The application of non-combinatorial chemistry to lead discovery

Jeremy Everett, Mark Gardner, Frank Pullen, Graham F. Smith, Mike Snarey and Nick Terrett

Non-combinatorial chemistry is a powerful technology for the synthesis of large numbers of compounds, with complete control over the properties of those compounds. We have developed a Library Creation, Registration and Automation system (LiCRA), which harnesses an efficient non-combinatorial chemistry design and synthesis engine, together with high-throughput automated purification. This LiCRA system also operates in a closed loop mode for hit-to-lead optimization, and contains an integrated IT system that controls and facilitates all aspects of the operation from design to registration. Quality has been our watchword, from the quality of compound design through to the quality of the products.

***Jeremy Everett**
Mark Gardner
Frank Pullen
Graham F. Smith
Mike Snarey
Nick Terrett
Medicinal Technologies
Pfizer Global R&D
Sandwich
Kent, UK
CT13 9NJ
*tel: +44 (0)1304 616161
fax: +44 (0)1304 656259
e-mail: jeremy_everett@sandwich.pfizer.com

▼ The launch of combinatorial chemistry onto an unsuspecting pharmaceutical industry in the early 1990s resulted in several frantic efforts as companies tried to maintain a competitive edge through the generation and screening of compounds in unprecedented numbers and at an unprecedented rate. The focus at that time was solid-phase synthesis and the generation of compound mixtures, resulting in library arrays of many millions of compounds. These combinatorial libraries required considerable analytical detective work to ascertain whether the desired compounds had been made. Frequently, the library approaches were frustrated by restrictive solid-phase chemistry of limited success or generality. Furthermore, the observation of biological activity in assays of mixtures frequently resulted in a wild-goose chase that employed sophisticated or tedious deconvolution or decoding, pursuing hits that often vanished: the false-positive problem. Having suffered several resource-intensive yet fruitless pursuits, in the mid- to late-1990s, many companies were forced to radically rethink their combinatorial strategies[1–4]. There was no doubt that the new technology and the ready availability of automated synthesis offered opportunities for increased productivity in both file enrichment and lead optimization. However, it was a certainty that chemistry would not permit easy success when there was such a casual disregard of the basic tenets of good science – careful experimental planning and meticulous observation of results. Indeed, we commented in 1995 that in the future of drug discovery through combinatorial chemistry, 'quality, rather than quantity, will become the new goal'[5].

Thus, library synthesis reverted towards the old values of quality, both in compound design and product analysis. Although this inevitably resulted in fewer compounds being prepared overall, we have now reached a stage where those that are made are the products of more considered design, meticulous protocol development and thorough structural analysis. Medicinal chemists who exploit high-speed chemical techniques are increasingly making single compounds, depending primarily on the proven legacy of solution-phase chemistry. The products are often purified, or subjected to rigorous analysis, before being sent for registration and screening.

We have now reached a phase in the development and application of high-speed chemistry that permits library compounds that are active in biological screens to be identified without deconvolution, fractionation or tag decoding, and, more significantly, there are much fewer false-positives. Because the preparation of targeted single compounds removes the requirement for the synthesis of combinatorial arrays, we refer to our approach as being 'non-combinatorial'.

1359-6446/01/$ – see front matter PII: S1359-6446(01)01876-1

This article describes the non-combinatorial approach developed and used by our Library Design and Production Group, and demonstrates that a combination of rational design, careful synthesis and compound purification before testing can generate high quality biological data that result in novel leads and new drug discovery programmes.

Non-combinatorial chemistry

Non-combinatorial chemistry is a phrase coined to express the difference in strategy between our technique and combinatorial chemistry as it is more commonly applied. Combinatorial chemistry involves the use of all combinations of a set of reagents to produce an ordered array of products. For example, a combinatorial amide library synthesis starting with 80 amines and 60 acids would make 4800 amide products (80 × 60). In non-combinatorial chemistry, the same reagent sets can be used, but the products would be individually chosen for synthesis, such that the 80 amines and 60 acids might be used to select any number of products from 80 up to the combinatorial limit of 4800, and still use each reagent. This approach is also called 'cherry-picking' by some groups. The key differential is to allow the design, synthesis, purification and screening of any group of individual compounds from a virtual library, not just arrays of compounds. The drive for this strategy is the desire to exploit virtual libraries that are as large as possible in a powerful and flexible manner, while retaining total control over the properties of each product. The approach is important in both file enrichment and lead optimization.

