Visual exploration of HTS databases: bridging the gap between chemistry and biology

Christopher Ahlberg

As pharmaceutical corporations are under pressure to shorten research and development (R&D) cycles for new drugs, novel technologies are being deployed, producing a data explosion in the research environment. Researchers can then become the bottleneck in the R&D cycle when they are unable to analyze data quickly enough. To deliver on the promise of these high-throughput technologies, pharmaceutical R&D must introduce novel decision support systems to support the mantra of R&D decision support: access, analyze and publish.

Pharmaceutical corporations are facing a fierce competitive environment and an increasingly rapid race to find novel drugs. Sustaining traditional growth levels and hence, traditional stock market valuations, will require significant improvements in product development[1]. This implies that research and development (R&D) throughput must improve with shorter leadtime and that the quality of drug candidates throughout the R&D process must improve, with the poorer candidates failing earlier.

Addressing these issues, pharmaceutical R&D has invested heavily both in instrumentation, enabling the quantity of measurements on target and lead compounds to increase dramatically, and in databases and inventories to store these measurements and annotations. Technologies typically invested in are high-throughput screening (HTS) instruments, DNA-sequence databases and combinatorial libraries of chemical compounds.

These investments have now shifted the bottleneck in R&D from being unable to measure the data quickly enough to being unable to analyze it quickly enough. The number of targets, compounds, assays and data generated from these investments has created a combinatorial explosion, where traditional methods and tools for accessing, analyzing and distributing the data and insight are no longer viable. In addition, demands for more rapid processes require teams to be larger and to comprise individuals across multiple disciplines. These cross-functional aspects create novel challenges requiring teams of chemists, biologists and pharmacologists to combine expertise and analyze results that bridge multiple domains, thereby requiring cross-functional decision support.

HTS is probably the greatest technological driving force in creating this challenge for pharmaceutical R&D. HTS technology is rapidly producing large databases, which store assay results for millions of compounds. These compounds must also be quality-assured, analyzed and correlated, not only against other assays, but also against physical properties and analytical, toxicological and structural data. Other technologies challenging the pharmaceutical R&D include gene expression and combinatorial chemistry.

Pharmaceutical corporations have now put HTS databases in place, but have still not adequately invested in the technology to support researchers in technical decision-support activities such as:

Christopher Ahlberg, Spotfire Inc., One Broadway, Cambridge, MA 02142, USA. tel: +1 617 621 0340, fax: +1 617 621 0381; e-mail: ahlberg@spotfire.com; Web: http://www.spotfire.com

1359-6446/99/$ – see front matter © Elsevier Science. All rights reserved. PII: S1359-6446(99)01373-2

- Accessing data existing in complex R&D databases
- Analysis of data
- Publishing and disseminating analysis results to colleagues

Drug discovery portfolio management

The basis for solid decisions is solid data combined with solid methods for analysis. Research organizations make hundreds of decisions every day but will not know if they are correct until much later. This severely delayed feedback loop requires R&D departments to 'bet' on several candidates that might be able to make it through the research 'funnel', and then watch these candidates carefully, much in the manner that financial analysts watch their portfolios.

