120]. Functional protein microarrays are constructed by printing a large number of individually purified proteins, and in principle, it is feasible to print arrays comprised of virtually all annotated proteins of a given organism, effectively comprising a whole proteome microarray [124]. Functional protein microarrays have been successfully applied to identify

protein–protein, protein–lipid, protein–antibody, protein–small molecules, protein–DNA, protein–RNA, lectin–glycan, and lectin–cell interactions, and to identify substrates or enzymes in phosphorylation, ubiquitylation, acetylation, and nitrosylation, as well as to profile immune response [114]. The first use of functional protein microarrays was demonstrated by Zhu et al. [125] to determine the substrate specificity of protein kinases in yeast. Since then, reported applications of functional protein microarrays in basic research, as well as in clinical applications, have increased rapidly [126]. Significant achievements in providing the whole proteome of several organisms (i.e., human, yeast, *E. coli*, virus) on arrays have provided the tools for many important biological discoveries [126].

4.3.1.3 Reverse-Phase Protein Microarrays

This format immobilizes an individual complex test sample in each array spot such that an array is comprised of hundreds of different patient samples or cellular lysates. Each array is incubated with one detection protein (i.e. antibody), and a single analyte end point is measured and directly compared across multiple samples. This method allows for the analysis of many samples obtained at different states by directly spotting tissue, cell lysates, or even fractionated cell lysates on a glass slide. Many different probes can be tested to specifically identify certain proteins in lysate samples [120]. This type of microarray was first established by Paweletz and colleagues to monitor histological changes in prostate cancer patients [127]. Using this method, they successfully detected microscopic transition stages of pro-survival checkpoint protein in three different stages of prostate cancer: normal prostate epithelium, prostate intraepithelial neoplasia, and invasive prostate cancer. The high degree of sensitivity, precision, and linearity achieved by reverse-phase protein microarrays enabled this method to quantify the phosphorylation status of some proteins (such as Akts and ERKs) in these samples; phosphorylation was statistically correlated with prostate cancer progression.

4.3.2 Protein Informatics

With the advent of high-throughput technologies like Next-Generation Sequencing (NGS), a wealth of information about genomic sequences from a variety of organisms has been amassed. This has led to a rapid buildup of protein sequence data in the form of new protein databases and updation of the existing ones. The exponential increase in the protein related data has prompted computational biologists to develop an advanced infrastructure that facilitates better organization, structural and functional annotation, and evolutionary analyses [128]. Along with a myriad of protein analysis tools, numerous protein related databases have been created that can be categorized as sequence databases, family and domain databases, 3D structure databases, gene expression databases, enzyme and pathway databases, PTM databases,

protein-protein interaction databases, etc. [129]. More information can be accessed at http://www.oxfordjournals.org/our_journals/nar/database/cap/, https://proteinformationresource.org/staff/chenc/MiMB/dbSummary2015.html or Expasy, a Swiss Bioinformatics Resource Portal.

4.3.3 Domain Analysis

Protein domains are defined as basic units of structure, function, and evolution. Ranging from 30 to 600 amino acids in length, these units are able to fold independently into stable tertiary structures [130–132]. Compact structures with separate hydrophobic cores represent structural domains wherein contacts between residues within the domain are found to be more extensive than between domains [133, 134]. Identification of domains forms the basis of protein classification and annotation. This is exemplified by many protein sequence databases like Pfam, SMART and Interpro, and protein structure databases such as SCOP, CATH, and PALI, which consider domains as the basis for their classification of proteins. The domain-level approach has strongly influenced our understanding in the areas of evolutionary history, homology detection and modeling, protein fold recognition, etc. Interestingly, a relatively random domain shuffling process is thought to have led to domain linkages during the course of evolution, resulting in a few beneficial domain associations being selected and propagated in the interest of cell fitness [135–138].

Most sequence-based domain recognition methods rely on the conservation of contiguous homologous segments, which is complicated by domain shuffling and recombination in multidomain proteins, accessory domains, and evolutionary divergence of sequences. Therefore, to get an insight into the functional and structural interplay of domains in multidomain proteins, all domains in the full-length amino acid sequence need to be considered simultaneously for protein classification. Keeping this in view, an alignment-free tool, named CLAP (CLAssification of Proteins), was developed for effective classification of multidomain proteins, bypassing the need for identification of domains and their sequential order [135].

Proteins are generally composed of one or more domains arranged in a distinct way that largely dictates the protein function. This is referred to as domain architecture [139, 140].A limited fraction of domain combinations have been found in proteins ruling out the possibility of random combinations. In accordance with power law distribution, the covalent linkage between domains is such that most domains have few partners with a smaller fraction of abundant domains being highly connected [141–143]. Based on domain co-occurrence or context, a novel approach, dPUC (Domain Prediction Using Context) was developed for domain prediction and identification. The scores are assigned by analyzing whether two domain families frequently co-occur (positive context) or have never been found as a pair (negative context) [144].

As we move from prokaryotes to eukaryotes or from unicellular eukaryotes to animals, the number of unique domains and the fraction of multidomain proteins

increase. This organismal complexity-associated trend, called domain accretion, is thought to play a significant role in evolution. While researchers have likened the genomes to natural language texts, protein domains are considered analogous to words, with domain architectures and amino acids representing sentences and letters, respectively [145]. So *n*-gram analysis, a well-known probabilistic language-modeling (linguistic) technique helpful in the identification of meaningful word combinations by treating consecutive words in sentences as a unit, is utilized to probe the rules of domain association leading to distinct domain architectures. The set of rules, termed "proteome grammar" is employed to study genome complexity and domain evolution [146]. Some domains tend to be involved in many different domain architectures, a phenomenon called protein domain promiscuity. A bigram analysis has been employed to study the evolution of this promiscuity. Bigram refers to a pair of domains on a protein sequence [139, 147]. Domain rearrangements and domain accretion are two important aspects of evolution. It has been found that there is a nearly universal value of information gain (loss of entropy) associated with a transition to the observed domain architectures from random domain combinations. This highly conserved constant value corresponds to the minimum complexity required to maintain a functioning cell and is governed by "quasi-universal grammar." However, two major groups viz. a subset (extremely simplified cells) of Archaea and animals (extreme complexity) have deviations from this constant value [146].

4.3.3.1 Domain Parsing

The accurate prediction of domain boundaries (domain parsing) is crucial for the design of chimeric proteins with multi-functional domains and the experimental structure determination of proteins where crystallization is adversely affected by flexible regions. It also makes the multiple sequence alignments more reliable [148, 149]. The defining feature of structural domains underlies some effective algorithms that assign domain boundaries using 3D structures. Some protein features such as signal peptide, trans-membrane helices, low-complexity, and disordered regions and coiled coils, which are not found in globular domains, can be easily predicted using relevant tools. These analyses performed subsequent to a sequence search with BLAST constitute initial steps in domain prediction. The domain boundary prediction employing templates with known structure involves three steps [150]. Sequence search against protein structure databases like PDB comprises the first step and helps in retrieving alignments between target sequence and template structure. The more accurate alignments with a percentage identity of at least 30% and enough coverage (at least 100 residues) with few gaps can be used for 3D model generation. Many methods including a highly sensitive HHPred server have been developed for detecting remote structural homologs by performing a search against a wide range of databases such as PDB, SCOP, Pfam, and COGs [151]. The final step involves 3D model generation using a modeling program like MODELLER (incorporated in HHSearch server) [152–154]. Phyre2 [105, 155], I-TASSER [156,

157], and ROBETTA [158] are some of the methods that detect templates and develop the model automatically.

4.3.4 Structure Prediction

A few advanced techniques, viz. X-ray crystallography, nuclear magnetic resonance (NMR), and recently developed cryo-electron microscopy (cryo-EM), have been instrumental in solving the 3D structures of proteins. In spite of this, a rapid non-proportional progress in the field of genomics has widened the already existing gap between the number of protein sequences identified and the number of available protein 3D structures [159]. To address this issue, a variety of computational methods that are faster, easier, and economical have been developed. One of the methods involves sampling the conformational space (c-space) of a protein through deterministic or heuristic approaches. With deterministic methods like homology modeling, entire or part of the c-space is scanned and sub-spaces excluded based on a priori knowledge. In heuristic algorithms (ab initio modeling, Monte Carlo, and molecular dynamics simulations), only a fraction of the c-space is sampled without a priori knowledge generating a representative set of Boltzmann-weighted conformations [160, 161].