Application of non-combinatorial chemistry to file enrichment

Non-combinatorial chemistry is important in file enrichment in that it allows the synthesis of diverse sets of compounds, using as many monomers as possible, not just a small subset. The reason why this is important is best illustrated by an example. Consider a three component library, represented as A–B–C. Let us assume that A are amines, B are amino acids and C are acids, so the products will be diamides. It can be easy to acquire 500 amines, 100 amino acids and 500 carboxylic acids from commercial sources, giving a virtual library size of 25 million compounds. Clearly, it is unlikely that 25 million single compounds will be synthesized using parallel synthesis. Therefore, a choice has to be made as to how to exemplify this virtual library (with perhaps 8000 examples) in a compound collection for screening purposes. A typical approach might be to select 20 amines, 20 amino acids and 20 carboxylic acids to give a combinatorial array of 8000 examples. Frequently, diversity-based methods might be used to select the representative building blocks. However, this is an approach with significant drawbacks; a typical situation that might be faced in selecting 20 amino acids from a set of 100 is presented in Fig. 1.

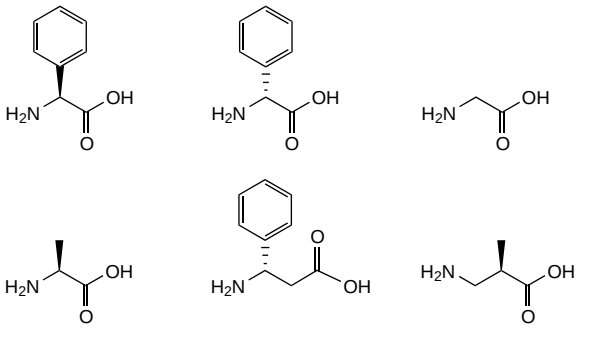

Figure 1. Six amino acids. Your task is to choose the two most diverse amino acids that should be selected as monomers for file enrichment. This is representative of the problem faced in combinatorial file enrichment, in which a small number of monomers have to be selected in an attempt to represent the properties of a much larger set.

As a subset of the whole exercise of choosing 20 from 100, two amino acids might have to be chosen from the set of six presented in Fig. 1. If you have been involved in such an exercise, you will recognize the similarity of this situation with examples from your own experience. Although software packages might be able to make a selection, the truth is that there is no correct way to makethis choice for a file enrichment library. Whichever pair you choose, they clearly cannot adequately represent all six. Once you have made your selection of a pair of amino acids, remember that in this library of 8000 compounds, each of those amino acids will be present in 400 products (in combination with the 20 selected amines and 20 selected carboxylic acids). The other four amino acids will be present in no products at all.

An alternative approach is to construct the 8000 compounds in a non-combinatorial manner, using all the available building blocks. In this example, if each of the components A, B and C is used equally, each amine and carboxylic acid would be present in 16 (8000/500) products, and each amino acid in 80 (8,000/100) products. Arbitrary choices of monomers no longer need to be made – all can be used in the file enrichment library – thus providing much greater diversity of products. The preparation for synthesis of such a library does take longer than for a combinatorial array (it takes longer to weigh and solubilize 1100 components than 60). In addition, automation is now essential for synthesis because neat arrays that could be handled manually are no longer being used.

A further significant benefit of the non-combinatorial approach is that you can choose not to synthesize specific products, and thus precisely tailor the physicochemical properties of the non-combinatorial set. For example, it is straightforward to calculate the properties of the products and to ensure that all of those to be synthesized should comply with certain distributions of molecular weight or clogP values, Lipinski's rule of five[6] or other desired properties. An example of the effect of this flexibility can be seen in Fig. 2.

