ELSEVIER

HTS technologies in biopharmaceutical discovery

Ge Wu and Stephen K. Doberstein

Five Prime Therapeutics, 1650 Owens St., Suite 200, San Francisco, CA 94158, USA

The concepts and philosophies of HTS can be productively applied to the discovery of new biopharmaceuticals. It is now possible, comprehensively and systematically, to enumerate, clone, produce and screen all secreted proteins, by building upon knowledge accumulated over the past two decades in HTS, genomics and parallel protein expression technologies. Each of the crucial operational components (comprehensive and high-quality cDNA library construction, proper protein-sequence classification, high-throughput protein production, medically relevant assays, state-of-the-art screening and data management) must be optimized to increase the chances of success. In this review, we draw comparisons between small-molecule and protein screening to illuminate common underlying principles as well as differences between the two operations.

Two decades after the launch of the first human recombinant protein (Humulin®) in 1982, protein biopharmaceuticals now account for a significant proportion of total FDA approvals [1–4]. The major 'blockbuster' biopharmaceuticals fall into two categories [5,6]: secreted proteins normally encoded by the genome, including cytokines, growth factors and hormones that activate cellular receptors; and monoclonal antibodies or soluble receptors capable of blocking the activity of cell-surface receptors, hence preventing activation by their native ligands. The commercial landscape is crowded and competitive, particularly in the field of antibody discovery and development [7]. Although overall approvals and sales of protein drugs have been increasing, R&D productivity (in terms of approvals per R&D spend) has dropped, and there is a clear need for new approaches to protein drug discovery.

It is estimated that there are roughly 3500 to 4000 gene loci encoding secreted proteins and single-pass cell-surface receptors in the human genome. These proteins represent the fundamental set of secreted protein therapeutic candidates and a large proportion of potential antibody targets. Although advances in genomics have impacted drug discovery considerably [8], functional understanding of gene products has lagged behind the sequencing effort and only a small fraction of proteins encoded by the human genome have been assigned functions [9,10]. For example, of ~4000 proteins in the secreted and single-pass transmembrane (STM) protein classes we fully understand the function and pharmacology of perhaps 15%, and nearly all existing protein therapeutics emerge from that fraction. It is not unorthodox to expect that more secreted proteins and antibody targets of therapeutic value are contained in the remaining, not fully characterized, fraction. Although the application of genomics (primarily transcriptional profiling) to secreted proteins and cell-surface receptors has attracted much attention [11–14], there remains the need to determine which, from the trove of uncharacterized secreted proteins, have medically relevant pharmacology.

Concurrent with the human genome sequencing effort, improvements in HTS technologies have solidified the role HTS has as a major driving force for small-molecule discovery [15–17]. HTS encompasses small-molecule library design and assembly, robotics, assay development and data handling. Early in the HTS era (1980s), compound screening libraries were primarily mixtures of compounds, robotics to handle the newly developed 96-well microtiter plate were not reliable, screening data were stored in flat files, and the throughput of even a large screening system was <100 plates per day. Twenty years later libraries of millions of compounds are screened routinely with single compounds of high purity in each well. Fully automated robotic systems greatly improve screening data quality and have virtually

Corresponding author: Wu, G. (ge.wu@FivePrime.com)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.06.010

eliminated human error with many operations running continuously, screening entire libraries of compounds with little or no human oversight. Most screens are now executed in 384- and 1536-well formats and data storage has evolved from primitive flat files to versatile relational databases. HTS-specific theories have also emerged, such as the Z' factor for quantitation of assay reproducibility and substrate conversion equations that aid enzymatic assay development [18,19]. By the end of 2003 HTS approaches had generated 74 leads in clinical development and two marketed drugs [20].

Application of HTS approaches to protein libraries

The application of HTS concepts and technologies to biopharmaceutical discovery demonstrated success even before the completion of the human genome project [13,14,21]. Although the universe of small molecules is estimated to approach 10^{200} [22], the total number of native non-antibody secreted proteins and cell-surface receptors is modest by HTS standards. The combination of new molecular biology and protein-expression technologies, and the judicious application of high-throughput and high-content screening tools, now make it possible to screen the entire set of candidates comprehensively for biotherapeutic lead molecules. This review describes the state-of-the-art discovery of novel biopharmaceuticals by HTS.

