



Data management solutions for protein therapeutic research and development

Jost Vielmetter, Jeff Tishler, Marie L. Ary, Peter Cheung and Richard Bishop

Protein therapeutics, including monoclonal antibodies, are a growing focus of drug discovery research organizations. High-throughput screening of large libraries of protein variants is therefore becoming increasingly important in R&D. As a result, there is a need to link large numbers of variant protein sequences with chemical and biological assay data. This integration will allow more efficient data mining and facilitate decision-making regarding hit identification, lead optimization and drug development. In this paper, we present an implementation in which a widely used small-molecule high-throughput screening data management system has been adapted to meet the unique needs of protein drug discovery and development.

► With the rapid advance of high-throughput screening (HTS) technology and the creation of large and diverse chemical libraries, the number and size of datasets created from routine screening campaigns has grown considerably in biopharmaceutical research and development. Managing and extracting valuable information from such datasets has therefore become a high priority for many pharmaceutical companies. Several data management platforms are now commercially available to enable small-molecule drug discovery by facilitating the organization, analysis and subsequent querying of high-volume HTS data. This technology allows efficient data mining and reporting and is widely used in the pharmaceutical industry to assist decision-making regarding hit identification, lead optimization and drug development.

Therapeutic proteins, including monoclonal antibodies, are gaining increased attention in drug discovery research organizations. Global sales in 2003 exceeded US\$30 billion compared with US\$12 billion in 2000. In this three year period, 30 new protein drugs were approved by regulatory agencies in the

US and Europe, accounting for more than a quarter of all new drug approvals [1]. With hundreds more proteins in clinical trials, these numbers will continue to grow, especially since biopharmaceuticals tend to move through clinical development more quickly than small-molecule drugs [2,3]. The expanding interest in protein therapeutics stems in part from their superb affinity and specificity for their clinical targets. In many cases, natural proteins serve as excellent 'lead' compounds. However, evolution has not shaped natural proteins to function as medicines; consequently, they often lack many of the characteristics desired in a therapeutic. Optimization through protein engineering can therefore be quite beneficial, and is now frequently employed in lead development and in the creation of second-generation products [4–7]. Engineering can be used to alter multiple protein properties including efficacy, specificity, solubility, stability, pharmacokinetics and immunogenicity. These can impact drug safety, potency, dosing frequency and route of administration; other considerations that can be affected include manufacturing cost and intellectual property.

Jost Vielmetter*

Marie L. Ary

Peter Cheung

Richard Bishop

Xencor,
111 W. Lemon Avenue,
Monrovia,
CA 91016,
USA

*e-mail: jvietmet@xencor.com

Jeff Tishler

ID Business Solutions Inc.,
1900 Powell Street,
Suite 1070,
Emeryville,
CA 94608,
USA

TABLE 1

Examples of companies generating protein libraries

Company	Focus	Technology/method of generating protein libraries
Abgenix	Therapeutic antibodies	High-throughput screening of human antibodies developed in transgenic mice, epitope mapping (XenoMouse™, XenoMax™)
Amgen	Therapeutic proteins and antibodies	Protein engineering, glycosylation engineering, antibody display
Applied Molecular Evolution/Eli Lilly	Therapeutic proteins and antibodies	Directed evolution (AMESystem™ technology)
Biovation	Therapeutic proteins and antibodies	T cell epitope identification and removal (site-directed mutagenesis) (Delimmunisation™ technology)
Cambridge Antibody Technology	Therapeutic antibodies	Antibody libraries, phage display, ribosome display
Centocor/Johnson and Johnson	Therapeutic proteins and antibodies	Monoclonal antibody technology, structure-based design
Diversa	Therapeutic proteins and antibodies, industrial enzymes	Directed evolution (Gene Site Saturation Mutagenesis™, Tunable Gene Reassembly™)
DNA2.0	Custom gene synthesis, expression optimization, protein engineering	Bioinformatics-based protein design (DeNovo Genes™ technology, custom gene synthesis)
Egea Biosciences/Johnson and Johnson	Therapeutic proteins and antibodies	High-throughput gene synthesis, rational protein design
FivePrime	Proteomics, target discovery	High-throughput cloning and expression of human cDNAs (ProScreen technology)
Genentech	Therapeutic proteins and antibodies	Bioinformatics/genomics (Secreted Protein Discovery Initiative, SPDI), phage display, structure-based design
Maxygen	Therapeutic proteins, vaccines	Directed evolution, PEGylation, glycosylation (Molecular Breeding™, Family Shuffling™)
Novozymes	Industrial enzymes	Directed evolution, combinatorial protein engineering/phage display, site-directed mutagenesis
Protein Design Laboratories	Humanized therapeutic antibodies	Rational structure-based design, antibody humanization
Roche	Therapeutic proteins and antibodies	High-throughput cloning and expression, site-directed mutagenesis
Xencor	Therapeutic proteins and antibodies	Rational structure-based design, rational PEGylation (Protein Design Automation®, ImmunoPDA™, XmaB™, and Rational PEGylation™ technologies)