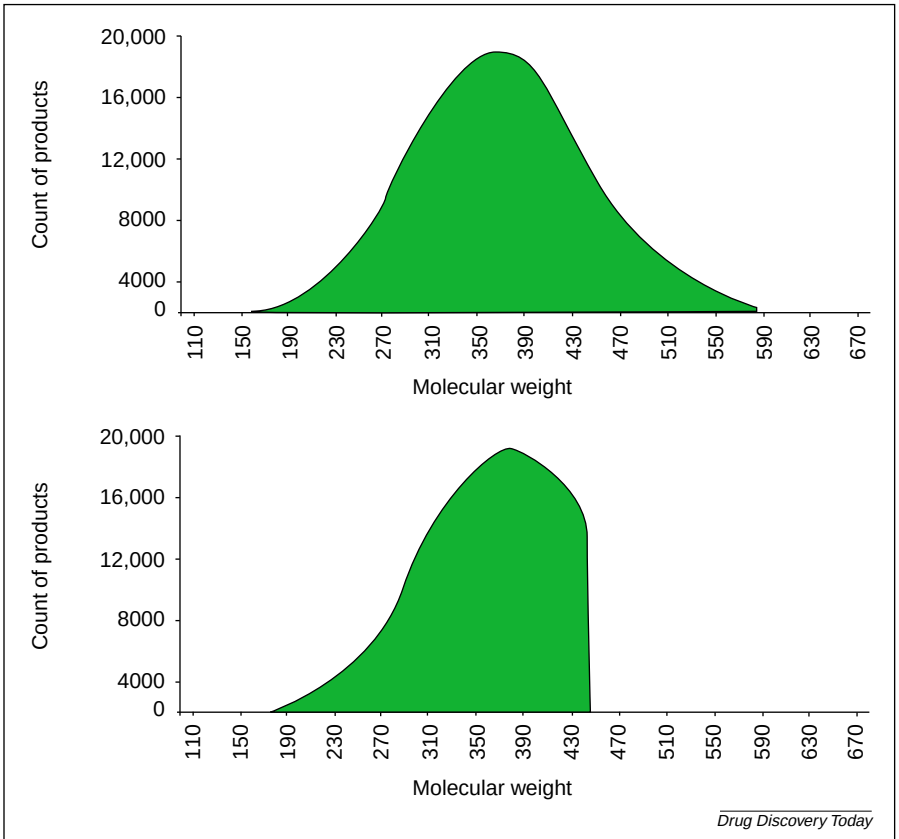

Figure 2. An illustration contrasting property distribution in a combinatorial selection with that in a non-combinatorial selection. Well-defined property cut-offs can be readily achieved via non-combinatorial approaches.

The standard combinatorial array would give something approaching a normal distribution of a property such as molecular weight. However, with non-combinatorial synthesis, you can synthesize only those products with molecular weights below a given value or in a given value-range. Our in-house library design software, S1D (selection in one dimension, i.e. by additive or approximately additive fragmental properties, such as molecular weight or clogP) allows for the application of these design principles and has been used to select diverse libraries of specific design for file enrichment for three years. S1D makes pseudo-random selections of products from a virtual library in a way that attempts to use each monomer an equal number of times, but with boundary conditions such that the calculated properties of each product (molecular weight, clogP, and so on) fall within predetermined limits. S1D links to the Library Creation, Registration and Automation system (LiCRA). Both LiCRA and S1D were written by the chemoinformatics and Discovery Research informatics groups in Pfizer Global R&D, Sandwich, UK.

Application of non-combinatorial chemistry to lead optimization

Although non-combinatorial synthesis offers significant benefits in file enrichment, the strategy is particularly powerful once there are some design principles to apply, such as when an active compound or a crystal structure is in hand. Non-combinatorial synthesis gives the medicinal chemist complete freedom to choose any selection of single compounds from a given virtual library to expand upon this structural information. Virtual libraries can also be searched by computational techniques, such as pharmacophore searching or crystal structure 'docking', and the hit-lists synthesized as they are or manipulated further.

Non-combinatorial synthesis using LiCRA

Our first significant problem was that at the time we started, nobody had considered that non-combinatorial synthesis would ever be a user requirement of 'combinatorial chemists', and still many do not. We were drawn to our solution by the knowledge that our HTS groups used 'cherry-picking' software on Tecan liquid handling robots (Tecan UK, Reading, UK) to choose active compounds from HTS screening plates. However, no software was available at that time for our library synthesis task, because of our unique in-house requirements, and we decided, therefore, to design and build LiCRA ourselves.

When our group started work in 1997, we used each piece of automation in isolation to try to understand what inputs and outputs were important for us to monitor and change. In general, most commercial automation runs on its own software with minimal opportunities for broader integration. A flat file can often be read to load sample data

and a further file is often written to the workstation to log automation activity. Our first attempt at synthesis software simply read flat files from a shared network drive and produced an input file for the next step of the process. This proved successful and was flexible in the event of machine errors. User intervention and file manipulation, however, was causing problems with tracking, and serious human errors were increasing.