Homology modeling (also called comparative modeling) consists of predicting 3D structure from the primary sequence of the protein. It is useful in identifying therapeutic targets, studying structure and function of proteins, protein interaction networks and signaling pathways, and mutagenesis associated with certain diseases [159, 162]. It also has applications in molecular modeling of protein complexes and in the refinement of cryo-EM 3D structures [163–165]. It involves multiple steps starting with the identification and selection of suitable templates by searching PDB (Protein Data Bank), an online database of known crystal structures. The protein Basic Local Alignment Search Tool (BLASTp) is employed to look for templates with a sequence identity of more than 40%. There are other algorithms available including PSI-BLAST (Position-Specific Iterated BLAST), hidden Markov models (HMMs), and profile–profile alignments, for templates with low homology [166]. Subsequent to the optimization of the selected alignments, 3D model is built using rigid-body assembly method (as in 3D-JIGSAW and SWISS-MODEL programs), segmented matching method (used by SegMod/ENCAD), spatial restraint method (used by MODELLER and DRAGON), or the artificial evolution method (used by NEST). Next, loop modeling is performed either by scanning a structure database like PDB (a knowledge-based approach used by MODELLER, 3D-JIGSAW and SWISS-MODEL) or by optimizing a scoring function through Monte Carlo or molecular dynamics methods for randomly chosen conformations (an energy-based method using an ab initio fold prediction approach). Addition of side chains to the main backbone requires rotamer libraries, which contain statistical distributions of side chain and backbone orientations extracted from known crystal structures. These are tested sequentially and scored using energy functions. Some of the tools

used for side chain packing include SCWRL, FASPR, and SCAP [160–162]. To improve the quality of the model thus generated, optimization is done by energy minimization through molecular mechanics force fields. Other ways of model refinement employ molecular dynamics and Monte Carlo simulations. The relationship between the energy of a protein and its conformation is described by the potential energy hyper-surface (PEHS) modeled with quantum or molecular mechanical methods. The native conformation of a protein in the funnel shaped PEHS is ideally represented by a global energy minimum conformation (GEMC), although a canonical ensemble of structures is required to describe the system state completely. The top portion of the PEHS funnel contains high energy conformations resulting from steric and hydrophilic/hydrophobic clashes and unoptimized bond lengths and angles, etc. These conformations are eliminated as a protein folds and GEMC is reached with the narrowing of the funnel [160, 167].

Finally, the model is evaluated and validated by considering stereochemistry, physical parameters, statistical mechanics, etc. To perform this task, Distance-matrix ALIgnment (DALI; http://ekhidna2.biocenter.helsinki.fi/dali/) or Verify3D online servers are used. Many homology modeling programs and online servers like SWISS-MODEL [168–170] and Phyre2 [105] have been developed that perform most of the aforementioned steps in an automated fashion. Other homology modeling programs include MODELLER [153], I-TASSER [157], Rosetta [171], Raptor X [172, 173], GalaxyTBM [174], AlphaFold [175], etc.

4.4 Expression Sequence Tags (ESTs), Application of Protein Detection Microarray with Examples

4.4.1 Expression Sequence Tags (ESTs)

Expressed sequence tags (ESTs) are partial cDNA sequences, resulting from single pass sequencing of clones obtained from cDNA libraries. They are used for decoding genome organization and determining gene expression profiles in specific tissues under different conditions. The foremost utilization of ESTs in genome organization studies is to regulate the chromosomal localization of analogous genes employing somatic hybrid cell panel. Furthermore, ESTs contribute in comparative genetics of different species to decipher their gene function. Overall the ESTs lead to integrated genomic approach by the combination of sequence, functional, and localization data. To date, over 45 million ESTs have been generated from over 1400 different eukaryotic species. They have been proven very useful in gene identification and predictions because they are low-cost alternative to whole genome sequencing [176]. This is particularly important for eukaryotes which tend to have "less gene-dense genomes" [177].

For the generation and processing of ESTs mRNA is collected either from whole organisms or specific tissues depending on the size of the organism. This is followed

by the extraction of pooled mRNAs and purified typically on the basis of their polyadenylation. Subsequently, a cDNA library is constructed from this pool and clones are randomly picked for a single pass sequencing read. The raw data are then processed to derive the underlying sequence which is followed by further processing that removes low-quality and contaminating sequences associated with the vectors. The purified sequence data is the submitted to EST database such as dbEST [178–180]. The continued generation of ESTs for different species along with the technological advancement has led to its exploitation in range of applications. For example, tandem mass spectrometry matches peptide fragments to known protein sequences. However, limited number of sequences in protein databases leads to computational bias against poorly characterized proteins. ESTs are beginning to have a widespread appeal in identifying and characterizing alternative spliced isoforms [181].

4.4.2 Application of Protein Detection Microarray

Microarray technology was developed in 1989 by Roger Ekins, based on ambient analyte immunoassay [182, 183]. Later, it was transformed into DNA microarray for simultaneous detection of mRNA expression levels in multiple genes. However, the mRNA expression in a cell does not always correspond to exact protein levels [184]. Therefore, to overcome these limitations of DNA microarrays, protein microarray was developed for functional analysis of proteins as they are the major driving forces behind all cellular processes. Immunoassays are first protein microarray, based on specific antigen-antibody interactions, later expanded to antibody microarray which enabled parallel detection of multiple proteins in minute sample quantity with high sensitivity and reproducibility [184]. High-throughput protein array was developed by immobilization of purified proteins on chip glass slide/bead/nitrocellulose membrane or microtiter plate chip [106].

As discussed earlier, the arrays are divided into three main categories; (1) analytical protein microarrays (2) reverse phase protein microarrays, and (3) functional protein microarrays (Fig. 4.1). Analytical protein microarrays (APMs) or capture microarrays are composed of antibodies, aptamers, or affibody libraries attached to a solid support that binds to a specific protein in cell lysate. The APMs provide information regarding protein expression, their binding affinities and specificity; however, cross-hybridization of antibodies is still the major challenge associated with these microarrays. Analytical microarrays are generally used for identification and profiling of treated/non-treated cells and diseased/non-diseased tissues. The reverse phase protein microarray (RPPA) separates complex mixture of proteins in tissue lysate; detected by fluorescent or chemiluminescent assays, and are useful to identify altered proteins or post-translational modifications in diseased cell. Unlike APMs and RPPMs, functional protein microarray (FPM) is used to study biochemical activities in entire proteome. These microarrays are composed of full-length purified functional protein or protein domain arrays immobilized on protein chip. FPMs are used to identify protein–protein, protein–DNA, protein–RNA,

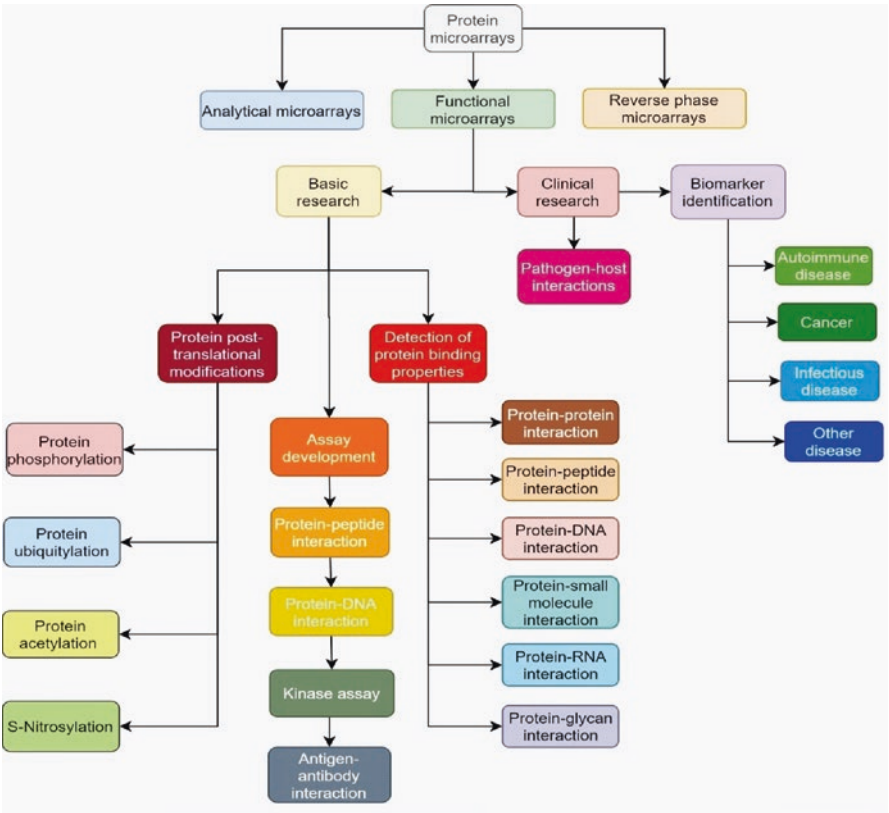

Fig. 4.1 Types of protein microarrays and their application in basic and clinical research

protein–phospholipids, and protein–small molecule interactions, detect antibodies, and determine enzyme activity and its specificity. Compared to other methods, FPMs are more capable in detecting low level of protein expression and weak interactions. On the basis of recent developments in FPMs, they are divided into four categories, namely, (1) Purified proteome microarray, (2) Purified protein family microarray, (3) Purified protein domain microarray and, (4) Cell-free protein/peptide microarray.