Improved properties can be achieved by modifying the protein's primary structure through sequence changes, or by incorporating chemical or post-translational modifications such as PEGylation (the attachment of polyethylene glycol) or glycosylation. The oligomerization state of the protein can also be altered, or the protein can be fused to other entities such as albumin or the Fc region of antibodies [4,5,7]. Strategies for implementing these changes include site-directed mutagenesis, random mutagenesis, recombination and other directed evolution methods, as well as rational protein design and structure-based computational approaches [5,8–10].

Many pharmaceutical and biotechnology companies are now using these and other strategies to generate large protein libraries and are taking advantage of HTS technology to rapidly assay these libraries for desired properties (Table 1). This creates the need to link large numbers of variant protein sequences with assay data. Ideally, one must analyze the results efficiently and display them so that patterns, particularly sequence–activity relationships, can be identified. Although database systems with HTS data management and structure–activity relationship (SAR) analytical capabilities are commercially available,

they are specifically designed to handle small-molecule chemical and biological data; they do not support the basic bioinformatics functions that proteins require, such as sequence registration and comparison. There is no integrated and efficient data management system available that provides all the capabilities required for protein sequence data.

One solution is to adapt an existing data management system to handle high-volume protein sequence data. In this article, we describe an implementation in which a widely used small-molecule screening, data management, and SAR analysis system has been adapted to serve the unique needs of protein drug discovery and development.

Data management requirements for protein drug discovery

In addition to linking protein sequences with HTS data and providing powerful analysis tools, a useful data management system for proteins should support graphics as well as numeric and string-based data types, enable connections between various sources of data, and allow efficient querying and reporting. To be effective, the system must not only include the required infrastructure and

data-mining tools, but should also support the needs of scientists engaged in discovery research. Scientists need to know where the information is located and how it can be accessed and exploited to enhance productivity, but should not be required to have in-depth knowledge of how database systems operate. The user interface must be understandable and clearly reflect the experimental entities and laboratory processes involved. The system must readily incorporate in-house data, be easy to navigate, and provide visualization tools that can transform vast amounts of information into usable knowledge. It should perform all of these functions in a transparent and integrated fashion, and be flexible enough to meet the evolving needs of research.

Relational database management systems—handling HTS data

Relational database management systems, such as Oracle (Oracle Corp), meet many of these requirements by providing a unified framework that allows data storage and access via standard development tools such as Microsoft Visual Tools, Open Database Connectivity (ODBC), Java, Java Database Connectivity (JDBC) and Oracle Objects for OLE. Compatibility with standard development tools facilitates the creation of customized plug-in modules, enhancing flexibility and expandability. These systems not only supply a superb platform for applications such as query and analysis tools and provide connections to multiple data sources, but they are also scalable, easy to maintain, and can afford simultaneous multi-user access.

Relational database management systems provide an ideal platform for tools designed to manage HTS data, and are the systems most frequently employed for this purpose. Examples of some commonly used HTS data management packages include Accord™ Enterprise Informatics suite (Accelrys), BioAssay (CambridgeSoft), ActivityBase (IDBS), Assay Explorer (MDL), and CBIS (ChemInnovation Software) [11]. All of these applications use Oracle databases to handle the high volumes of data generated by HTS. They employ custom scripts to access proprietary databases, facilitate storage or import data (such as chemical structure information) into Oracle tables. These products have enjoyed widespread use in the pharmaceutical industry. However, they do not address the data management issues associated with screening large protein libraries.