Every protein screening effort has three main stages: the selection and establishment of a cDNA library encoding the individual proteins to be screened; the conversion of the cDNA library into a functional protein library; and screening of the protein library for therapeutically relevant proteins. The multi-step nature of the process means that the failure rate in each step multiplies (i.e. a 50% success rate in cloning and a 50% success rate in expression yields only 25% of the total possible proteins in the screening step). It is crucial to maximize yield when developing the cDNA library and protein expression system, and to emphasize quality in screening operations.

cDNA library construction and classification

The first step of a protein screening effort is the establishment of a library of cDNA clones that encode the proteins to be assayed. Early efforts to catalogue the proteins encoded in the human genome were based on expressed sequence tags (ESTs), which are useful for tagging the location of expressed gene sequences. However, establishment of a robust screening library requires accurate and complete knowledge of the coding sequence of each gene, and EST-based prediction remains subject to the limitations of gene prediction. Computational gene predictions from human genomic sequences, even with the aid of ESTs, are error-prone because of long introns and short exons [23]. Gene prediction based on full-length cDNA sequencing remains the most accurate method for gene-sequence prediction [24–27]. RefSeq and LocusLink are useful resources for complete genomic nucleic acids, assembled contigs, transcripts and proteins [28]. A modest number of full-length physical clones for protein expression can be obtained from the NIH Mammalian Gene Collection (MGC), they were initially generated by random selection from a large number of cDNA libraries [29]. Tissue-specific expression of some cDNAs and splice variants could limit their availability in cDNA collections derived from limited tissue sources. The most accurate way to establish a comprehensive cDNA collection for screening purposes remains careful construction of a well-curated full-length cDNA library from a large number of tissue sources, coupled with the most advanced full-length cDNA cloning technologies [30–33].

The next task is accurate curation of the cDNA collection to select the components of the library. Not all proteins represent candidate biotherapeutics and, similar to drug-like small molecules, can be selected using computational chemistry. Proteins can be classified to focus on those most relevant for discovery research. Secreted proteins and cell-surface STM proteins are the two categories most relevant to biopharmaceutical development. Most proteins targeted for secretion (or translocation to the plasma membrane) have N-terminal signal peptides of 20–40 amino acids in length, and distinct characteristics that enable their prediction by computer programs. There are many computational algorithms available for prediction of signal sequences [34–36]. However, each algorithm has limitations. SignalP, the best public domain algorithm, can predict secreted proteins with an accuracy of 78.1% [37]. The transmembrane proteins typically form hydrophobic α-helical structures that are long enough to traverse the membrane, and transmembrane domains can be predicted with some accuracy. The TMpred program can predict membrane-spanning regions and their orientation [38]. The algorithm is based on the statistical analysis of TMbase, a database of naturally occurring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring. Because a functional domain within a protein class can be conserved in evolution, motif annotation programs, such as Pfam, Prosite and eMatrix, can also aid in protein classification [39]. In addition to signal-sequence-containing proteins, there are increasing numbers of proteins that are secreted through a non-signal sequence-mediated secretion pathway [40–42], and the computational tools to predict these have only just begun to be developed [43]. Further issues in library protein selection arise when one considers that many important signaling ligands, such as those in the epidermal growth factor (EGF) and tumor necrosis factor (TNF) families, are encoded as STM proteins and are only released by proteolysis at the cell surface. Ultimately, completeness is essential in library selection and a combination of increasingly advanced computer models (for example hidden Markov techniques) and careful manual sequence curation is required for robust library design.

STM proteins that are on the cell surface that are represented in screening libraries as extracellular domains (ECDs), present challenges and opportunities. The opportunities for this set of proteins are broad because some ECDs represent potential signaling ligands, whereas others represent receptors, which would emerge from screening as inhibitors (through neutralization of their cognate ligands). For example, we recently discovered a novel cytokine involved in monocyte differentiation and rapidly determined its signaling receptor by HTS of ECDs for inhibitors of that cytokine (manuscript in preparation). Each STM protein sequence must be carefully curated to determine the most likely localization of the protein (the plasma membrane components being most important for pharmaceutical purposes) and each sequence must be subcloned as a truncated cDNA, deleting the transmembrane and intracellular domains. Some ECDs are best expressed as fusion proteins [e.g. fused to the Fc (constant) domain of immunoglobulin] for stability and high-level expression. Although there is no

guarantee that all ECDs will fold properly, there exist many examples of STM ligands that are functionally expressed as ECDs (TNF family members, EGF family members) and receptor ECD–Fc fusions neutralizing their cognate ligands [TNF receptors, EGF receptors, fibroblast growth factor (FGF) receptors].

As noted above, the difficulty of comprehensive protein library design is compounded by the need to include those proteins that could have extracellular pharmacology, despite the lack of a canonical signal sequence. Some important signaling proteins are actively secreted without a signal sequence [40–42] and models are being developed to predict additional members of this family. Remarkably, some proteins known to have intracellular functions also have extracellular signaling roles [44,45]. The challenge for the protein library designer is to be as comprehensive as possible in the face of new data, at the same time also being cost-effective and limiting the library to those proteins that might reasonably be turned into drugs or drug targets.

Converting cDNA collections to protein screening libraries

Although the construction and maintenance of a protein library superficially resembles the corresponding process for a small-organic-molecule library, there are important differences in operational details for working with protein libraries. Ultimately, one would like to express all the predicted secreted proteins in the native cells (where they are made and processed), purify the secreted proteins and store them in a universal medium that is compatible with downstream assays. Operational and scientific considerations make this infeasible, and typically a single or at most a small number of complementary expression hosts are used for converting the cDNA collection to proteins for screening.