The purified proteome microarray, consisting of genome-wide expressed proteins immobilized on a microarray, is widely utilized in *E. coli*, *S. cerevisiae,* and human system to study their functional and biochemical properties. Two DNA repair proteins, namely YbaZ and YbcN, were identified by Chen et al. using *E. coli* proteome microarray consisting of 4256 unique proteins [185]. In another study, using similar microarray, Spr phase switch and DNA binding proteins were identified in type 1 fimbria [186]. Unique antimicrobial peptide and glycosaminoglycans protein targets were identified using *E. coli* proteome microarray [187, 188]. Additionally, *CobB* deacetylation enzyme was identified as a strong binder of cyclic di-GMP (bacterial second messenger) while Yojl was found to be involved in

bacterial cell invasion by probing human brain microvascular endothelial cells on *E. coli* proteome microarray [189, 190]. Besides this, *E. coli* proteome microarray has also been applied for identification of glycoproteins [191], tyrosine sulfation [192], and ClpYQ protease [193] to study bacterial physiology and host–microbial interactions.

A total of 33 novel calmodulin and more than 150 phospholipid binding proteins were identified with biotinylated calmodulin and fluorescently labeled liposomes, respectively, utilizing yeast proteome microarray [125]. The same microarray was used for the identification of SMIR3 and SMIR4 rapamycin inhibitors, Arg5,6 mitochondrial enzyme, and Pus4 and App1 brome mosaic virus antiviral proteins [194–196]. Lin et al. further demonstrated two signaling pathways, NuA4 complex-mediated protein acetylation reactions involved in yeast aging and substrates for HECT domain ubiquitin E3 ligase Rsp5 [197]. All these studies demonstrate the usefulness of bacterial and yeast proteome microarray in basic research. However, human proteome microarray is still the most widely used in clinical research, pharmaceutical industry, and translational research. HuProt composed of ~21,000 full-length purified human proteins, ProtoArray ~9000 purified human proteins from insect cells, and NAPPA with 10,000 human proteins are the three popular human proteome microarrays. Human proteome microarray is broadly applied in five major areas: (a) diagnostics, (b) proteomics, (c) protein functional analysis, (d) antibody characterization, and (e) treatment development. Diagnostics includes profiling of sera to discover new disease biomarkers and monitoring of disease states/responses to therapy in personalized medicine. In 2010, Song et al. identified and validated three highly specific anti-hepatitis biomarkers RPS20, Alba-like, and dUTPase with 47.5%, 45.5%, and 22.7% sensitivity, respectively, using human protein microarray consisting of 5011 non-redundant proteins [198]. In another study, a microarray with 1626 purified human recombinant proteins was utilized for validation of six highly specific biomarkers against autoimmune hepatitis with 82% sensitivity and 92% specificity [199]. Also, six highly specific biomarkers, namely PTPRN2, MLH1, MTIF3, PPIL2, NUP50, QRFPR associated with type 1 diabetes were recently validated using Nucleic Acid Programmable Protein Array (NAPPA) [200]. In addition to this, Protoarray was used for validation of transglutaminase 4 (TGM4) biomarker specific infertility causing autoimmune polyendocrine syndrome type 1 (APS1) in males [201]. Furthermore, validation of RNA Polymerase II subunit A C-terminal domain phosphatase (CTDP1) biomarker specific to Behcet disease was performed using HuProt array [202]. Three highly specific p53, PTPRA, and PTGFR biomarkers were validated against ovarian cancer using NAPPA 5177 tumor antigens microarray with 98.3% specificity [203]. In other study, four SNX1, PQBP1, IGHG1, and EYA1 biomarkers specific to glioma were identified by probing ~17,000 human protein microarray [204]. HuProt array was used for identification of COPS2, NT5E, TERF1, and CTSF biomarkers for diagnosis of gastric cancer and validation of p53, HRas, and ETHE1 for early detection of lung cancer [205, 206]. Moreover, three specific FGFR2, CALM1, and COL6A1 prostate cancer biomarkers were identified and validated using 123 purified antigens microarray platform [207]. Human proteome array has been utilized to validate IGHG4, STAT6,

CRYM, HDAC7A, EFCAB2, SELENBP1, and CCNB1 biomarkers against Meningiomas [208]. In a more recent study, a highly specific biomarker panel, identified and validated using protein array based approach, was employed to discriminate Zikavirus and Dengue virus infections [209]. High-Density Nucleic Acid Programmable Protein Array (HD-NAPPA) of the pathogen *Mycobacterium tuberculosis* (*Mtb*) was used in the identification of eight antibody targets, viz. Rv0054, Rv0831c, Rv2031c, Rv0222, Rv0948c, Rv2853, Rv3405c, Rv3544c, for tuberculosis serology [210].

Currently, there are two methods to detect protein signals (1) labeled and (2) label free. The ideal protein array detection method should produce low background noise and generate high signal frequency. Therefore, the most common and widely used detection method is fluorescence labeling which is highly sensitive, safe, and compatible with readily available microarray laser scanners. Other labels used are affinity, photochemical, or radioisotope tags. As these labels are attached to the probe itself which can interfere in the probe-target protein reaction; thus, a number of label free detection methods are developed such as surface Plasmon resonance (SPR), carbon nanotubes, carbon nanowire sensors, and microelectromechanical system cantilevers. Most of these methods are relatively new and not very suitable for detection of high-throughput protein interactions; however, they do offer much promise for the future.

4.5 Data Analysis and Interpretation of Protein Detection Arrays

Protein microarrays provide wealth of information regarding protein interactions, protein functions, and signaling pathways which could be used for clinical diagnosis. However, data translation requires automated data processing and interpretation for generation of meaningful information. Protein microarray data analysis mainly depends on the design of surfaces, content, detection method, data preprocessing, inference, classification, and validation (Fig. 4.2). The design of array is a crucial step as it significantly affects data analysis and its final interpretation. Inclusion of biological replicates is recommended as they provide higher statistical confidence; however, they also make results more complex to evaluate. Data preprocessing, which includes image analysis, normalization, and data transformation, also greatly affects data analysis and interpretation. Image-processing algorithms distinguish foreground and background intensities and inference based on data analysis variability [211]. Different data analysis strategies for different types of array generate variable results. These arrays provide variety of tools for disease analysis but lack standard analytical and data processing strategy which enhance complexity in data analysis.

The data analysis strategies like spot-finding on slide images, *Z*-score calculations, and significance analysis of microarrays (SAM) have their origin in DNA microarray analysis; however, concentration-dependent analysis (CDA) is specific for protein microarrays (Fig. 4.3) [212].