Protein-specific requirements

A common strategy in small-molecule lead optimization is to screen libraries of compounds whose members are related to each other on the basis of their chemical or structural similarity to a given lead candidate, with the goal of identifying pharmacophores and ultimately compounds with improved therapeutic properties. Similarly, in engineering therapeutic proteins, libraries are typically built as sequence variations derived from a promising lead protein. As mentioned above, the goal might be to obtain

protein variants with improved efficacy, specificity or pharmacokinetics, or decreased toxicity or immunogenicity. The typical process flow in optimizing a lead protein is protein library design and construction, followed by screening, analysis and lead identification.

Registering and comparing sequences, linking mutations to assay data

Efficient task management in a protein-engineering environment requires that protein sequences be registered in the database with unique IDs as they are designed and produced in the laboratory. Variations from the lead protein must be calculated and stored in the database, and this data must be linked to assay results. Sequence comparison allows differences to be identified, so that data can be described in terms of mutations, deletions, insertions or other modifications such as PEGylation and glycosylation, and correlated with biological activities and other assay data. Assay results must also be registered and processed, so that SAR and other analyses can be performed and useful queries made.

Handling string-based data

The data management challenges faced in accomplishing these tasks are, in many ways, analogous to those encountered in managing data from small-molecule libraries and assays. However, differences in the type of data and the way it is stored require that protein sequence information be handled somewhat differently. For example, small-molecule libraries consist of chemical inventories and structures and are usually stored in proprietary databases such as ISIS (MDL). On the other hand, protein sequences and the DNA sequences that encode them are string-based data objects and are typically stored in flat files. Sequence-based data operations, such as sequence registration and sequence comparisons, are usually only provided for by software platforms centered on bioinformatics and molecular biology applications such as Vector NTI® (Invitrogen Corp.), DS SeqStore and DS AtlasStore (Accelrys). Conversely, bioinformatics platforms typically lack assay result processing and storage functions. Consequently, none of the data management systems currently in use provide all the functions required to efficiently manage the screening of large protein libraries.

Adapting a data management system for protein variant data

Our solution to this problem was to start with a currently existing data management system and adapt it to handle protein sequence data. The relational systems mentioned above have many desirable features, including SAR analysis and the ability to manage the high volumes of data generated by HTS. We chose ActivityBase for these reasons, but also because it has an excellent track record in the pharmaceutical industry [12], and custom modules can easily be constructed and integrated into its platform.

BOX 1

Data structure elements and relationships for a typical HTS data management system

Objects: Entities measured or tested in an experiment (small molecules, compounds). Objects have unique IDs and are often registered into the database (e.g. Oracle) in large numbers; they can be grouped together as lists, libraries or plates.

Conditions: Experimental parameters or variables (e.g. type or concentration of reagent). Conditions are entered into the database and linked with experimental Results.

Raw Data: Numbers, character strings, or other data generated directly from an assay or experiment; typically stored as tables in files.

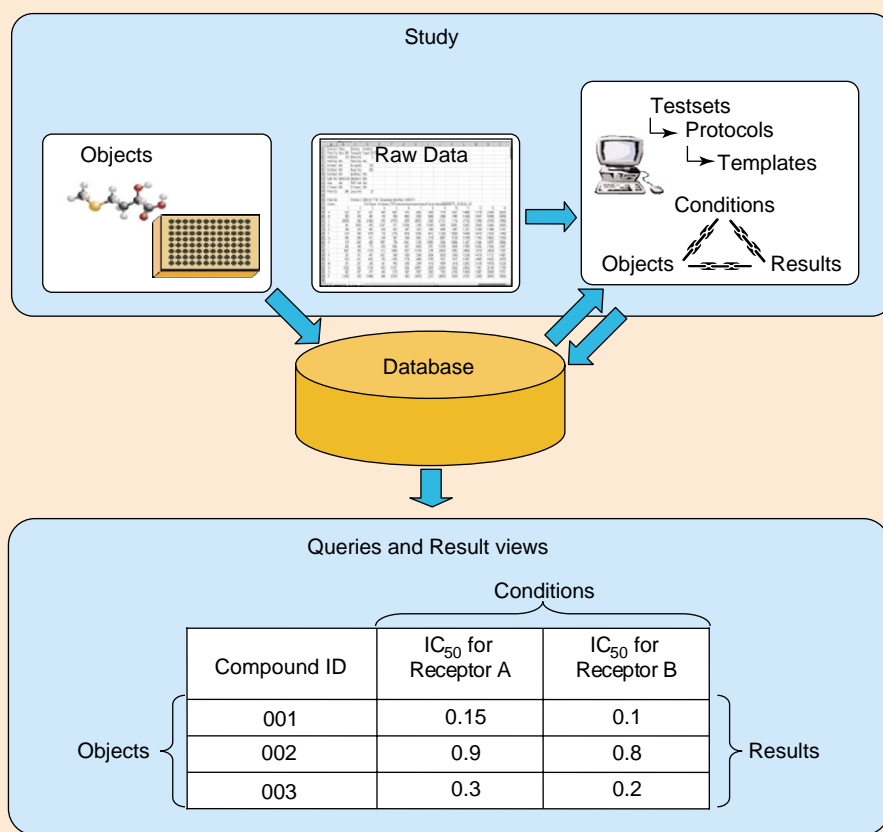
Template: Custom Excel spreadsheet or other template with input filters and data processing instructions that perform calculations and convert Raw Data into more useful Results. Can be extended with a curve-fitting engine such as XIFit (IDBS) to generate graphs.