One difference is scale. As noted above, although drug-like small-molecule structural space is almost infinite, there is a relatively moderate number of proteins in humans. The primary components of a protein screening library, the canonically secreted proteins, are encoded by <4000 gene clusters. The total number of protein variants and isoforms will be slightly higher when one considers splice variation and post-translational modifications, such as enzymatic cleavage, phosphorylation, sulfation, lipidation and glycosylation. Nonetheless, it remains possible to screen all of the secreted and STM proteins in cell-based assays, from a library that is modest by HTS standards.

Protein stability in protein libraries is an issue with profound operational implications. Small molecules are generally stable both in dry form and in dimethyl sulfoxide (DMSO) – a near universal solvent for small organic molecules. There is no proven medium that can store all proteins for lengthy periods of time (as would be necessary for the development of a permanent protein library). Degradation and precipitation are major problems facing long-term storage of proteins and it is, therefore, necessary to produce the proteins fresh from a cDNA-source collection for robust and comprehensive screening results.

The main goal when designing a protein screening process is to ensure that a sufficient concentration of the protein to be screened is available (in the assay well) in a soluble, active form. Automated parallel expression of a large number of proteins has been described for protein-structure studies [46]. In this approach proteins were tagged, allowing for a similar affinity purification procedure for most targets and the implementation of high-throughput parallel purification methods. In most cases, these affinity-purified proteins are sufficiently homogeneous so that a single subsequent gel-filtration chromatography step is adequate to produce protein preparations that are >98% pure. Using this platform, >1000 different proteins were cloned, expressed and purified in milligram quantities over a 36-month period. Techniques for expression of tagged proteins in a 96-well format have also been described [47]. Although epitope tagging aids in the generalization of purification protocols and can be used to increase protein concentrations in the assay plate, it could also render the protein less active or, in extreme cases, ablate activity altogether.

Systems that obviate the need for purification, and therefore allow screening of untagged proteins, are preferable. Many proteins have been expressed in an active form in microbial systems. However, on a genome-wide basis, expression of native proteins in cultivated mammalian cells is superior for proper protein folding, assembly and post-translational modification [48]. High-throughput transfection of a large number of full-length human cDNA clones in mammalian expression vectors has been demonstrated in HeLa and HEK293 cells on microscope slides [10]. A larger-scale example demonstrated the parallel transfection of ~20,000 full-length cDNAs arrayed in 384-well microtiter plates, with each well containing one target gene and a reporter gene [12]. These genes were transiently co-transfected in serum-free medium into HEK293, HCT116 and HepG2 cells, and the reporter assays were performed in the same well. Although this technique has an advantage of simplicity by combining transfection, expression and the assay in one well (to improve throughput and reliability), a limitation of this approach is that assay formats are restricted by the need for compatibility among transfection, expression and assay conditions.

Separation of assay readout from the expression system permits single-well transfection and expression while allowing flexibility with respect to assay formats and cell types. High-throughput transient transfection of each cDNA can be performed in 96- or 384-well microtiter plates, using an automated robotic system such as the one shown in Figure 1. The supernatants are then separated from the transfected cells to form the protein library that can be used for assays in diverse formats, using different medically relevant cell types. Because of the idiosyncratic stability of proteins noted above, it is preferable to use the protein library immediately after it has been produced. Our total number of library proteins is close to 4000, a number that can be produced in a single day in a 384-well format. Although this system is robust, maximum accuracy in liquid-handling and sample tracking is essential because even a modest spillover from well to well at the cDNA stage can amplify to very large signals in the screening step.

HTS assay development

Development of assays for HTS requires high standards of reproducibility, scalability, robustness and compatibility with automation [17,49]. HTS assays are usually performed in microtiter plates with 96–1536 wells. Typical detection methods suitable for HTS include fluorescence, absorbance, luminescence, fluorescence polarization (FP), time-resolved fluorescence (TRF), fluorescence resonance energy transfer (FRET), time-resolved-FRET, scintillation proximity assays (SPA), AlphaScreen™, electrochemiluminescence (ECL) and