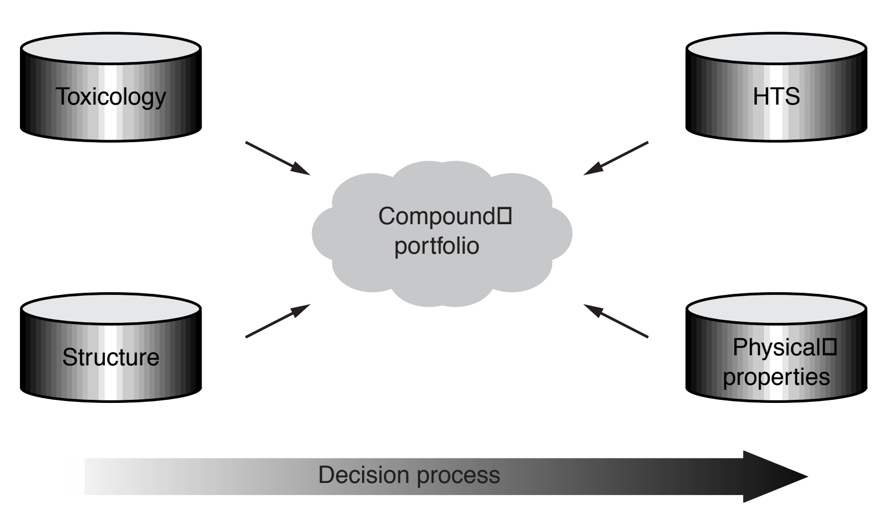


Figure 1. Database infrastructure and decision support systems must enable a rapid decision process enabling scientists to quickly identify strong candidates whilst discarding the weak ones.

The portfolio itself is no better than the data supporting the analysis of the portfolio, which has then driven novel instrumentation technology for rapidly measuring the properties of large sets of candidates. HTS allows pharmaceutical companies to run large libraries of compounds (in the magnitude of millions) through relatively few assays (in the magnitude of tens). Attractive compounds might be those with high efficiency in 1 out of 40 assays, but with little activity in the other 39 assays (i.e. this is a high-dimensional analysis problem).

The challenge is made more difficult by the fact that the assays are typically not binary, i.e. the range of possible outcomes in each dimension is large. Additionally, decisions will be made, not only on the basis of a set of biological assays, but also on factors such as physical properties, efficacy, toxicity and chemical structures. As biological assay results typically reside in relational databases that have been developed in-house, whilst the other information is usually stored in separate databases, the scientists and committees that make decisions on which compounds make it through the selection process need to combine a large number of decision support materials. Bringing these materials together is often very complex and typically requires the involvement of specialized Information Technology (IT) personnel. However, the existing IT infrastructure is not designed to support decisions on compounds, but rather to efficiently store experimental results from individual activities.

For a cross-functional team to successfully work on a portfolio of compounds, they must be continuously enabled to pull together decision support materials which can then be discussed, analyzed, disseminated and propagated throughout the process (Fig. 1).

Lead portfolio decisions: Access→Analyze→Publish

Managing the portfolios of candidate compounds involves three main IT-related aspects:

- Accessing relevant data for decision-making, which is stored both in-house and in externally acquired databases
- Analyzing the accessed data to make informed decisions, an activity involving numerous people and varied disciplines
- Publishing analysis results for distribution to colleagues and committees taking stop–go decisions, so that results can be quickly used as the basis for a novel analysis.

Accessing compound decision support materials

The ability to access databases is well known to be a complex task[2], as can formulating queries into single databases, even for the more sophisticated users. Those involved in pharmaceutical research are facing situations where gathering decision support materials for stop–go decisions requires access to 4–5 disparate databases and systems, a task which is obviously even more complex.

Tasks might include aggregating assay, physical property, toxicology and structural data, which, in addition to being located in physically and disparate databases, are structured in a great variety of different ways. For example, data on compounds might be stored in relational databases (e.g. Oracle) but at different levels of aggregation, which therefore significantly complicates the task of formulating information across two databases.

The set of compounds, or the portfolio, is central to executing queries or searches on leads in the discovery phase, and such queries focus on gathering materials that support decisions made concerning these compounds. Conversely, typical databases and legacy systems (old computer systems and mainframe computers) in R&D are not structured to allow users to access data on compound portfolios.