The latest version of LiCRA is based around an Oracle database (Oracle UK, Reading, UK) of the individual compound batches that we have chosen to make. Their non-combinatorial selection means that, compared with array synthesis, there is much less common data from one batch in a library to the next. Our main reasons for moving to a database-centred system were to increase storage and security and to facilitate data analysis and trend recognition. Compliance with data storage and retrieval requirements for patent purposes made us look into this more robust approach.

The LiCRA user deals with a Visual Basic interface, which treats all the compounds in a library together (e.g. reaction temperature, molar equivalents) but tracks most other data at the individual compound level (e.g. yield, purity and protecting groups). A simple input interface that requires the product structure in MDL (Molecular Design Ltd, Camberley, UK) Structure definition (SD) file format, and an ASCII (American Standard Code for Information Interchange) file containing compound numbers with constituent monomers, is all that is required to load a library. Typical libraries range from 200 to 2000 compounds. In planning a synthetic scheme, volumes, equivalents and so on, can be entered for a whole library but can also be altered on a monomer-by-monomer basis. An automatic reservation can be made against the monomer inventory system and lists of quantities printed if required. Monomers that appear in the library design but are not present at all, or are in insufficient quantity in the monomer store, are flagged for intervention. The LiCRA inventory control system interfaces directly with accurate balances during weighing and calculates the exact amount of each monomer weighed. The weighed monomer samples are bar-coded and loaded into racks for synthesis. LiCRA produces a report that warns when monomers are below a predetermined re-order level, so that the inventory can be maintained.

The LiCRA synthesis system uses several standard and robust pieces of commercial hardware. All solution dispensing activities are performed by Tecan Genesis liquid handling robots, which have a proprietary interface for non-combinatorial sample distributions. Specific racks and boxes have been custom-made to enable inert atmosphere, corrosive reagent and solid-phase chemistry capabilities. All chemical incubations are performed off-line in simple shakers or ovens, and all products are sent through our in-house autopurification system (see later). Samples are reformatted into 96-well Micronics plates (Micronics, Redmond, WA, USA) for registration that allows a final quantitation by weighing to be performed using a Bohdan® Weighing Station (Mettler-Toledo Myriad, Royston, UK). The user is then presented with quality-control data on a plate-by-plate basis to allow review of spectra and individual sample data before registration.

LiCRA has built-in links to several chemical structure databases to allow easy analysis of monomer solubility, reactivity and other properties. Product structures are also stored within the LiCRA database to allow more chemically intuitive quality control. Because all synthesis information is stored within the database, trends in solubility and reactivity of individual monomers can be tracked to enhance the quality of future libraries. When the chemist is satisfied with the library products, a structure definition file, which contains all relevant information, is produced and sent to our compound control centre for registration to the corporate database.

Automated purification of compounds from non-combinatorial synthesis

The non-combinatorial chemistry approach to file enrichment and lead optimization, as described earlier, generates large numbers of single compounds, all of which need to be purified to minimize the presence of impurities and the generation of false-positives in biological screens. Analysis, purification and quantitation of large numbers of compounds in 96-well plate format has been viewed by many as a highly labour intensive process. Systems are available for high-throughput purification (e.g. the Biotage Parallex Flex™ system – Dyax, Charlottesville, VA, USA). However, there are currently no commercial purification systems on the market that can effectively process, quantify and track large numbers of compounds in an automated way. The key to any automated purification system is a fully integrated sample tracking process, coupled with proven purification technology and linked to a user-friendly software interface. A useful review of automated purification methodologies and emerging supercritical fluid chromatography-based methods is available[7].