Results: Numbers, strings of characters, graphs or pictures derived from experimental Raw Data. Results are automatically stored in the database and linked to their corresponding Objects, Conditions and Testsets/Test Occasions.

Protocol: Maps the associations between Results and their experimental Conditions that will be used within Templates; links Templates to Testsets or Test Occasions.

Testset, Test Occasion: Defines which Protocol and Template is to be applied to a set of Objects; launches the Template and when execution is complete, uploads the Results into the database with links to Objects and Conditions.

Study: The group of Objects, Testsets, Test Occasions, Protocols and Templates associated with a given research theme.



Drug Discovery Today

FIGURE 1

Data management process flow. Raw Data are parsed and uploaded to the database through Testsets, Protocols, and Templates, which process the data and link together Objects, Conditions, and Results. Objects are registered into the database separately using registration tools. Specialized modules are used to query the database and view Objects with their associated Results under various Conditions, facilitating data mining (e.g. structure-activity relationship analysis).

Data structure and data mining tools of a HTS data management system

One of our goals was to enable simultaneous querying of sequence variations and assay results, thus assisting the identification of protein sequence-activity relationships. ActivityBase provides flexible data structures that can be chosen to reflect the experimental entities and laboratory processes employed. One of the first steps in adapting ActivityBase for proteins was therefore to determine the appropriate field choices, object dependencies, and assignments required so that the data structure would support our needs.

Definitions of the different data structure elements provided by ActivityBase and how they relate to each other are shown in Box 1 and Figure 1. Objects are the tangible items that are measured or tested, Conditions are the variables in the experiment, and the experimental measurements themselves are the Raw Data. The Template

contains input filters and data processing instructions that perform calculations and convert the Raw Data to more useful Results. Curve-fitting engines such as XIFit (IDBS) can be incorporated to create graphs. The Results are automatically stored in the database and linked to tables containing their associated Conditions according to a Protocol, while Testsets or Test Occasions define the association between Objects and their Results and Conditions. Finally, the Objects, Test Occasions, Testsets, Protocols and Templates associated with a particular theme or research question are all grouped together under a so-called Study. For example, one might organize a Study around a particular drug family or class of compounds, where the Objects being tested are structurally related to each other.

The Objects and Testsets or Test Occasions can be chosen so that they virtually mirror the experimental objects and assays within a screening project, such as the compounds