In recent years, services on the Internet (e.g. AltaVista, Yahoo and Amazon) have demonstrated how applications can be integrated with complex databases and systems. This has enabled end-users to carry out sophisticated tasks spanning complex and vast databases and systems in a single user interface, the web browser (Fig. 2). Hence, researchers gathering decision support materials can now perform complex tasks as simply as searching millions of web pages using Yahoo, or searching and buying books from the millions of books at Amazon. The technology enabling this process includes:

- Data integration technology that provides unified interfaces into numerous databases
- A semblance of the world, reflecting underlying data to users in terms of meaningful actions, objects and context
- Intranet technology that provides a unified user interface and data communications protocol for exposing those interfaces.

End-user applications for accessing databases should provide researchers with seamless interfaces enabling tasks such as:

- The retrieval of physical property data and chemical structures concerning this portfolio of compounds
- The retrieval of compounds from these experiments, grouped by the biological assay, as well as the chemical structures available for those compounds
- The retrieval of all compounds, including substructures with physical property and toxicology data on those compounds
- The retrieval of assay data and structures on selected portfolios, grouping data by portfolios and aggregate assays based on average activity levels and compound identification within the portfolio.

Analyzing compound portfolios

As discussed previously, the data supporting portfolio decisions will come from multiple sources. Facilitating analysis across these multiple sources requires researchers to deal simultaneously with several variables when performing tasks of trending (analyzing data to find a trend), outlier detection (finding anomalies in data points) and filtering. The average human being is not adept at reasoning when more than three or four variables are present at any one time, yet successful approaches must support scientists in reasoning significantly more variables. Therefore, in an attempt to reduce the number of variables, computers need to be able to represent data in such a way as to enable multidimensional reasoning, as well as to compute relationships and present the results of these computations. The particular situation discussed here requires additional thought, as the data is presented in a number of formats including as (simply) numbers and categories, as complex full-text documents (e.g. patent information) and as chemical structures, and therefore appropriate methods will also need to span across these types of data. Below, a series of

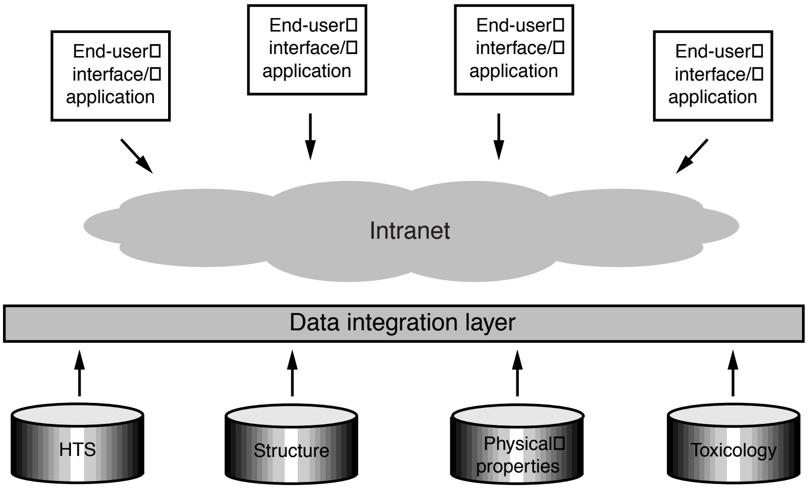

End-user interface/application
End-user interface/application
End-user interface/application
End-user interface/application
Intranet
Data integration layer
HTS
Structure
Physical properties
Toxicology

Figure 2. *Integrating disparate databases allows end-users to effectively access decision support materials through the Intranet in a variety of applications.*

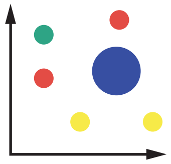


Figure 3. *The visual properties position (x, y), color and size are used to encode data in a simple scatterplot.*

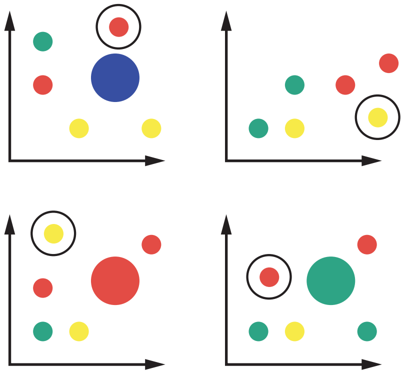


Figure 4. *Visual correlation through the brushing-and-linking technique, where moving a marker over a compound in one visualization highlights it in the other three.*

techniques is described, demonstrating the effective management of portfolios, based on interactive visual displays of data.