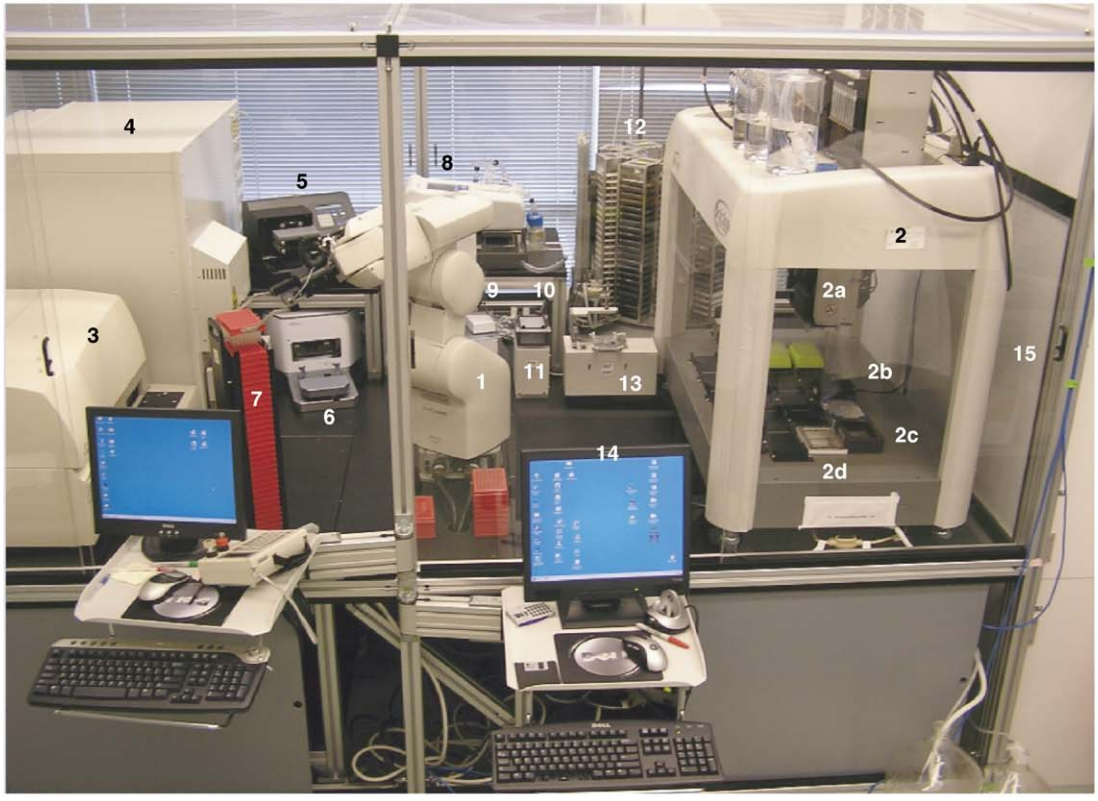

FIGURE 1

A fully integrated multifunctional robotic screening system. The system is fully enclosed and comprises the following components: **(1)** Mitsubishi MELFA RV2A 6-axis robot; **(2)** Caliper Sciclone ALH3000 with interchangeable 96- or 384-tip pipetting head, an independent 8-channel pipettor, two bulk-reagent dispensers, and plate gripper **(2a)**. The following accessories are integrated into the Sciclone: microtiter plate shaker **(2b)**; positive-pressure filtration system **(2c)**; and ultrasonic tip-wash station **(2d)**. **(3)** PerkinElmer Fusion with 11-mode detection, which includes absorbance, fluorescence, fluorescence polarization, time-resolved fluorescence, time-resolved fluorescence–resonance energy transfer, AlphaScreen™, etc. **(4)** Kendro Cytomat6001 with humidity, temperature and CO_2 controls, and 189 normal microtiter plate storage capacity. **(5)** Biotek ELX-405 plate washer, which can be used for 96-well and 384-well plates. **(6)** Volecity11 Vspin centrifuge can be used for normal- and deep-well plates. **(7)** Thermo CRS high-capacity stacker is used to store up to 32 stacked tip boxes. **(8)** PerkinElmer Flexdrop equipped with four individual dispensing heads that can dispense four bulk reagents in a broad volume range for each head (from 200 nl to 2 ml). **(9)** Velocity11 Vcode automatic barcode labeler. **(10)** MicroScan MS-3 barcode reader. **(11)** Caliper plate regrip station that changes the plate orientation to facilitate the interaction between the robot arm and individual components. **(12)** Kendro room temperature incubator that stores 189 regular microtiter plates. **(13)** Caliper plate-lid-handling station. **(14)** Six Variomag shaker station that provides an independent plate-shaking operation (behind the monitor, not visible). **(15)** Liberty Industry air purifier provides ultra-dust-free conditions for the enclosed system and prevents the introduction of contaminants, from the surrounding air, to the work area. The reader table is modular and can be swapped for another reader if necessary.

automated cellular fluorescence imaging, among others. Assays can be cell-based or biochemical and can be homogeneous or heterogeneous. The assays are designed to screen hundreds of thousands of compounds per run, so each plate must contain internal control wells to ensure quality and comparability of results between plates. Because the samples are typically run as singlets, high-quality assays ($Z' > 0.5$) are required to identify active compounds unequivocally [18].

Protein screening presents special challenges and opportunities. As noted previously, the modest library size allows more-complex and time-consuming assays to be efficiently handled. Typically, assays are based on cellular readouts (phenotypic screening) rather than on direct molecular interactions, hence, careful selection of media and conditions compatible with cell-type to be assayed are essential. Finally, 'you get what you screen for'. It is essential to define the desired disease-modifying activity before starting assay development; otherwise a focused, rapid drug-screening effort can be reduced to mere signal-seeking in search of a therapeutic hypothesis.