We have developed an autopurification system based on gradient HPLC separation, off-line MS analysis of fractions and quantification by light scattering that has the capacity to purify up to 100,000 compounds per year at the 10–50 mg scale. The software interface (OmniTrak, developed by Aitken Scientific, Thame, UK) is fully integrated with LiCRA, and controls all the HPLC, MS and robotic hardware

involved in the purification process. There are eight semi-preparative HPLC systems [Gilson 306 gradient systems, with Gilson 215 liquid handlers configured as both autosamplers and fraction collectors, operating under Unipoint control software (Gilson, Middleton, WI, USA)]. The HPLCs all use simple generic gradients to effect separation and purification (acetonitrile–0.1% trifluoroacetic-acid gradients running from 5–95% organic over 8 min at 4 ml min^{-1} with a total run time of 15 min). Each 96-well compound plate has a barcode generated from LiCRA linked to database records describing the 96 expected compounds and assigning an empirical formula to each well. Each of the 96 compounds is fractionated by gradient HPLC into bar-coded 48-well plates, collecting up to five fractions per compound using UV-triggered fraction collection. The 48-well plates are manually loaded onto a carousel and an ORCA® robot then automatically transfers the plate to a Gilson 231 autosampler for analysis (Fig. 3). Each fraction well is analyzed by flow injection MS (Micromass® Platform LC, Micromass® UK, Manchester, UK) that operates in parallel with evaporative light scattering (Sedex Sedere 55, Thomson Instrument Company, Chantilly, VA, USA). The MS identifies the fraction or fractions with the expected empirical formula for each of the 96 compounds throughout the series of 48-well fraction plates. The evaporative light-scattering data provides an estimate of the amount of compound present in the fraction by reference to a calibration file[8]. The correctly identified fractions that have sufficient sample mass are then automatically selected and compressed into a bar-coded plate of tared 5 ml Micronics tubes to be evaporated to dryness (Genevac, Ipswich, UK). After autopurification, all the MS and chromatographic data on any compound can be accessed from within LiCRA.

Figure 3. A view of part of the autopurification system: the linear track ORCA® robot plate-handler (Beckman Coulter, Fullerton, CA, USA) is housed in a recirculating safety cabinet. The MS and evaporative light scattering unit are situated below the bench on the right-hand side. Some of the eight semi-prep HPLCs can be seen in the background.

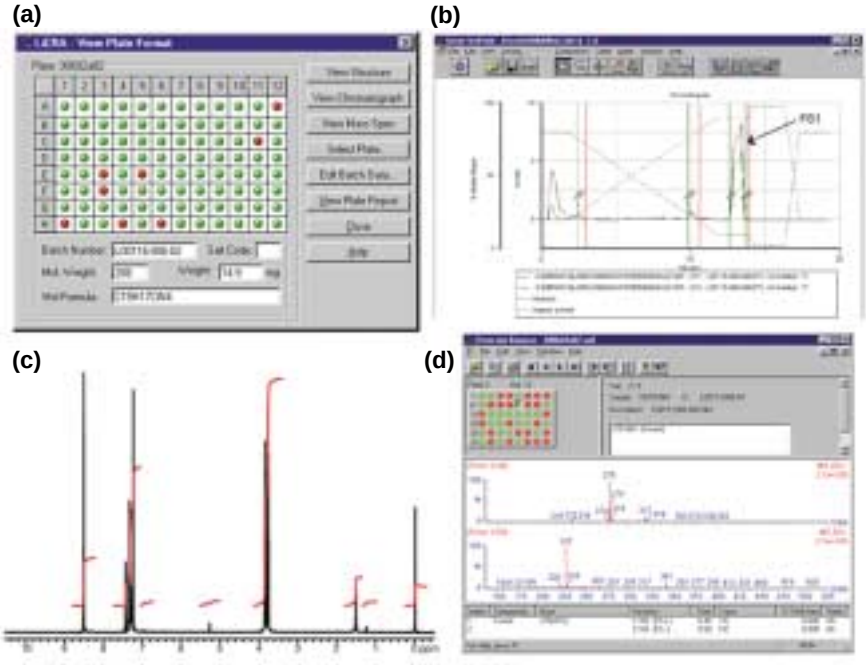

Figure 4. (a) A plate view from the LiCRA system, illustrating the connectivity between the 96-well plate of products and all of the associated structural, analytical and synthetic information. The colour coding illustrates success (green) or failure (red) of the chemistry at the well level. **(b)** A typical dual wavelength UV chromatogram of the crude product from a library synthesis. **(c)** Positive- and negative-ion electrospray MS of the desired fraction (fraction 61, vial 13) from the purification of this crude product. **(d)** A typical ^{1}[H] NMR showing the quality of compounds obtained using this autopurification system.