A simple display might, for example, encode two dimensions by their position in a scatterplot, a third dimension by color and a fourth by size. This type of display is very effective for trending as well as for outlier detection, while also forming a strong basis for performing filtering operations (Fig. 3). Shape, pattern, rotation and labels etc., might encode further dimensions. However, it is apparent that three to four dimensions in a single display is close to the human cognitive limit. Therefore, two alternative methods of data display are explored below: linking and interactive filtering of multiple visualizations.

Brushing-and-linking. In multiple visualizations, the encircled data point in Fig. 4 is highlighted across all four visualizations. Users can select a compound in one of the visualizations, which will in turn highlight that compound in the other visualizations. This approach potentially enables the rapid scanning of a single compound for 12 dimensions (four per visualization), where the four visualizations might depict biological assays, physical properties, toxicology and development data. This allows scientists to view multiple sources of background data, correlate them without difficulty and improve communication between, for example, chemists and biologists. This technique is referred to as brushing-and-linking[3].

Note that this technique does not enable the researcher to reason over a large number of compounds across 12 dimensions ‘in his/her head’, but rather to externalize the problem to a computer screen. The task is supported by providing rapid access into parts of the problem space and making the problem more of a perceptual (vision-oriented), rather than a cognitive problem.

Dynamic queries. Another method is to link visualization with interactive filters for each dimension that is available in the decision support materials, enabling dynamic (continuously changing) queries into the data (Fig. 5). Each of the sliders on the right of each diagram can be used to filter out only those compounds fulfilling certain criteria. The visualizations are updated in real time (<150 ms), again providing an effective manner to perform trending, outlier detection and filtering[4]. Here, a biologist’s view into the data (e.g. displaying results of four biological assays) can be effectively correlated with the chemist’s view into the data (e.g. physical property data) by adjusting the sliders representing the physical property variables. Of

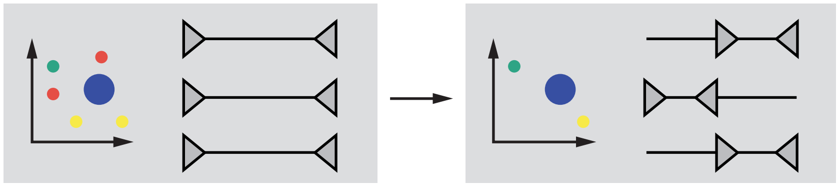


Figure 5. *Visual correlation through the dynamic queries technique, where each of the dimensions of a chemical compound, for example, is displayed visually to the left, with the action of filtering these compounds being represented on the right. Each dimension is represented by a slider that can be adjusted to instantly filter out interesting compounds on the left.*

course, this technique can be very effectively correlated with the brushing-and-linking technique described above.

Underlying both the approaches described above is the technique for substituting for the lack of human ability to reason in multidimensional problems by using visual display and interactivity. By providing interactive control over the displays with immediate feedback, scientists can visually explore the problem space. In a step-by-step fashion, aspects of the space can be examined, as opposed to the traditional approach of formulating hypotheses up-front and then testing them statistically and experimentally, even when the problem is ill-defined. Managing portfolios across high-dimensional databases is a very ill-defined problem and in many instances, finding the right question is as challenging, if not more challenging, as finding the answer.

Linking to related information. The display types used above, as well as similar displays such as histograms, bar charts and pie charts, can all be used for table-oriented information consisting of numbers and text. However, for information such as chemical structure and full-text, they are less useful.