Multiple parallel assays

Screening assays can be divided into two main categories: target-based screening and phenotype-based screening. In target-based screening a target protein is selected after the disease biology is, at least partially, defined (validated). The target protein is assayed by either measuring the interaction of a purified protein(s) with a compound in the screening library (e.g. immunoassay, enzyme assay, binding between protein and protein–nucleic-acid and/or carbohydrate interactions) or via measurement of output signals from a cell line that expresses the target protein (e.g. reporters, Ca^{2+} flux, labeled protein translocation, and so on). The target-based screening approach is typical of most modern small-molecule

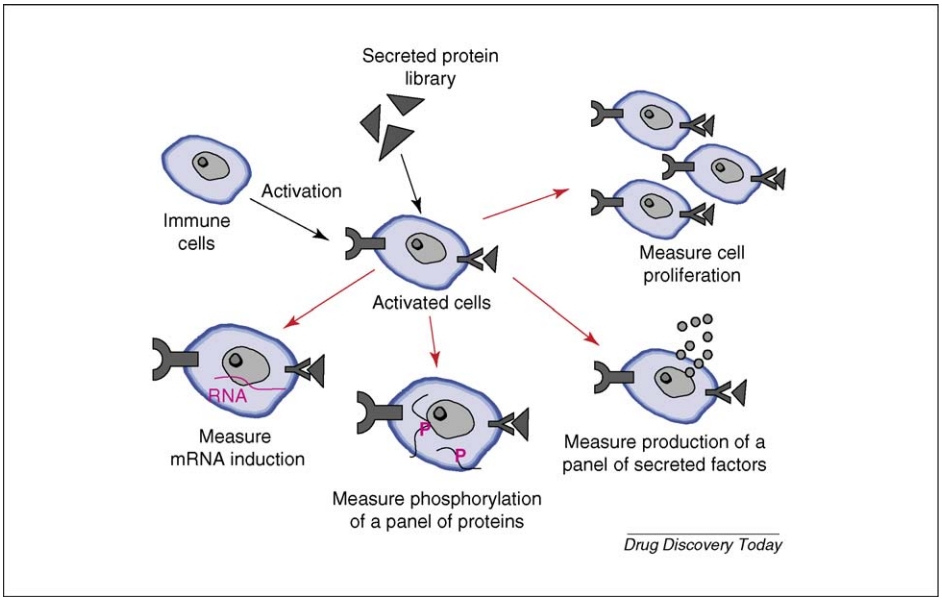

FIGURE 2

Parallel screening for secreted proteins that act on immune cells (B cells, T cells, monocytes and natural killer cells). First, immune cells are partially activated by different reagents, such as bacterial cell wall components or co-activating proteins. This activation exposes many receptors on the cell surface that the naïve cell does not express and which bind to different ligands. The activated cells are then incubated with secreted library proteins. Multiple parallel assays are then performed that include cell proliferation, changes in cytokine secretion, phosphorylation of specific proteins and induction of specific mRNAs.

discoveries (e.g. targeting proteases, kinases, G-protein-coupled receptors, transferases, etc.). In phenotype-based screening a target cell or a pathway is hypothesized to be important in pathophysiology. Changes in cellular characteristics (e.g. morphology, cell number, metabolic rate) or in a focused metabolic or signaling pathway (e.g. gluconeogenesis, MAP kinase, cAMP, apoptosis) are measured in live cells [9,50].

In protein therapeutic discovery, nearly all assay development is focused on phenotype-based screening. For example, a screening campaign for discovery of novel secreted factors that modulate the immune system is illustrated in Figure 2. Because there is no defined target or pathway it is essential to perform multiple parallel assays to catch all possible regulators of the target cells. Of course, a protein screening system might also be very useful in the discovery of new drug targets amenable to small-molecule screening through judicious assay selection. Careful consideration should be given to each new hit with respect to the best pharmacological mode of attack, whether it is a protein, an antibody or a small molecule.

Cell targets

Owing to the fact that secreted proteins primarily act at the plasma membrane and their natural targets in the body are native cells, primary human cells should be considered first in assay development. If they are not available, primary mammalian cells with functions known to be preserved in evolution should be the second choice. However, lack of cross-species reactivity can be an issue, particularly for proteins active in the control of the immune system, which are generally poorly conserved between human and rodents. Another choice is cell lines; however these are transformed cells with profoundly perturbed biology and, in many cases, lack receptors that can interact with secreted proteins. For example, the L6 cell line, widely used to mimic muscle cells, does not express, or expresses at a dramatically reduced level, EGF-family receptors – in contrast to primary muscle cells. In addition, adipocyte cell lines rapidly lose expression of the insulin receptor and other metabolically important receptors. In the absence of a human primary cell source it is generally useful to screen rodent primary cells as well as human cell lines, to maximize the likelihood of success.