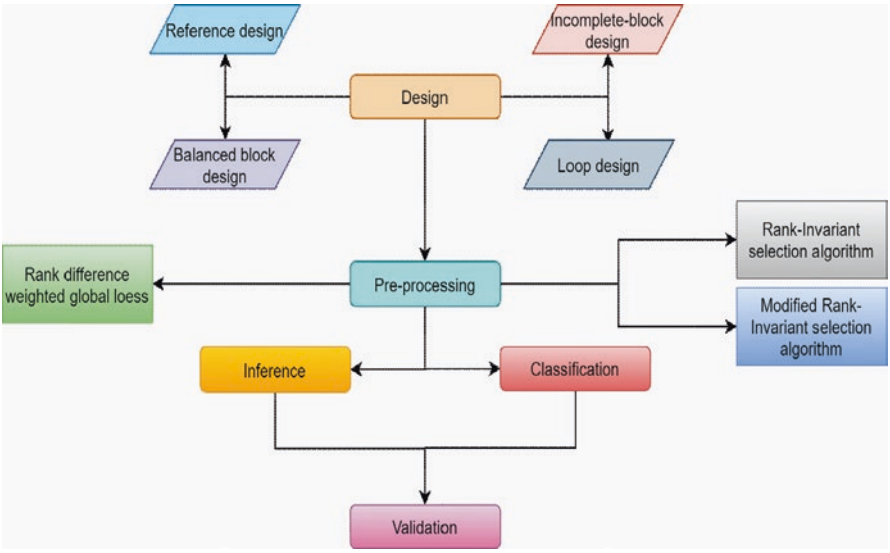

Fig. 4.2 Protein microarray strategy for data analysis

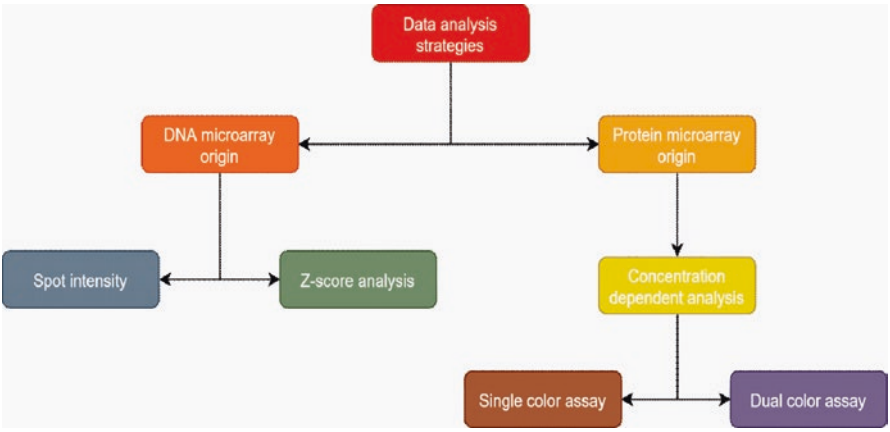

Fig. 4.3 Methods for data analysis of protein microarrays

In spot intensity determination method, microarray image analysis starts with the fixing of spot intensity. A grid of circles with adjusted position and size are placed over protein spots to get reliable intensity data. The output file is created by GenePix Pro software (Molecular Devices, CA). In *Z*-score analysis, *Z*-score equation ($Z_s = S_s - \mu/\sigma$) is analyzed to determine the significantly different from the expected values, where Z_s is the *Z*-score for the *s*th spot, S_s is the signal for that spot', μ is the mean signal across all spots, and σ is the standard deviation across all spots. In concentration-dependent analysis (CDA), absolute signals generated depend on protein concentration in the sample. To overcome this issue an iteration process is

used to calculate *Z*-score ($Z_s = S_s - \mu_w/\sigma_w$) and remove outliers. In the equation, Z_s represents *Z*-score for the *s*th spot, S_s spot signal, μ_w mean signal, and σ_w standard deviation (Fig. 4.3).

The signals produced are detected by fluorescent dyes or colorimetric assays. There are two types of assays for colorimetric detection of proteins: single color assay and dual color assay. Single color assay is a single antibody based microarray which uses internal control system based on two colors for quantification of antigen and antibody. In the dual color assay, each sample is labeled with different fluorescent dyes and their signal intensity is measured using fluorescence image scanners. Dual color assays have better reproducibility and discriminatory efficiency then single color assays [213]. To prevent undesired technical artifacts caused by electric charges, different protein sizes, hydrophobic protein interaction of proteins and antibody/antigen binding kinetics in dual color assays the data pre-processing protocols like filtering, background correction and data normalization are required [214]. Furthermore, it employs four different microarray designs: (a) Reference in which sample of interest and reference sample is labeled with different fluorescent fluorochromes. This design is generally used for comparative studies; (b) Balanced-block where two samples bearing two different fluorochromes are hybridized to make a single block; (c) Incomplete block, more than two samples bearing only two fluorescent fluorochromes are co-hybridized on microarray; (d) Loop design where samples are hybridized in different arrays using different fluorochromes which leads to duplication of arrays.

For data normalization, different algorithms, rank-invariant selection, modified rank-invariant selection and rank difference weighted global loess are used to define the set of probes. Rank-invariant selection algorithm is used in the absence of housekeeping controls. However, its major limitation is that it does not cover entire intensity range [215]. Modified Rank-Invariant Selection Algorithm corrects intensity values through extrapolation of curve to lower and upper intensity limits. Rank Difference Weighted Global Loess is applied to whole probes on array to get global normalization. Despite the differences among data processing methods in microarray analysis the general recommendations which need to be considered for data processing in microarray analysis are: Bayesian approaches to examine intersections, quality-control, validation methods, and standardized testing platforms.

4.6 Summary

Medical biotechnology has provided several products essential for research, therapeutics, and diagnostic purpose. Recombinant protein technology is the connecting link between medical biotechnology, and mass production of therapeutic and diagnostic products. For a long time, *E. coli* has served as the cost-effective and low-maintenance expression system for the production of pharmaceutically important recombinant proteins. The limitation of expressing several human proteins with specific post-translational modifications, essential for their biological activity, in

E. coli (which lacks post-translational modification machinery), has forced biochemists to look for alternative expression systems for large recombinant protein production. Using new and improved recombinant methods, humanization of *E. coli,* which has failed so far, could be tried and made successful. Lately, insect cell lines, non-human mammalian cell lines, and human cell lines with engineered genome are extensively used to produce therapeutic proteins such as growth factors, vaccines against infectious diseases, monoclonal antibodies, and IFNs to treat cancers and other diseases. The advantages and disadvantages of the different expression systems are reviewed in this chapter. We have also discussed the significance of protein tags used during protein purification following its expression. Some of these tags are also known to increase the solubility of the proteins, ultimately leading to high protein yields required for commercial purpose.

Ever since informative machineries started to evolve, proteomics technologies have been aimed at the comprehensive detection of the downstream proteins to evaluate complex disease diagnosis, allied mechanism and concerned therapy for effective management of the diseases. Moreover, to understand the complex biological organization, it is imperative to understand regulatory interconnections between DNA, RNA, and protein. For instance, microarrays, automated sequencing, and mass spectrometry have significantly contributed to systems biology approach by investigating protein–protein interactions, signal pathway analysis, studies of PTMs, or/and detection of toxins. It also has a wide array of opportunities in disease biomarker discovery to enable better disease management through improved diagnostics. More importantly, various forms of protein microarray have gradually evolved for proteomics research. With the progressive development, standardization of the experimental workflow and data interpretation, protein microarray holds promises in diagnostic applications. For protein microarray data analysis, various techniques including spot intensity determination method, *z*-score calculation, and concentration-dependent analysis have been used. Moreover, colorimetric assays involving fluorescent dyes are used for detecting signals.

Expressed sequence tags and cDNA provide direct evidence for all sample transcripts and are the most important resources for transcriptome exploration. ESTs have proven useful in different applications along with individual tools and pipelines for EST analysis.

Finally, an increasing number of bioinformatics methods have been developed to meet the needs of researchers for rigorous analysis of a vast amount of data generated through high-throughput genomic and proteomic techniques. From sequence-based analysis to protein structure prediction, analysis tools have been developed that focus on individual steps of the process or perform the whole process in an automated way. For instance, different protein databases along with certain analysis tools have been used for protein data analysis. Correct identification of domain boundaries for elucidating domain architecture and predicting protein structure from primary sequences using homology modeling have become possible with some of the finest tools developed recently. All these methods have facilitated the research on therapeutic agents that could be used in drug designing and other areas.