Here, the brushing-and-linking approach described above is again effective. Tracking a cursor over the compounds in the visualization can be used as a technique not only to correlate data points between multiple displays of data (as done above) but also to trigger the display of alternative data, such as chemical structure. Figure 6 illustrates how the selection of a compound in the visualization triggers display of a chemical structure. Having identified a chemical structure of interest, an equivalent operation of brushing-and-linking in the chemical view will enable the division of the visualization into subsets (Fig. 7). Identifying a substructure can be a query operation, as depicted below, providing chemists and biologists with the ability to:

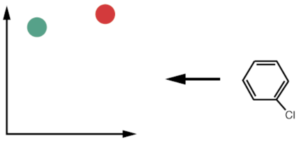


Figure 7. *A substructure search, specified on the left, is used as an alternative filtering method, allowing chemists and biologists to mix query criteria across domains of chemistry and biology.*

- Gain insight into biological assays through visualizations
- Identify compounds of interest and interactively access related chemical structures
- Perform substructure searches
- Highlight those compounds fulfilling the substructure search in visualizations.

In summary, interactive visualizations and dynamic queries effectively combine multiple dimensions and data types, enabling chemists and biologists to make stop–go decisions in a collaborative fashion, based on a single display of data. The interactive visualization system, Spotfire Pro, can be used to display more than ten dimensions of multiple databases whilst simultaneously linking chemical structure, in the manner described above (Fig. 8).

Intranet-based compound portfolio reports

After researchers have carried out the analysis, results must be efficiently shared with colleagues. Project teams are growing in size, and, rapid selection of compounds together with increasing the quality of compounds passing through the process is becoming increasingly important. Traditionally, research reports that reflect the results of analyses carried out by one or more researchers are printed on paper. Paper obviously has many advantages, including legibility and ease of adding annotations. However, distinct disadvantages of using paper include:

- Figures and tables representing analysis results cannot be expanded into the underlying analysis. Even if the underlying data were provided, it would still be in printed

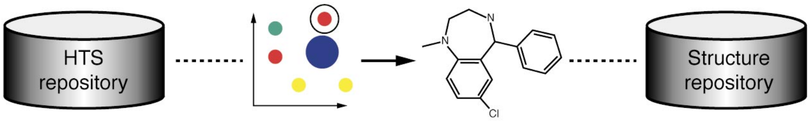
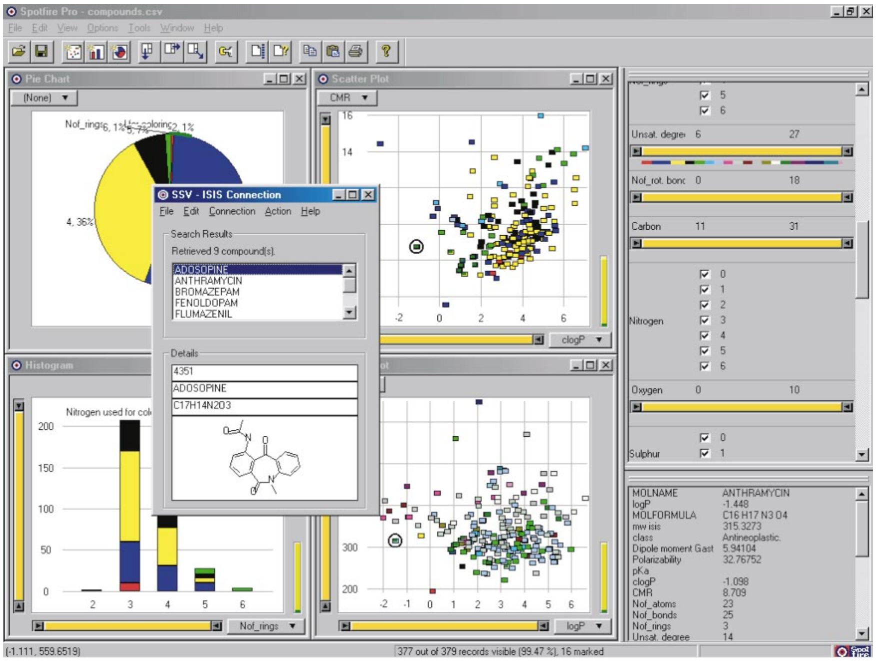