Assay media

The medium in which the protein library is generated can cause artifacts in cell-based assays, leading to high backgrounds, false-negatives or false-positives. In small-molecule screening, compounds stored in 100% DMSO are diluted to a final concentration that exerts minimal effects on the assay (usually <5% DMSO for enzymatic assays and <0.5% DMSO for cell-based assays). For proteins produced for screening in mammalian cell lines, the supernatants in which the proteins were originally secreted usually contain ~5% serum for optimal protein production. Because the proteins from the transfection supernatant are typically used in a nonconcentrated form for screening, dilution significantly reduces assay concentration and might risk missing poorly expressed proteins or less-potent hits. Generally supernatants should be assayed with a dilution factor of 2–5. By contrast, 100-fold dilution from original small-molecule master libraries is common, reducing the final concentration of DMSO to <1%. Because some assay formats and cell types cannot accommodate even low concentrations of serum, it might be necessary to compromise protein production by the use of serum-free media, which typically results in lower protein concentrations. For those proteins produced in microbial cells serum is not an issue, but contaminants (such as cell wall components and bacterial DNA) can

elicit potent artifact responses by activating receptors of the innate immune system. In addition, non-protein metabolite composition (such as glucose concentration) might differ from well to well because of idiosyncratic differences in metabolic rate, caused by the transfected genes. In all cases, careful assay development is essential and attention must be paid to possible artifacts at every step in the process.

Assay format

The use of primary cells, rather than cell lines, creates challenges throughout assay development. Scarce cell sources, donor variability, cell viability and stromal environment are all crucial variables that must be understood and controlled. For example, only a limited quantity of immune cells (such as natural killer cells) can be obtained from a single donor at one time. Immune cells from different donors will react with each other and cannot be pooled, so primary cells from multiple donors have to be used separately in screening and can result in variation, as a result of genetic background and health of the donor, which is not an issue when using cell lines. Most primary cells can only be cultured for limited times. For example, freshly isolated primary adipocytes survive for $<$24 h in culture, requiring an assay format that can be executed within hours rather than days. In such cases, higher-sensitivity assay formats will offer advantages. For example, choosing a fluorogenic substrate over a chromogenic substrate will reduce assay time by a factor of 10–100, as a result of the greater sensitivity in detecting a fluorescence signal compared with an absorbance signal. Some primary cells (such as cardiomyocytes and chondrocytes) will rapidly dedifferentiate into fibroblast-like cells after being plated on a 2D surface. 3D matrices can be used to maintain cellular phenotypes, but this imposes challenges in liquid handling and cell dispensing [51]. Finally, some cellular activities only occur in intact tissues. For example, pancreatic β cells will secrete insulin from excised islets but rapidly lose this characteristic upon disaggregation. Dispensing of aggregated cells and tissue slices into microtiter plates can be challenging, with the risk of large well-to-well variation. Application of new technologies can help to overcome some of these difficulties. For example, high-throughput automated imaging systems enable the study of specific cell types within a mixed population of cells [52]. Because it remains difficult to transfect reporter genes into some primary cells without changing the cells' characteristics, label-free technologies can play a unique role in primary cell screening because no genetic manipulation of the cells is required [53–55].

Screening operation

After two decades of innovation in HTS technology with a major focus on increased throughput, technology development efforts have recently shifted to enable the generation of more-reliable data and to broaden the application of the HTS concept to new areas [49,56,57]. In small-molecule screening, a typical compound library usually contains 0.5–2.0 million compounds. The assays are usually designed for well-defined target proteins, either isolated in pure form or overexpressed in a cell line. Modern large-scale integrated screening systems can be used to screen a million compounds in one assay in a few days, using 1536-well (or higher density) plates. Commercial database packages specifically designed for small-molecule HTS can be used to handle the screening data. Such an operation is not only expensive to build but also not necessarily ideal for a screening program focused on protein therapeutics. However, the fundamental concepts and practices of HTS can be of significant value in designing screening operations for proteins [58].

Because of the moderate size of the protein library and the large number of assays that have to be performed for each protein simultaneously (owing to protein stability), a screening operation for proteins should not be aimed primarily at maximizing throughput but rather at achieving high-quality data and maximum flexibility in handling a variety of assays. An integrated robotics system with a single articulated robotic arm that can randomly access multiple functional components and is 96- and/or 384-well-plate compatible is generally sufficient. HTS planners should first decide what functionalities are required in a system and then carefully pick the most reliable individual components that perform multiple tasks; the system should also ideally have a small footprint. It is very important to study the specifications of each component, and to ask competing vendors for on-site demonstrations. Most importantly, the best way to determine historical reliability of components is through a discussion with a network of previous end-users. A fully automated multifunctional screening system built at Five Prime Therapeutics (San Francisco, CA, USA) is shown in Figure 1 to demonstrate these principles.