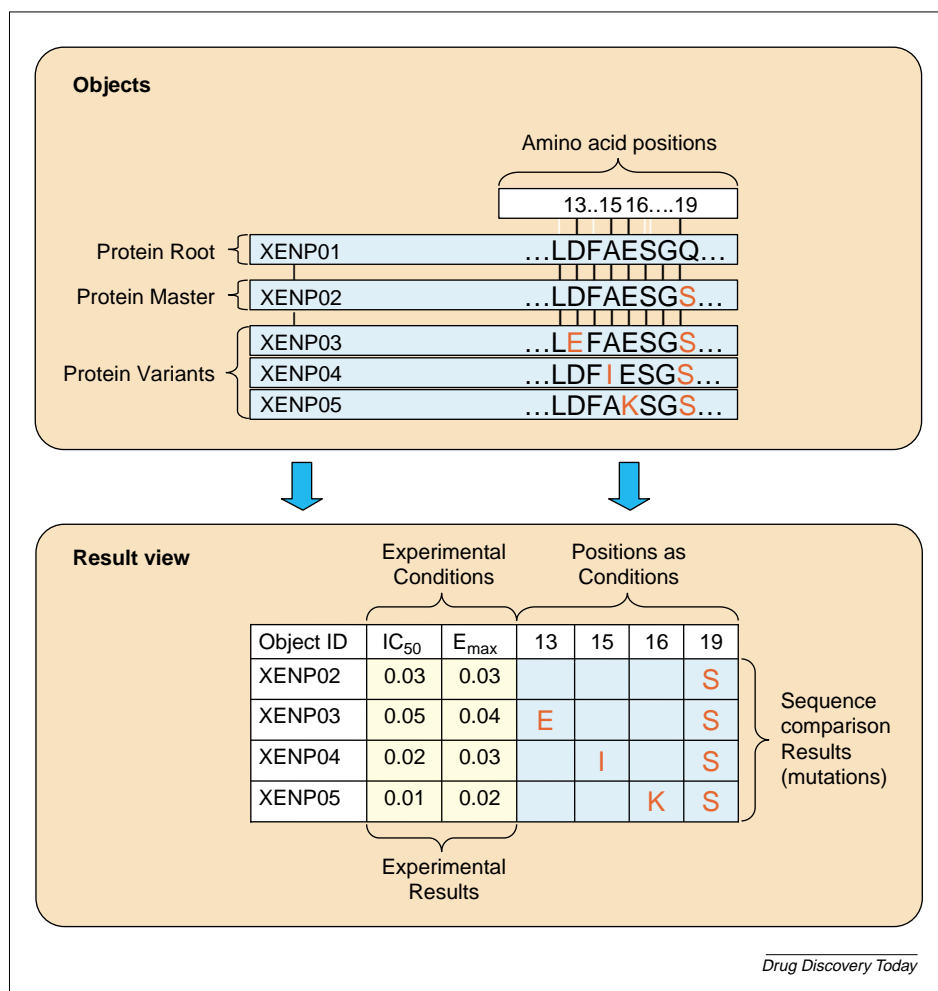
**FIGURE 1**

Diagram illustrating how protein sequence comparison Results (mutations) are mapped to amino acid positions (Conditions) and correlated with experimental Results. The original Protein Root sequence serves as a reference for the amino acid positions in the Protein Master, which in turn serves as a template for constructing a protein variant library. The positions at which amino acid identities have changed relative to the Protein Root are uploaded into the database as Conditions; the amino acid changes themselves are entered as Results and are associated with their corresponding Object IDs. Associated bioassay Results can now be viewed with the corresponding structural information (amino acid mutations) and sequence-activity relationships can be determined.

to be screened, the assay plates containing subsets of compounds, and the assays performed with these compounds. A Study can similarly be designed to reflect laboratory activities. All these virtual entities are visually presented in a customized graphical user interface called a workbench and represent a subset of the table structure in the relational database (Oracle).

This type of data structure is easy to understand and navigate, and provides an elegant means to enable comparative querying and display of results under different experimental conditions. Most HTS data management systems provide query and visualization tools for these purposes. For example, IDBS provides SARgen to retrieve assay results, while SARview allows nested sorting, filtering and parsing of the data to organize and display it in a useful format. Proper use of these tools can facilitate data mining, reporting and decision-making.

Data structure assignments and dependencies for protein engineering

In many protein-engineering projects, protein libraries are produced by altering amino acid sequences or by introducing chemical modifications to a parent protein. Accordingly, in adapting ActivityBase for proteins, we chose a data structure in which protein variant members of a protein library relate to a single parent protein sequence called the Protein Root. In the database, this is implemented by having the Object record of each protein variant reference the Object ID of its Protein Root. Protein variant libraries are thus registered referencing a unique Protein Root Object ID. A particular variant may show improvement over the Protein Root and be selected to serve as the parent sequence of a second-generation protein variant library; this new parent is called a Protein Master. The members of the second-generation library reference their Protein Master as well as the original Protein Root. Several protein variant libraries can relate to each other by referencing a single Protein Root. All the variants, including their Protein Masters and the Protein Root, are grouped together under one Study, which also contains the Test Occasions and Testsets carried out on this set of Objects.