Figure 6. *The brushing-and-linking technique is used to link structured and unstructured data, e.g. biological activity in the scatterplot on the left and chemical structure in the display on the right. Rapidly moving the cursor over compounds in the scatterplot will trigger rapid display of chemical structure, allowing chemists to read 'structures as music' in the context of biological data.*



MOLNAME ANTHRAMYCIN
logP -1.448
MOLFORMULA C16 H17 N3 O4
mw isis 315.3273
class Antineoplastic
Dipole moment Gast 5.94104
Polarizability 32.76752
pKa
clogP -1.098
CMR 8.709
Nof_atoms 23
Nof_bonds 25
Nof_rings 3
Unsat. degree 14

Figure 8. *Spotfire Pro, an interactive visualization system where disparate data sources can be brought together visually, with query capabilities readily available on the left. The chemical structure of the compound currently being pointed at is displayed in the separate window, and the biological data on the compound is displayed in the bottom righthand corner.*

form and be virtually impossible for a colleague to pick up from where someone else left the work without going back to the original author to ask for the data.

- Paper reports quickly ‘age’ and are logistically expensive to keep updated as new data is made available.
- Paper reports can be expensive to distribute, especially in virtual project teams in major pharmaceutical corporations where teams can spread across several countries and continents.
- Paper reports are hard to locate.

An attractive alternative approach would be to make Intranet-based reports the primary information carrier (Fig. 9), as they can:

- Link back to the underlying data giving information on the tools and the state of the tools that were used to support the analysis, thereby enabling a researcher to continue the research of a colleague from where it had been previously stopped.
- Be identified by searching with a search-engine, which can identify a report based on any keyword, author or annotation used in the report, as well as on the underlying data supporting the report.
- Easily be distributed over e-mail.
- Serve as an effective carrier of a saved, sharable database query, representing both the underlying data and the analysis work done on the data.

Going back to the process of Access→Analysis→Publish, the research report representing the query into the database is, of course, the most effective database query from an end-user perspective, as well as an effective template that can be applied to new data. The interactive research report becomes the ultimate carrier of knowledge, and the best practice for data analysis and sharing in the research process.

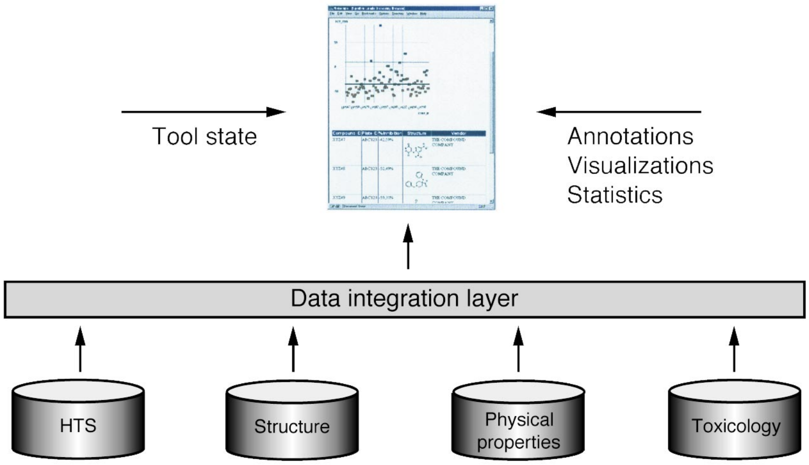


Figure 9. *Example of a web-based interactive research report, where beyond the actual report, pointers to data sources, annotations, visualizations, statistics, as well as the current status of the system used to define the report, are brought together.*