Staggering quantities of data are continuously generated in Five Prime's typical HTS laboratory, requiring the use of a stable and fail-safe database system. An HTS-specific application from a third party, such as ActivityBase from IDBS (Guildford, Surrey, UK), or a 'home-made' application based on reliable and scalable relational databases, such as Oracle, is commonly employed in HTS operations. The application must be capable of automated data storage, retrieval and analysis to aid hit-picking. The database should ensure data integrity, be secure and flexible, and should have a user-friendly web-based graphic user interface (GUI) to provide an integrated view of data [59]. Data quality control is as important as compound quality control in a well-run HTS system [60], and close relationships between the network, database and assay-execution teams is essential for success.

Conclusion

For the first time, it is possible to enumerate accurately an almost complete list of secreted and STM proteins, from which most future protein therapeutics and antibodies will be derived. Biopharmaceutical discovery productivity can be improved, and timelines shortened, through the assembly and screening of proteins in a manner analogous to small-molecule HTS. Only by taking advantage of the vast storehouse of HTS knowledge in the small-molecule world can such an effort be maximally effective. However, there are significant operational and scientific differences between the two screening areas in terms of scale, library construction, library stability and assay formats. Although much progress has been made, current parallel protein expression and assay systems will benefit from further improvement. For example, the current state-of-the-art technologies in protein expression and purification do not permit the assembly and long-term storage of a protein library of native sequences (with known concentrations of purified protein in each well). Ultimately, new understanding of

the underlying biology of disease will be needed to enable assay design. Furthermore, we expect protein library screening for new stem-cell factors to be of tremendous value in the future.

Acknowledgement

We would like to thank our team at Five Prime Therapeutics, especially Dr Rusty Williams for his scientific guidance.