Amino acid variations are identified by comparing variant sequences with the Protein Root; these changes are registered in the database as Results. The positions where the changes have occurred are linked to the Results and registered as Conditions. These data structure assignments allow simultaneous querying of

sequence variants and experimental Results, thus assisting the identification of protein sequence-activity relationships (see Figure 1).

Custom modules for sequence-based data operations

Because ActivityBase does not provide functions to handle sequence-based data operations, we constructed a set of custom modules for this purpose. Our solution included sequence registration and sequence comparison modules, which were developed as external programs and integrated into the system, as well as customized Test Occasions built inside the ActivityBase platform. We implemented all these custom procedures using standard Visual Basic programming tools. Registering protein variants with unique IDs as they are designed and produced in the laboratory can be a challenge, especially when large libraries containing several thousand variants are involved. Our


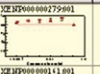

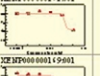

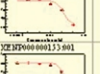



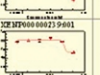

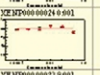

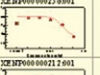

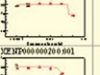


Amino acid positions											
Object ID	Batch	IC50 receptor-I	Dose response receptor-I	IC50 receptor-II	Dose response receptor-II	31	32	33	86	146	
XENP000000279	001	0.0012		ND					R		
XENP000000161	001	0.0012		0.1787		D					
XENP000000169	001	0.0015		0.1383		E					
XENP000000153	001	0.0018		0.1439		I					
XENP000000185	001	0.0026		0.1434			E				
XENP000000239	001	0.0042		0.1651					T		
XENP000000240	001	0.0047		ND			W		T		
XENP000000238	001	0.0048		0.1225			W				
XENP000000212	001	0.0056		0.3669						R	
XENP000000200	001	0.0060		0.1513				E			

FIGURE 2

Result view obtained in a study to identify receptor-specific protein variants. Variants were screened for binding to receptor-I and receptor-II and our protein-customized data management system was used to register, process and link the data. SARgen was employed to query the database for IC₅₀ values and corresponding dose–response curves. SARview was then used to visualize and filter the data. To identify sequences with specific binding to receptor-I, the output was restricted to only show records with receptor-I IC₅₀ values less than 0.007 μ M and receptor-II IC₅₀ values greater than 0.1 μ M. Data were sorted to list records in order of increasing receptor-I IC₅₀ values; individual Xencor protein variants are listed under Object ID (e.g. XENP000000279); ND indicates IC₅₀ values could not be determined because receptor binding was not detectable. Reporting the data in this way allowed sequence–activity relationships to be identified: amino acid changes at positions 31, 32, 33, 86 and 146 result in significant loss of binding to receptor-II, whereas binding to receptor-I is retained or only moderately affected. (Wild-type protein binds to both receptors.)

sequence registration module provides scripts that perform these tasks. We also included features such as data input validation to minimize errors and ensure data consistency. Researchers can label protein samples with unique IDs as soon as they are produced in the laboratory; these IDs reference variant sequences that were registered in the database when the protein library was designed. The unique IDs are maintained during all phases of research, enabling consistent identification of samples throughout protein library production, HTS, follow-up assays, and preclinical testing.

We built a sequence comparison module that compares variant sequences with their Protein Root and creates a file listing the variations and the positions where these changes occur. The comparison algorithm scans the variant and Root sequences and notes substitutions with native or non-native amino acids; insertions, deletions and chemical modifications are identified with previously encoded symbols. For example, deletions are encoded by

replacing the letter for each deleted amino acid with the symbol '#'; insertions are encoded by delimiting the inserted sequence with curly braces (e.g. '{PRTN}'), and modifications use square brackets as delimiters (e.g. '[PEG20]'). Our customized Test Occasion uploads the file generated by the sequence comparison module, enters the sequence variations into the database as Results, and registers the changed positions as Conditions. Annotations are also entered into the Object record of each protein variant. The annotations include the start and end positions of the protein variant sequence, and for each changed position, the position of change, the wild-type amino acid, and the new amino acid or modification.

Anonymous objects

Our adaptation also allows protein variants to be registered without knowing their sequence (as anonymous Objects); the anonymous Objects can be correlated with bioassay data, and only those identified as promising can be sent for sequencing. This feature is useful when large combinatorial [13] or randomly mutated libraries are screened so that only hit sequences need be registered. This strategy can save time and resources, and may be beneficial when information about inactive sequences is not required.