References

1. Swartz JR (2001) Advances in Escherichia coli production of therapeutic proteins. Curr Opin Biotechnol 12:195–201
2. Walsh G (2014) Biopharmaceutical benchmarks 2014. Nat Biotechnol 32(10):992–1000
3. Fakruddin M, Mohammad Mazumdar R, Bin Mannan KS, Chowdhury A, Hossain MN (2013) Critical factors affecting the success of cloning, expression, and mass production of enzymes by recombinant E. coli. Biotechnology 2013:1–7
4. Huang CJ, Lin H, Yang X (2012) Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements. J Ind Microbiol Biotechnol 39:383–399
5. Studier FW (2005) Protein production by auto-induction in high density shaking cultures. Protein Expr Purif 41(1):207–234
6. Horga LG, Halliwell S, Castiñeiras TS, Wyre C, Matos CFRO, Yovcheva DS et al (2018) Tuning recombinant protein expression to match secretion capacity. Microb Cell Factories 17(1):199
7. Edwards RA, Bryan J (1995) Fascins, a family of actin bundling proteins. Cell Motil Cytoskeleton 32(1):1–9
8. Ramón A, Señorale-Pose M, Marín M (2014) Inclusion bodies: not that bad…. Front Microbiol 5:298
9. Baneyx F, Mujacic M (2004) Recombinant protein folding and misfolding in Escherichia coli. Nat Biotechnol 22:1399–1407
10. Rosano GL, Morales ES, Ceccarelli EA (2019) New tools for recombinant protein production in Escherichia coli: a 5-year update. Protein Sci 28:1412–1422
11. Lipinszki Z, Vernyik V, Farago N, Sari T, Puskas LG, Blattner FR et al (2018) Enhancing the translational capacity of E. coli by resolving the codon bias. ACS Synth Biol 7(11):2656–2664
12. Feldman MF, Wacker M, Hernandez M, Hitchen PG, Marolda CL, Kowarik M et al (2005) Engineering N-linked protein glycosylation with diverse O antigen lipopolysaccharide structures in Escherichia coli. Proc Natl Acad Sci U S A 102(8):3016–3021
13. Dumon-Seignovert L, Cariot G, Vuillard L (2004) The toxicity of recombinant proteins in Escherichia coli: a comparison of over expression in BL21(DE3) C41(DE3) and C43(DE3). Protein Expr Purif 37(1):203–206. https://doi.org/10.1016/j.pep.2004.04.025
14. Browning DF, Richards KL, Peswani AR, Roobol J, Busby SJW, Robinson C (2017) Biotechnol Bioeng 114(12):2828–2836. https://doi.org/10.1002/bit.26434
15. San-Miguel T, Pérez-Bermúdez P, Gavidia I (2013) Production of soluble eukaryotic recombinant proteins in E. coli is favoured in early log-phase cultures induced at low temperature. Springerplus 2(1):1–4
16. Bjerga GEK, Lale R, Williamson AK (2016) Engineering low-temperature expression systems for heterologous production of cold-adapted enzymes. Bioengineered 7(1):33–38
17. de Marco A, Deuerling E, Mogk A, Tomoyasu T, Bukau B (2007) Chaperone-based procedure to increase yields of soluble recombinant proteins produced in E. coli. BMC Biotechnol 7:32
18. Kyratsous CA, Panagiotidis CA (2012) Heat-shock protein fusion vectors for improved expression of soluble recombinant proteins in Escherichia coli. Methods Mol Biol 824:109–129
19. Amann T, Schmieder V, Faustrup Kildegaard H, Borth N, Andersen MR (2019) Genetic engineering approaches to improve posttranslational modification of biopharmaceuticals in different production platforms. Biotechnol Bioeng 116:2778–2796
20. Georgiou G, Segatori L (2005) Preparative expression of secreted proteins in bacteria: status report and future prospects. Curr Opin Biotechnol 16:538–545
21. Choi JH, Lee SY (2004) Secretory and extracellular production of recombinant proteins using Escherichia coli. Appl Microbiol Biotechnol 64:625–635
22. Landeta C, Boyd D, Beckwith J (2018) Disulfide bond formation in prokaryotes. Nat Microbiol 3(3):270–280

23. Karyolaimos A, Ampah-Korsah H, Hillenaar T, Mestre Borras A, Dolata KM, Sievers S et al (2019) Enhancing recombinant protein yields in the E. coli periplasm by combining signal peptide and production rate screening. Front Microbiol 10:1511
24. Burgess RR (2009) Chapter 17 Refolding solubilized inclusion body proteins. In: Methods in enzymology. Academic Press, Boston, pp 259–282
25. Simpson RJ (2010) Solubilization of Escherichia coli recombinant proteins from inclusion bodies. Cold Spring Harb Protoc 5(9):pdb.prot5485
26. Ban B, Sharma M, Shetty J (2020) Optimization of methods for the production and refolding of biologically active disulfide bond-rich antibody fragments in microbial hosts. Antibodies 9(3):39
27. Baghban R, Farajnia S, Rajabibazl M, Ghasemi Y, Mafi AA, Hoseinpoor R et al (2019) Yeast expression systems: overview and recent advances. Mol Biotechnol 61:365–384
28. Kim HJ, Kim H-J (2017) Yeast as an expression system for producing virus-like particles: what factors do we need to consider? Lett Appl Microbiol 64(2):111–123
29. Huertas MJ, Michán C (2019) Paving the way for the production of secretory proteins by yeast cell factories. Microb Biotechnol 12:1095–1096
30. Çelik E, Çalik P (2012) Production of recombinant proteins by yeast cells. Biotechnol Adv 30:1108–1118
31. Hou J, Tyo KEJ, Liu Z, Petranovic D, Nielsen J (2012) Metabolic engineering of recombinant protein secretion by Saccharomyces cerevisiae. FEMS Yeast Res 12:491–510
32. Vieira Gomes A, Souza Carmo T, Silva Carvalho L, Mendonça Bahia F, Parachin N (2018) Comparison of yeasts as hosts for recombinant protein production. Microorganisms 6(2):38
33. Daly R, Hearn MTW (2005) Expression of heterologous proteins in Pichia pastoris: a useful experimental tool in protein engineenring and production. J Mol Recognit 18:119–138
34. Damasceno LM, Huang CJ, Batt CA (2012) Protein secretion in Pichia pastoris and advances in protein production. Appl Microbiol Biotechnol 93:31–39
35. Krainer FW, Dietzsch C, Hajek T, Herwig C, Spadiut O, Glieder A (2012) Recombinant protein expression in Pichia pastoris strains with an engineered methanol utilization pathway. Microb Cell Factories 11(1):22
36. Karbalaei M, Rezaee SA, Farsiani H (2020) Pichia pastoris: a highly successful expression system for optimal synthesis of heterologous proteins. J Cell Physiol 235:5867–5881
37. Idiris A, Tohda H, Kumagai H, Takegawa K (2010) Engineering of protein secretion in yeast: strategies and impact on protein production. Appl Microbiol Biotechnol 86:403–417
38. Looser V, Bruhlmann B, Bumbak F, Stenger C, Costa M, Camattari A et al (2014) Cultivation strategies to enhance productivity of Pichia pastoris: a review. Biotechnol Adv 33:1177–1193
39. Yang Z, Zhang Z (2018) Engineering strategies for enhanced production of protein and bioproducts in Pichia pastoris: a review. Biotechnol Adv 36:182–195
40. Han M, Yu X (2015) Enhanced expression of heterologous proteins in yeast cells via the modification of N-glycosylation sites. Bioengineered 6(2):115–118
41. Hamilton SR, Davidson RC, Sethuraman N, Nett JH, Jiang Y, Rios S et al (2006) Humanization of yeast to produce complex terminally sialylated glycoproteins. Science 313(5792):1441–1443
42. Gerngross TU (2004) Advances in the production of human therapeutic proteins in yeasts and filamentous fungi. Nat Biotechnol 22:1409–1414
43. Li H, Sethuraman N, Stadheim TA, Zha D, Prinz B, Ballew N et al (2006) Optimization of humanized IgGs in glycoengineered Pichia pastoris. Nat Biotechnol 24(2):210–215
44. Choi BK, Bobrowicz P, Davidson RC, Hamilton SR, Kung DH, Li H et al (2003) Use of combinatorial genetic libraries to humanize N-linked glycosylation in the yeast Pichia pastoris. Proc Natl Acad Sci U S A 100(9):5022–5027
45. Bobrowicz P, Davidson RC, Li H, Potgieter TI, Nett JH, Hamilton SR et al (2004) Engineering of an artificial glycosylation pathway blocked in core oligosaccharide assembly in the yeast Pichia pastoris: production of complex humanized glycoproteins with terminal galactose. Glycobiology 14(9):757–766