References

1 Nagle, T. *et al*. (2003) The further evolution of biotech. *Nat. Rev. Drug Discov.* 2, 75–79
2 Pavlou, A.K. and Reichert, J.M. (2004) Recombinant protein therapeutics success rates, market trends and values to 2010. *Nat. Biotechnol.* 22, 1513–1519
3 Brekke, O.H. and Sandlie, I. (2003) Therapeutic antibodies for human diseases at the dawn of the twenty-first century. *Nat. Rev. Drug Discov.* 2, 52–62
4 Walsh, G. (2005) Current status of biopharmaceuticals: approved products and trends in approvals. In *Modern biopharmeceuticals: design, development and optimization* (Knablein, J., ed.), pp. 1–34, Wiley-VCH Verlag
5 Walsh, G. (2003) *Biopharmaceuticals: biochemistry and biotechnology* (2nd edn), John Wiley & Sons
6 Ho, R.J.Y. and Gibaldi, M. (2003) *Biotechnology and biopharmaceuticals: transforming proteins and genes into drugs*. Wiley–Liss
7 Mitchell, P. (2005) Next-generation monoclonals less profitable than trailblazers? *Nat. Biotechnol.* 23, 906
8 Emilien, G. *et al*. (2000) Impact of genomics on drug discovery and clinical medicine. *QJM* 93, 391–423
9 Jackson, P.D. and Harrington, J.J. (2005) High-throughput target discovery using cell-based genetics. *Drug Discov. Today* 10, 53–60
10 Hodges, E. *et al*. (2005) Accelerated discovery of novel protein function in cultured human cells. *Mol. Cell. Proteomics* 4, 1319–1327
11 Harada, J.N. *et al*. (2005) Identification of novel mammalian growth regulatory factors by genome-scale quantitative image analysis. *Genome Res.* 15, 1136–1144
12 Chanda, S.K. *et al*. (2003) Genome-scale functional profiling of the mammalian AP-1 signaling pathway. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12153–12158
13 Clark, H.F. *et al*. (2003) The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. *Genome Res.* 13, 2265–2270
14 Fiscella, M. *et al*. (2003) TIP, a T-cell factor identified using high-throughput screening increases survival in a graft-versus-host disease model. *Nat. Biotechnol.* 21, 302–307
15 Janzen, W.P. (2002) *High throughput screening: methods and protocols*. Humana Press
16 Devlin, J. (1997) *High throughput screening*. CRC Press
17 Seethala, R. and Fernandes, P.B. (2001) *Handbook of drug screening*. Marcel Dekker
18 Zhang, J.-H. *et al*. (1999) A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* 4, 67–73
19 Wu, G. *et al*. (2003) Determining appropriate substrate conversion for enzymatic assays in high-throughput screening. *J. Biomol. Screen.* 8, 694–700
20 Fox, S. *et al*. (2004) High-throughput screening: searching for higher productivity. *J. Biomol. Screen.* 9, 354–358
21 Chen, C. *et al*. (2003) An integrated functional genomics screening program reveals a role for BMP-9 in glucose homeostasis. *Nat. Biotechnol.* 21, 294–301
22 Schneider, G. and Fechner, U. (2005) Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discov.* 4, 649–663
23 ENCODE Project Consortium, (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* 306, 636–640
24 Imanishi, T. *et al*. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2, e162
25 Brent, M.R. (2005) Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Res.* 15, 1777–1786
26 The FANTOM Consortium, (2005) The transcriptional landscape of the Mamm. Genome. *Science* 309, 1559–1563
27 Konno, H. *et al*. (2001) Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library. *Genome Res.* 11, 281–289
28 Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* 29, 137–140
29 The MGC Project Team, (2004) The status, quality, and expansion of the NIH full-length cDNA project: the mammalian gene collection (MGC). *Genome Res.* 14, 2121–2127
30 Carninci, P. *et al*. (1998) Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc. Natl. Acad. Sci. U. S. A.* 95, 520–524
31 Carninci, P. *et al*. (2000) Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* 10, 1617–1630
32 Carninci, P. *et al*. (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* 13, 1273–1289
33 Carninci, P. *et al*. (1997) High efficiency selection of full-length cDNA by improved biotinylated cap trapper. *DNA Res.* 4, 61–66
34 Nielsen, H. *et al*. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.* 8, 581–599
35 Chen, Y. *et al*. (2005) SPD – a web-based secreted protein database. *Nucleic Acids Res.* 33, D169–D173
36 Klee, E.W. *et al*. (2004) Identifying secretomes in people, pufferfish and pigs. *Nucleic Acids Res.* 32, 1414–1421
37 Zhang, Z. and Henzel, W.J. (2004) Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Sci.* 13, 2819–2824
38 Hofmann, K. and Stoffel, W. (1993) TMBASE - a database of membrane spanning protein segments. *Biol. Chem. Hoppe-Seyler* 374, 166
39 Bateman, A. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.* 32, D138–D141
40 Revest, J.-M. *et al*. (2000) Fibroblast growth factor 9 secretion is mediated by a non-cleaved amino-terminal signal sequence. *J. Biol. Chem.* 275, 8083–8090
41 Gewurz, B.E. *et al*. (2002) US2, a human cytomegalovirus-encoded type I membrane protein, contains a non-cleavable amino-terminal signal peptide. *J. Biol. Chem.* 277, 11306–11313
42 Miyakawa, K. and Imamura, T. (2003) Secretion of FGF-16 requires an uncleaved bipartite signal sequence. *J. Biol. Chem.* 278, 35718–35724
43 Bendtsen, J.D. *et al*. (2004) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* 17, 349–356
44 Fukuhara, A. *et al*. (2005) Visfatin: a protein secreted by visceral fat that mimics the effects of insulin. *Science* 307, 426–430
45 Goldstein, A.L. *et al*. (2005) Thymosin beta4: actin-sequestering protein moonlights to repair injured tissues. *Trends Mol. Med.* 11, 421–429
46 Acton, T.B. *et al*. (2005) Robotic cloning and protein production platform of the northeast structural genomics consortium. *Methods Enzymol.* 394, 210–243
47 Vincentelli, R. *et al*. (2005) Automated expression and solubility screening of His-tagged proteins in 96-well format. *Anal. Biochem.* 346, 77–84
48 Wurm, F.M. (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* 22, 1393–1398
49 Walters, W.P. and Namchuk, M. (2003) Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* 2, 259–266
50 Iourgenko, V. *et al*. (2003) Identification of a family of cAMP response element-binding protein coactivators by genome-scale functional analysis in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12147–12152
51 Kunz-Schughart, L.A. *et al*. (2004) The use of 3-D cultures for high-throughput screening: the multicellular spheroid model. *J. Biomol. Screen.* 9, 273–285
52 Giuliano, K.A. *et al*. (1997) High-content screening: a new approach to easing key bottlenecks in the drug discovery process. *J. Biomol. Screen.* 2, 249–259
53 Comley, J. (2005) Label-free detection. *Drug Discovery World* 6, 63–74
54 Ciambrone, G.J. *et al*. (2004) Cellular dielectric spectroscopy: a powerful new approach to label-free cellular analysis. *J. Biomol. Screen.* 9, 467–480
55 Cunningham, B.T. *et al*. (2004) Label-free assays on the bind system. *J. Biomol. Screen.* 9, 481–490
56 Dove, A. (1999) Drug Screening–beyond the bottleneck. *Nat. Biotechnol.* 17, 859–863
57 Mullin, R. (2004) Drug Discovery: As high-throughput draws fire, researchers leverages science to put automation into perspective. *C&EN* 26, 23–32
58 Molloy, C. (2003) The industrial evolution of screening infrastructure. *Drug Discovery World* 4, 23–32
59 Ladd, B. (2000) Intuitive data analysis: the next generation. *Modern Drug Discovery* 3, 46–51
60 Sun, D. *et al*. (2005) Adopting a practical statistical approach for evaluating assay agreement in drug discovery. *J. Biomol. Screen.* 10, 508–516