Conclusion

As pharmaceutical corporations are focusing efforts on maintaining traditional growth levels, managing the R&D process through effective portfolio management becomes crucial. Supporting rapid evaluation and selection of compounds with increased quality requires large cross-functional teams where chemists and biologists jointly bring successful compounds forward. Cross-functional teams require cross-functional decision support where both chemists and biologists can:

- Easily gather effective decision-support materials from chemical and biological databases
- Analyze across multiple disparate data sources including assays, toxicological, structural and analytical data
- Report results linking back to the original data sources.

To enable such decision support systems, pharmaceutical research must build IT systems focused on cross-functional interaction and integration. This in turn requires investments in database-integration technology, data-analysis software and Intranet-based systems for interactive report generation and distribution. Access→Analysis→Publish becomes the mantra driving efficient decision support in pharmaceutical R&D.

REFERENCES

1 Rosofsky, M. and Banerjee, P.K. (1996) *SCRIP*

2 Shneiderman, B. (1992) *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (2nd edn), Addison–Wesley Publishing Co.

3 Cleveland, W. (1993) *Visualizing Data*, Hobart Press

4 Ahlberg, C., Williamson, C. and Shneiderman, B. (1992) Dynamic Queries for Information Exploration: An Implementation and Evaluation. *Proceedings ACM CHI'92: Human Factors in Comp. Systems*, pp. 619–626

News in short…

Enlisting HTS and genomics to defeat the 'superbugs'

EVOTEC BioSystems (Hamburg, Germany) and **Genome Pharmaceuticals Corporation** (GPC; Munich, Germany) are to collaborate on the development of a new generation of broad-spectrum antibiotics based on a novel class of genomics-derived targets. This collaboration will involve GPC supplying an undisclosed novel antibacterial target that has been identified and validated using their integrated second-generation genomics technologies. The appropriate HTS assays will be developed by EVOTEC, who will also perform small-molecule compound library screening for lead identification. Both companies anticipate that this project will evolve through pre-clinical discovery and envisage forming an alliance with a pharmaceutical company to perform the clinical development work.

At present, the tide of antibacterial resistance is increasing. New generations of bacteria, commonly known as 'superbugs' because of their resistance to current therapies, are causing major concern and are creating serious healthcare problems. Berndt Seizinger, GPC's CEO commented, 'The development of a new generation of antibiotics with different mechanism(s) of action is urgently needed'. He added '…the powerful synergy between GPC's target identification methods and EVOTEC's unique assay development and screening systems will significantly accelerate the development of a new generation of genomics-derived antibiotics.' If this collaboration is successful, these companies may enter a broader alliance to search for additional innovative targets for antibacterial research.

Cantab Pharmaceuticals (Cambridge, UK) have announced the results of their Phase I clinical trial with the potential anti-cocaine vaccine, TA-CD. This vaccine comprises a cocaine derivative coupled to a carrier protein, rCTB, and is intended to generate cocaine-specific antibodies, interfere with the transport of cocaine from the bloodstream to the brain and neutralize its psychoactive effect.

The trial involved 34 subjects with a history of cocaine addiction, who were given three injections of the vaccine at four-weekly intervals. The results of the trial indicated that all the patients mounted a dose-related antibody response to the vaccine, which persisted for at least 84 days and was capable of recognising free cocaine in the blood. Furthermore, no serious vaccine-associated adverse effects were reported and the trial demonstrated a positive safety and tolerability profile. The company now intend to enter TA-CD into randomised, placebo-controlled Phase II trials in the latter half of 1999, to evaluate the effect of TA-CD on reducing patient cocaine use, whilst gaining further knowledge of the safety and immunogenicity of the product. The success of this trial has also provided the encouragement required for the company to enter TA-NIC, a potential anti-nicotine vaccine based on a similar scientific concept, into Phase I trials.