46. Baghban R, Farajnia S, Ghasemi Y, Mortazavi M, Zarghami N, Samadi N (2018) New developments in Pichia pastoris expression system, review and update. Curr Pharm Biotechnol 19(6):451–467
47. Jacobs PP, Geysens S, Vervecken W, Contreras R, Callewaert N (2009) Engineering complex-type N-glycosylation in Pichia pastoris using GlycoSwitch technology. Nat Protoc 4(1):58–70
48. Jarvis DL (2009) Chapter 14 Baculovirus-insect cell expression systems. In: Methods in enzymology. Academic Press, Boston, pp 191–222
49. Adeniyi AA, Lua LHL (2020) Protein expression in the Baculovirus-insect cell expression system. In: Methods in molecular biology. Humana Press, Totowa, pp 17–37
50. Hu YC (2005) Baculovirus as a highly efficient expression vector in insect and mammalian cells. Acta Pharmacol Sin 26:405–416
51. Han X, Huang Y, Hou Y, Dang H, Li R (2020) Recombinant expression and functional analysis of antimicrobial Siganus oramin L-amino acid oxidase using the Bac-to-Bac baculovirus expression system. Fish Shellfish Immunol 98:962–970
52. Berger I, Garzoni F, Chaillet M, Haffke M, Gupta K, Aubert A (2013) The MultiBac protein complex production platform at the EMBL. J Vis Exp 77:e50159
53. Neuhold J, Radakovics K, Lehner A, Weissmann F, Garcia MQ, Romero MC et al (2020) GoldenBac: a simple, highly efficient, and widely applicable system for construction of multi-gene expression vectors for use with the baculovirus expression vector system. BMC Biotechnol 20(1):26
54. Contreras-Gómez A, Sánchez-Mirón A, García-Camacho F, Molina-Grima E, Chisti Y (2014) Protein production using the baculovirus-insect cell expression system. Biotechnol Prog 30(1):1–18
55. Felberbaum RS (2015) The baculovirus expression vector system: a commercial manufacturing platform for viral vaccines and gene therapy vectors. Biotechnol J 10:702–714
56. Aucoin MG, Mena JA, Kamen AA (2010) Bioprocessing of Baculovirus vectors: a review. Curr Gene Ther 10(3):174–186
57. Van Oers MM, Pijlman GP, Vlak JM (2015) Thirty years of baculovirus-insect cell protein expression: from dark horse to mainstream technology. J Gen Virolo 96:6–23
58. Cox MMJ (2012) Recombinant protein vaccines produced in insect cells. Vaccine 30:1759–1766
59. Bill RM (2015) Recombinant protein subunit vaccine synthesis in microbes: a role for yeast? J Pharm Pharmacol 67(3):319–328
60. Airenne KJ, Hu YC, Kost TA, Smith RH, Kotin RM, Ono C et al (2013) Baculovirus: an insect-derived vector for diverse gene transfer applications. Mol Ther 21:739–749
61. Harrison RL, Jarvis DL (2016) Transforming lepidopteran insect cells for continuous recombinant protein expression. In: Methods in molecular biology. Humana Press, Totowa, pp 329–348
62. Chen S (2016) Alternative strategies for expressing multicomponent protein complexes in insect cells. In: Methods in molecular biology. Humana Press, pp 317–326
63. Shi X, Jarvis D (2007) Protein N-glycosylation in the Baculovirus-insect cell system. Curr Drug Targets 8:1116–1125
64. Chavez-Pena C, Kamen AA (2018) RNA interference technology to improve the Baculovirus-insect cell expression system. Biotechnol Adv 36:443–451
65. Jarvis DL (2014) Recombinant protein expression in baculovirus-infected insect cells. In: Methods in enzymology. Academic Press, Boston, pp 149–163
66. Kost TA, Kemp CW (2016) Fundamentals of baculovirus expression and applications. In: Advances in experimental medicine and biology. Springer, New York, pp 187–197
67. Fabre LL, Arrías PN, Masson T, Pidre ML, Romanowski V (2019) Baculovirus-derived vectors for immunization and therapeutic applications. In: Emerging and reemerging viral pathogens, vol 2: Applied virology approaches related to human, animal and environmental pathogens. Elsevier, Amsterdam, pp 197–224

68. Dumont J, Euwart D, Mei B, Estes S, Kshirsagar R (2016) Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives. Crit Rev Biotechnol 36(6):1110–1122
69. Kim JY, Kim YG, Lee GM (2012) CHO cells in biotechnology for production of recombinant proteins: current state and further potential. Appl Microbiol Biotechnol 93:917–930
70. Swiech K, Picanço-Castro V, Covas DT (2012) Human cells: new platform for recombinant therapeutic protein production. Protein Exp Purif 84:147–153
71. Ghaderi D, Taylor RE, Padler-Karavani V, Diaz S, Varki A (2010) Implications of the presence of N-glycolylneuraminic acid in recombinant therapeutic glycoproteins. Nat Biotechnol 28(8):863–867
72. Zhu J, Hatton D (2018) New mammalian expression systems. In: Advances in biochemical engineering/biotechnology. Springer, Berlin, pp 9–50
73. Stanberry LR, Spruance SL, Cunningham AL, Bernstein DI, Mindel A, Sacks S et al (2002) Glycoprotein-D–adjuvant vaccine to prevent genital herpes. N Engl J Med 347(21):1652–1661
74. Null D, Pollara B, Dennehy PH, Steichen J, Sánchez PJ, Givner LB, et al. Safety and immunogenicity of palivizumab (synagis) administered for two seasons. Pediatr Infect Dis J 2005;24(\):1021–1023
75. Lai T, Yang Y, Ng SK (2013) Advances in mammalian cell line development technologies for recombinant protein production, vol. 6, Pharmaceuticals. MDPI AG, Basel, pp 579–603
76. Lin PC, Chan KF, Kiess IA, Tan J, Shahreel W, Wong SY et al (2019) Attenuated glutamine synthetase as a selection marker in CHO cells to efficiently isolate highly productive stable cells for the production of antibodies and other biologics. MAbs 11(5):965–976
77. Bandaranayake AD, Almo SC (2014) Recent advances in mammalian protein production. FEBS Lett 588:253–260
78. Cost GJ, Freyvert Y, Vafiadis A, Santiago Y, Miller JC, Rebar E et al (2010) BAK and BAX deletion using zinc-finger nucleases yields apoptosis-resistant CHO cells. Biotechnol Bioeng 105(2):330–340
79. Gräslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O et al (2008) Protein production and purification. Nat Methods 5:135–146
80. Walls D, Loughran ST (2011) Tagging recombinant proteins to enhance solubility and aid purification. Methods Mol Biol 681:151–175
81. Gutiérrez R, Martín Del Valle EM, Galán MA (2007) Immobilized metal-ion affinity chromatography: status and trends. Sep Purif Rev 36(1):71–111
82. Gómez-Arribas LN, Urraca JL, Benito-Penìa E, Moreno-Bondi MC (2019) Tag-specific affinity purification of recombinant proteins by using molecularly imprinted polymers. Anal Chem 91(6):4100–4106
83. Schmidt TGM, Batz L, Bonet L, Carl U, Holzapfel G, Kiem K et al (2013) Development of the Twin-Strep-tag® and its application for purification of recombinant proteins from cell culture supernatants. Protein Exp Purif 92:54–61
84. Peroutka RJ, Orcutt SJ, Strickler JE, Butt TR (2011) SUMO fusion technology for enhanced protein expression and purification in prokaryotes and eukaryotes. Methods Mol Biol 705:15–30
85. Rosano GL, Ceccarelli EA (2014) Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol 5:172
86. Nallamsetty S, Waugh DS (2006) Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. Protein Expr Purif 45(1):175–182
87. Costa S, Almeida A, Castro A, Domingues L (2014) Fusion tags for protein solubility, purification, and immunogenicity in Escherichia coli: the novel Fh8 system. Front Microbiol 5:63
88. de Marco A (2006) Two-step metal affinity purification of double-tagged (NusA-His6) fusion proteins. Nat Protoc 1(3):1538–1543
89. Arnau J, Lauritzen C, Petersen GE, Pedersen J (2006) Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins. Protein Exp Purif 48:1–13

90. Terpe K (2003) Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biotechnol 60:523–533
91. Nautiyal K, Kuroda Y (2018) A SEP tag enhances the expression, solubility and yield of recombinant TEV protease without altering its activity. New Biotechnol 42:77–84
92. Decaprio J, Kohl TO (2019) Tandem immunoaffinity purification using anti-FLAG and anti-HA antibodies. Cold Spring Harb Protoc 2019(2):147–153
93. Raran-Kurussi S, Waugh DS (2017) Expression and purification of recombinant proteins in Escherichia coli with a His6 or dual His6-MBP tag. In: Methods molecular biology. Humana Press, Totowa
94. Paraskevopoulou V, Falcone F (2018) Polyionic tags as enhancers of protein solubility in recombinant protein expression. Microorganisms. 6(2):47
95. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH (2017) Proteomics: technologies and their applications. J Chromatogr Sci 55:182–196
96. Zabel MD, Reid C (2015) A brief history of prions, vol. 73, Pathogens and disease. Oxford University Press, Oxford
97. Wilkins M (2009) Proteomics data mining. Expert Rev Proteomics 6:599–603
98. Xiao GG, Recker RR, Deng HW (2008) Recent advances in proteomics and cancer biomarker discovery, vol. 2: Clinical medicine: oncology. Libertas Academica Ltd., Los Angeles, pp 63–72
99. Xu D, Xu Y (2004) Protein databases on the internet. Curr Protoc Protein Sci, Chapter 2: Unit 2.6
100. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R et al (2011) Ongoing and future developments at the universal protein resource. Nucleic Acids Res 39(suppl 1):D214–D219
101. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ et al (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res 47(D1):D442–D450
102. Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V et al (2017) Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinformatics. 18(1):142
103. Grunberg R, Nilges M, Leckner J (2007) Biskit A software platform for structural bioinformatics. Bioinformatics 23(6):769–770
104. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. Nucleic Acids Res 33(suppl 2):W382–8
105. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10(6):845–858
106. Chen C-S, Zhu H (2006) Protein microarrays. BioTechniques 40(4):423–429
107. Da Gama DJ, Goosen RW, Lawry PJ, Blackburn JM (2018) PMA: protein microarray analyser, a user-friendly tool for data processing and normalization. BMC Res Notes 11(1):156
108. Díez P, Dasilva N, González-González M, Matarraz S, Casado-Vela J, Orfao A et al (2012) Data analysis strategies for protein microarrays. Microarrays 1(2):64–83
109. Ramachandran N, Srivastava S, LaBaer J (2008) Applications of protein microarrays for biomarker discovery. Proteomics Clin Appl 2:1444–1459
110. Huang Y, Zhu H (2017) Protein array-based approaches for biomarker discovery in cancer. Genomics Proteomics Bioinformatics 15:73–81
111. Hall DA, Ptacek J, Snyder M (2007) Protein microarray technology. Mech Ageing Dev 128(1):161–167
112. Ding X, Sun YS (2015) Use of microarrays as a high-throughput platform for label-free biosensin. J Lab Autom 20:334–353
113. Shen M, Rusling JF, Dixit CK (2017) Site-selective orientated immobilization of antibodies and conjugates for immunodiagnostics development. Methods 116:95–111
114. Hu S, Xie Z, Qian J, Blackshaw S, Zhu H (2011) Functional protein microarray technology. Wiley Interdiscip Rev Syst Biol Med 3(3):255–268