Case study: receptor-specific variants

One of our research objectives was to identify protein variants that show receptor binding specificity. Specificity can be studied using competitive receptor binding assays; a dose–response curve is obtained and the dose producing 50% inhibition (IC₅₀) is calculated for each receptor. Starting from the wild-type protein as the lead, a library of variants was designed and constructed, binding assays were performed, and IC₅₀ values calculated. Our protein-customized adaptation of ActivityBase was used to register variant sequences in the database, compare them to the Protein Root (wild-type sequence) to identify mutations, and link these Results to receptor binding data. The raw assay data were processed to generate dose–response curves and IC₅₀ values. Query and visualization tools (SARgen and SARview) were then used to retrieve the data and present it in a useful format. Figure 2 shows tables generated by SARview: amino acid changes of individual variants are presented next to their receptor binding dose–response curves and IC₅₀ values. In this case, the data were sorted and filtered to view protein variants with selective binding for receptor-I compared

with receptor-II. Presenting the data in this way allowed us to correlate structure (sequence) with activity and assisted us in making decisions regarding future designs.

Challenges and future applications

The development of data-management systems for high-throughput protein engineering is just beginning to evolve. Though one of the first to emerge, our adaptation takes advantage of the benefits of relational database systems and thus provides many valuable functions. Additional features that would be useful include the ability to make correlations between the three-dimensional structures of protein variants (not just their sequences) and biological properties. This certainly seems feasible as many programs are available that can model protein structure from a given sequence, particularly if the coordinates of the starting sequence are known. Integration of a structure viewer would be required, but if done efficiently, this should accelerate access to three-dimensional information and thus facilitate future cycles of structure-based protein design.

Adaptations similar to the one described here could be

used for the management of other string-based datasets and may prove beneficial in applications such as those employing genomics or proteomics data. In these cases, the addition of a feature providing the ability to directly cross-reference protein or DNA sequences with public data sources such as GenBank, UniProt, the Protein Data Bank or dbSNP would greatly facilitate data mining.

Conclusions

The high-throughput screening of large protein libraries poses unique data management challenges. Our adaptation of ActivityBase provides the first data management system to integrate high-volume protein sequence data, biological activities and SAR analysis to allow efficient data mining and querying. This application will significantly speed the development of protein therapeutics.

Acknowledgements

We would like to thank John Desjarlais, David Szymkowski, David Vielmetter and Bassil Dahiyat for helpful editorial advice.

References

- Walsh, G. (2003) Biopharmaceutical benchmarks-2003. *Nat. Biotechnol.* 21, 865–870
- Reichert, J.M. (2003) Trends in development and approval times for new therapeutics in the United States. *Nat. Rev. Drug Discov.* 2, 695–702
- Bibby, K. *et al.* (2003) Biopharmaceuticals—Moving to centre stage. *Bioprocess North American Biotechnology Industry and Supplier's Guide*, 3–11
- Lazar, G.A. *et al.* (2003) Designing proteins for therapeutic applications. *Curr. Opin. Struct. Biol.* 13, 513–518
- Marshall, S.A. *et al.* (2003) Rational design and engineering of therapeutic proteins. *Drug Discov. Today* 8, 212–221
- Steed, P.M. *et al.* (2003) Inactivation of TNF signaling by rationally designed dominant-negative TNF variants. *Science* 301, 1895–1898
- Walsh, G. (2004) Second-generation biopharmaceuticals. *Eur. J. Pharm. Biopharm.* 58, 185–196
- Graddis, T.J. *et al.* (2002) Designing proteins that work using recombinant technologies. *Curr. Pharm. Biotechnol.* 3, 285–297
- Brekke, O.H. and Loset, G.A. (2003) New technologies in therapeutic antibody development. *Curr. Opin. Pharmacol.* 3, 544–550
- Vasserot, A.P. *et al.* (2003) Optimization of protein therapeutics by directed evolution. *Drug Discov. Today* 8, 118–126
- Vaschetto, M. *et al.* (2003) Enabling high-throughput discovery. *Curr. Opin. Drug Discov. Devel.* 6, 377–383
- Zaborowski, M. (2004) Powering pharmaceutical productivity. *IDBS In Silico Summer* (1), 9–12
- Hayes, R.J. *et al.* (2002) Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15926–15931