115. Sutandy FXR, Qian J, Chen CS, Zhu H (2013) Overview of protein microarrays. Curr Protoc Protein Sci Chapter 27(suppl 72):Unit 27.1
116. Baldelli E, Calvert V, Hodge A, VanMeter A, Petricoin EF, Pierobon M (2017) Reverse phase protein microarrays. In: Methods in molecular biology. Humana Press, Totowa, pp 149–169
117. Alhamdani MSS, Schröder C, Hoheisel JD (2009) Oncoproteomic profiling with antibody microarrays. Genome Med 1:68
118. Wulfkuhle JD, Edmiston KH, Liotta LA, Petricoin EF (2006) Technology insight: pharmacoproteomics for cancer - promises of patient-tailored medicine using protein microarrays. Nat Clin Pract Oncol 3:256–268
119. Liotta LA, Espina V, Mehta AI, Calvert V, Rosenblatt K, Geho D et al (2003) Protein microarrays: meeting analytical challenges for clinical applications. Cancer Cell 3:317–325
120. Poetz O, Schwenk JM, Kramer S, Stoll D, Templin MF, Joos TO (2005) Protein microarrays: catching the proteome. Mech Ageing Dev 126:161–170
121. Brichta J, Hnilova M, Viskovic T (2005) Generation of hapten-specific recombinant antibodies: antibody phage display technology: a review. Vet Med-Czech 50:231
122. Kricka LJ, Master SR, Joos TO, Fortina P (2006) Current perspectives in protein array technology. Ann Clin Biochem 43:457
123. Berrade L, Garcia AE, Camarero JA (2011) Protein microarrays: novel developments and applications. Pharm Res 28:1480–1499
124. Zandian A, Forsström B, Häggmark-Månberg A, Schwenk JM, Uhlén M, Nilsson P et al (2017) Whole-proteome peptide microarrays for profiling autoantibody repertoires within multiple sclerosis and narcolepsy. J Proteome Res 16(3):1300–1314
125. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P et al (2001) Global analysis of protein activities using proteome chips. Science 293(5537):2101–2105
126. Zhu H, Qian J (2012) Applications of functional protein microarrays in basic and clinical research. Adv Genet 79:123–155
127. Pjaweletz CP, Charboneau L, Bichsel VE, Simone NL, Chen T, Gillespie JW et al (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. Oncogene 20(16):1981–1989
128. Chen C, Huang H, Wu CH (2017) Protein bioinformatics databases and resources. Methods Mol Biol 1558:3–39
129. Chen C, Huang H, Wu CH (2011) Protein bioinformatics databases and resources. Methods Mol Biol 694:3–24
130. Islam SA, Luo J, Sternberg MJE (1995) Identification and analysis of domains in proteins. Protein Eng Des Sel 8(6):513–526
131. Liu J, Rost B (2003) Domains, motifs and clusters in the protein universe. Curr Opin Chem Biol 7:5–11
132. Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420:218–223
133. Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261(5561):552–558
134. Chothia C, Janin J (1975) Principles of protein-protein recognition. Nature 256(5520):705–708
135. Mehrotra P, Ami VKG, Srinivasan N (2018) Clustering of multi-domain protein sequences. Proteins Struct Funct Bioinformatics 86(7):759–776
136. Jin J, Xie X, Chen C, Park JG, Stark C, James DA et al (2009) Eukaryotic protein domains as functional units of cellular evolution. Sci Signal 2(98):ra76
137. McGuffin LJ, Bryson K, Jones DT (2001) What are the baselines for protein fold recognition? Bioinformatics 17(1):63–72
138. Chakrabarti S, Sowdhamini R (2004) Regions of minimal structural variation among members of protein domain superfamilies: application to remote homology detection and modelling using distant relationships. FEBS Lett 569(1–3):31–36
139. Basu MK, Poliakov E, Rogozin IB (2009) Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform 10(3):205–216

140. Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein Sci 7(2):233–242
141. Apic G, Huber W, Teichmann SA (2003) Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. J Struct Funct Genom 4(2–3):67–78
142. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA (2004) Supra-domains: evolutionary units larger than single protein domains. J Mol Biol 336(3):809–823
143. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. Science 300:1701–1703
144. Ochoa A, Llinás M, Singh M (2011) Using context to improve protein domain identification. BMC Bioinformatics 12:90
145. Scaiewicz A, Levitt M (2018) Unique function words characterize genomic proteins. Proc Natl Acad Sci U S A 115(26):6703–6708
146. Yu L, Tanwar DK, Penha EDS, Wolf YI, Koonin EV, Basu MK (2019) Grammar of protein domain architectures. Proc Natl Acad Sci U S A 116(9):3636–3645
147. Basu MK, Carmel L, Rogozin IB, Koonin EV (2008) Evolution of protein domain promiscuity in eukaryotes. Genome Res 18(3):449–461
148. Gracy J, Argos P (1998) Automated protein sequence database classification: II. Delineation of domain boundaries from sequence similarities. Bioinformatics 14(2):174–187
149. Suyama M, Ohara O (2003) DomCut: prediction of inter-domain linker regions in amino acid sequences. Bioinformatics 19(5):673–674
150. Ezkurdia I, Tress ML (2011) Protein structural domains: definition and prediction. Curr Protoc Protein Sci 1(suppl 66):UNIT 2.14
151. Söding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7):951–960
152. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M et al (2006) Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics 15(1):UNIT 5.6
153. Webb B, Sali A (2016) Comparative protein structure modeling using MODELLER. Curr Protoc Bioinforma. 2016:5.6.1–5.6.37
154. Fiser A, Šali A (2003) MODELLER: generation and refinement of homology-based protein structure models. Methods Enzymol 374:461–491
155. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the web: a case study using the phyre server. Nat Protoc 4(3):363–373
156. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 23:9
157. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4):725–738
158. Chivian D, Kim DE, Malmström L, Bradley P, Robertson T, Murphy P et al (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53:524–533
159. Muhammed MT, Aki-Yalcin E (2019) Homology modeling in drug discovery: overview, current applications, and future perspectives. Chem Biol Drug Design 93:12–20
160. Hatfield M, Lovas S (2014) Conformational sampling techniques. Curr Pharm Des 20(20):3303–3313
161. Hameduh T, Haddad Y, Adam V, Heger Z (2020) Homology modeling in the time of collective and artificial intelligence. Comput Struct Biotechnol J 18:3494–3506
162. Jalily Hasani H, Barakat K (2017) Homology modeling: an overview of fundamentals and tools. Int Rev Model Simul 10:129–145
163. Zhu J, Cheng L, Fang Q, Zhou ZH, Honig B (2010) Building and refining protein models within cryo-electron microscopy density maps based on homology modeling and multiscale structure refinement. J Mol Biol 397(3):835–851
164. Yip KM, Fischer N, Paknia E, Chari A, Stark H (2020) Atomic-resolution protein structure determination by cryo-EM. Nature 587(7832):157–161

165. Egelman EH (2016) The current revolution in Cryo-EM. Biophys J 110:1008–1012
166. Haddad Y, Adam V, Heger Z (2020) Ten quick tips for homology modeling of high-resolution protein 3D structures. PLoS Comput Biol 16:e1007449
167. Zhang Y (2009) Protein structure prediction: when is it useful? Curr Opin Struct Biol 19:145–155
168. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T et al (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42(W1):W252
169. Bienert S, Waterhouse A, De Beer TAP, Tauriello G, Studer G, Bordoli L et al (2017) The SWISS-MODEL repository-new features and functionality. Nucleic Acids Res 45(D1):D313–D319
170. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31(13):3381–3385
171. Rohl CA, Strauss CEM, Chivian D, Baker D (2004) Modeling structurally variable regions in homologous proteins with Rosetta. Proteins Struct Funct Genet 55(3):656–677
172. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H et al (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7(8):1511–1522
173. Källberg M, Margaryan G, Wang S, Ma J, Xu J (2014) Raptorx server: a resource for template-based protein structure modeling. Methods Mol Biol 1137:17–27
174. Ko J, Park H, Seok C (2012) GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. BMC Bioinformatics 13(1):198
175. Alquraishi M (2019) AlphaFold at CASP13. Bioinformatics 35(22):4862–4865
176. Adams MD, Kelley JM, Gocayne JD, Dubnick MAK, Polymeropoulos MH, Xiao H et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252(5013):1651–1656
177. Parkinson J, Blaxter M (2009) Expressed sequence tags: an overview. Methods Mol Biol 533:1–12
178. Wasmuth JD, Blaxter ML (2004) prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics 5:187
179. Ranjit N, Jones MK, Stenzel DJ, Gasser RB, Loukas A (2006) A survey of the intestinal transcriptomes of the hookworms, Necator americanus and Ancylostoma caninum, using tissues isolated by laser microdissection microscopy. Int J Parasitol 36(6):701–710
180. Boguski MS, Lowe TMJ, Tolstoshev CM (1993) dbEST — database for "expressed sequence tags". Nat Genet 4:332–333
181. Edwards NJ (2007) Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol 3:102
182. Ekins RP (1989) Multi-analyte immunoassay. J Pharm Biomed Anal 7(2):155–168
183. Ekins RP (1998) Ligand assays: from electrophoresis to miniaturized microarrays. In: Clinical chemistry. American Association for Clinical Chemistry Inc., Washington, DC, pp 2015–2030
184. Kopf E, Zharhary D (2007) Antibody arrays-an emerging tool in cancer proteomics. Int J Biochem Cell Biol 39:1305–1317
185. Chen CS, Korobkova E, Chen H, Zhu J, Jian X, Tao SC et al (2008) A proteome chip approach reveals new DNA damage recognition activities in Escherichia coli. Nat Methods 5(1):69–74
186. Chen CS, Sullivan S, Anderson T, Tan AC, Alex PJ, Brant SR et al (2009) Identification of novel serological biomarkers for inflammatory bowel disease using Escherichia coli proteome chip. Mol Cell Proteomics 8(8):1765–1776
187. Hsiao FSH, Sutandy FR, Da SG, Chen YW, Lin JM, Chen CS (2016) Systematic protein interactome analysis of glycosaminoglycans revealed YcbS as a novel bacterial virulence factor. Sci Rep 6:28425
188. Ho YH, Shah P, Chen YW, Chen CS (2016) Systematic analysis of intracellular-targeting antimicrobial peptides, bactenecin 7, hybrid of pleurocidin and dermaseptin, proline-

arginine-rich peptide, and lactoferricin b, by using Escherichia coli proteome microarrays. Mol Cell Proteomics 15(6):1837–1847
189. Xu Z, Zhang H, Zhang X, Jiang H, Liu C, Wu F et al (2019) Interplay between the bacterial protein deacetylase CobB and the second messenger c-di-GMP. EMBO J 38(18):e100948
190. Feng Y, Chen CS, Ho J, Pearce D, Hu S, Wang B et al (2018) High-throughput chip assay for investigating Escherichia coli interaction with the blood-brain barrier using microbial and human proteome microarrays (dual-microarray technology). Anal Chem 90(18):10958–10966
191. Xiu WZ, Ping DR, Jiang HW, Guo SJ, Ying LH, Dong ZX et al (2012) Global identification of prokaryotic glycoproteins based on an Escherichia coli proteome microarray. PLoS One 7(11):e49080
192. Huang BY, Chen PC, Chen BH, Wang CC, Liu HF, Chen YZ et al (2017) High-throughput screening of sulfated proteins by using a genome-wide proteome microarray and protein tyrosine Sulfation system. Anal Chem 89(6):3278–3284
193. Tsai CH, Ho YH, Sung TC, Wu WF, Chen CS (2017) Escherichia coli proteome microarrays identified the substrates of clpyq protease. Mol Cell Proteomics 16(1):113–120
194. Hall DA, Zhu H, Zhu X, Royce T, Gerstein M, Snyder M (2004) Regulation of gene expression by a metabolic enzyme. Science 306(5695):482–484
195. Huang J, Zhu H, Haggarty SJ, Spring DR, Hwang H, Jin F et al (2004) Finding new components of the target of rapamycin (TOR) signaling network through chemical genetics and proteome chips. Proc Natl Acad Sci U S A 101(47):16594–16599
196. Zhu J, Gopinath K, Murali A, Yi G, Hayward SD, Zhu H et al (2007) RNA-binding proteins that inhibit RNA virus infection. Proc Natl Acad Sci U S A 104(9):3129–3134
197. Yi LY, Ying LJ, Zhang J, Walter W, Dang W, Wan J et al (2009) Protein acetylation microarray reveals that NuA4 controls key metabolic target regulating gluconeogenesis. Cell 136(6):1073–1084
198. Song Q, Liu G, Hu S, Zhang Y, Tao Y, Han Y et al (2010) Novel autoimmune hepatitis-specific autoantigens identified using protein microarray technology. J Proteome Res 9(1):30–39
199. Zingaretti C, Arigò M, Cardaci A, Moro M, Crosti M, Sinisi A et al (2012) Identification of new autoantigens by protein array indicates a role for IL4 neutralization in autoimmune hepatitis. Mol Cell Proteomics 11(12):1885–1897
200. Bian X, Wasserfall C, Wallstrom G, Wang J, Wang H, Barker K et al (2017) Tracking the antibody immunome in type 1 diabetes using protein arrays. J Proteome Res 16(1):195–203
201. Landegren N, Sharon D, Shum AK, Khan IS, Fasano KJ, Hallgren Å et al (2015) Transglutaminase 4 as a prostate autoantigen in male subfertility. Sci Transl Med 7(292):292ra101
202. Hu CJ, Pan JB, Song G, Wen XT, Wu ZY, Chen S et al (2017) Identification of novel biomarkers for behcet disease diagnosis using human proteome microarray approach. Mol Cell Proteomics 16(2):147–156
203. Anderson KS, Cramer DW, Sibani S, Wallstrom G, Wong J, Park J et al (2015) Autoantibody signature for the serologic detection of ovarian cancer. J Proteome Res 14(1):578–586
204. Syed P, Gupta S, Choudhary S, Pandala NG, Atak A, Richharia A et al (2015) Autoantibody profiling of glioma serum samples to identify biomarkers using human proteome arrays. Sci Rep 5:13895
205. Zhang HN, Yang L, Ling JY, Czajkowsky DM, Wang JF, Zhang XW et al (2015) Systematic identification of arsenic-binding proteins reveals that hexokinase-2 is inhibited by arsenic. Proc Natl Acad Sci U S A 112(49):15084–15089
206. Pan J, Song G, Chen D, Li Y, Liu S, Hu S et al (2017) Identification of serological biomarkers for early diagnosis of lung cancer using a protein array-based approach. Mol Cell Proteomics 16(12):2069–2078
207. Adeola HA, Smith M, Kaestner L, Blackburn JM, Zerbini LF (2016) Novel potential serological prostate cancer biomarkers using CT100+ cancer antigen microarray platform in a multi-cultural South African cohort. Oncotarget 7(12):13945–13964

208. Gupta S, Mukherjee S, Syed P, Pandala NG, Choudhary S, Singh VA et al (2017) Evaluation of autoantibody signatures in meningioma patients using human proteome arrays. Oncotarget 8(35):58443–58456
209. Song G, Rho HS, Pan J, Ramos P, Yoon KJ, Medina FA et al (2018) Multiplexed biomarker panels discriminate Zika and dengue virus infection in humans. Mol Cell Proteomics 17(2):349–356
210. Song L, Wallstrom G, Yu X, Hopper M, Van Duine J, Steel J et al (2017) Identification of antibody targets for tuberculosis serology using high-density nucleic acid programmable protein arrays. Mol Cell Proteomics 16(4):S277–S289
211. Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 7:55–65
212. DeLuca DS, Marina O, Ray S, Zhang GL, Wu CJ, Brusic V (2011) Data processing and analysis for protein microarrays. Methods Mol Biol 723:337–347
213. Olle EW, Sreekumar A, Warner RL, McClintock SD, Chinnaiyan AM, Bleavins MR et al (2005) Development of an internally controlled antibody microarray. Mol Cell Proteomics 4(11):1664–1672
214. Eckel-Passow JE, Hoering A, Therneau TM, Ghobrial I (2005) Experimental design and analysis of antibody microarrays: applying methods from cDNA arrays. Cancer Res 65:2985–2989
215. Sill M, Schröder C, Hoheisel JD, Benner A, Zucknick M (2010) Assessment and optimisation of normalisation methods for dual-colour antibody microarrays. BMC Bioinformatics. 11:556