Using this technology, library chemists can purify their compounds without involving themselves in the detailed understanding of HPLC or MS, as all the identification, analysis and quantification is carried out in an automated way. Compounds purified in this high-throughput and automated fashion are of comparable quality with those produced by traditional synthesis and purification technologies; they are typically of 90–95% purity (Fig. 4). All the compounds are tracked from their synthesis to registration using a combination of OmniTrak and LiCRA, and all the analytical data for a particular compound are linked to that tracking process.

Application of S1D and LiCRA non-combinatorial chemistry to hit-to-lead optimization

We have applied the S1D and LiCRA systems to several target types. The example that follows illustrates the power of the technology to optimize some weak hits coming from a kinase screen to generate multiple, potent leads for chemistry follow-up. Five thousand compounds were selected using S1D for non-combinatorial synthesis from a virtual library of 300,000 compounds that was designed specifically for a kinase screen. Figure 5 illustrates the type of chemistry used in this library subset. Figure 6a shows that the only activity observed on screening these compounds came from three members of a subset of 700 compounds made from just two templates: all other templates (4300 compounds) were inactive. The three active compounds were repeatedly tested and had IC_{50} values of 5–10 μM. The hits were followed up by non-combinatorial synthesis of a closed-loop subset of 693 compounds from the same virtual library, focussing in on the regions of chemical space occupied by the three original hits and related regions (Fig. 6b). This subset gave a respectable 2.3% hit rate at 10 μM. Furthermore, repeating the screening with solid samples for some of the more active compounds gave the same K_i values, reinforcing our confidence in the identity and quality of the parallel synthesis-derived liquid samples used in the screens.

The early SAR information generated from this first closed-loop prompted the synthesis of a second loop of 345 compounds, which gave an impressive 66% hit rate against the enzyme at 10 μM (Fig. 6c). These compounds had improved solubility, and exhibited weaker activity and independent SARs against key kinase-selectivity targets. Another 1400 compounds were made in four succeeding closed-loops, in the course of which the virtual library was expanded from 300,000 to 4,600,000 compounds by the addition of further custom and commercial monomers. A summary of the outcome of all six closed-loops is shown in Fig. 6d. It is clear from Fig. 6a–d that S1D has been used to avoid the selection of high molecular weight and lipophilic compounds that would occupy the top right-hand side of the charts, because of the way in which the monomer lists were ordered. The fifth and sixth loops delivered three different series of selective and potent compounds with activity against the kinase in the region of 10–50 nM. All of the 2500 rule-of-five[6] compliant compounds made in the six closed-loops were synthesized as singletons, each one of which was analyzed and autopurified as described above; all with less than one 'person-year' of effort. The leads we discovered have provided the basis for a chemistry-staffed lead-optimization project, which is now under way.

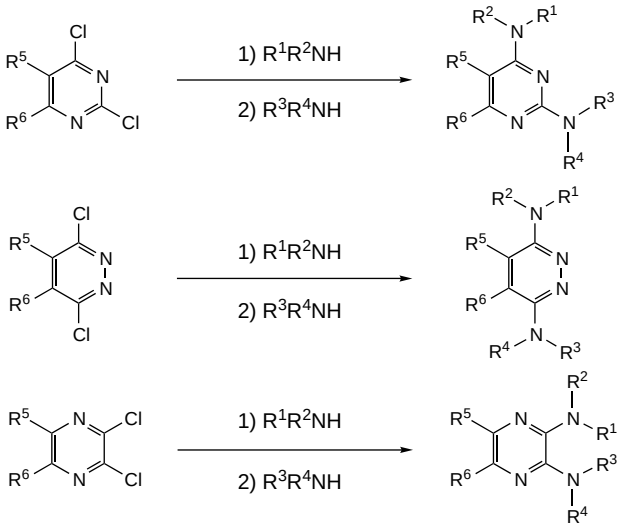

Figure 5. The design for some of the diamino-substituted aromatic heterocycles used in the kinase virtual library. This library employed simple, solution-phase, chlorine displacement chemistry using large available monomer sets in 96-well polytetrafluoroethylene plates.

Conclusions

We have demonstrated that non-combinatorial chemistry is an effective method for screening-file enrichment and for lead optimization. We believe that whatever strategy is used for library chemistry, the quality of design and the quality of the products are of paramount importance; we have achieved this with our S1D and LiCRA systems. We intend to build upon this system in the future and add new design tools and new synthetic capabilities to expand the scope and quality of the virtual chemistry space accessible to our chemists.

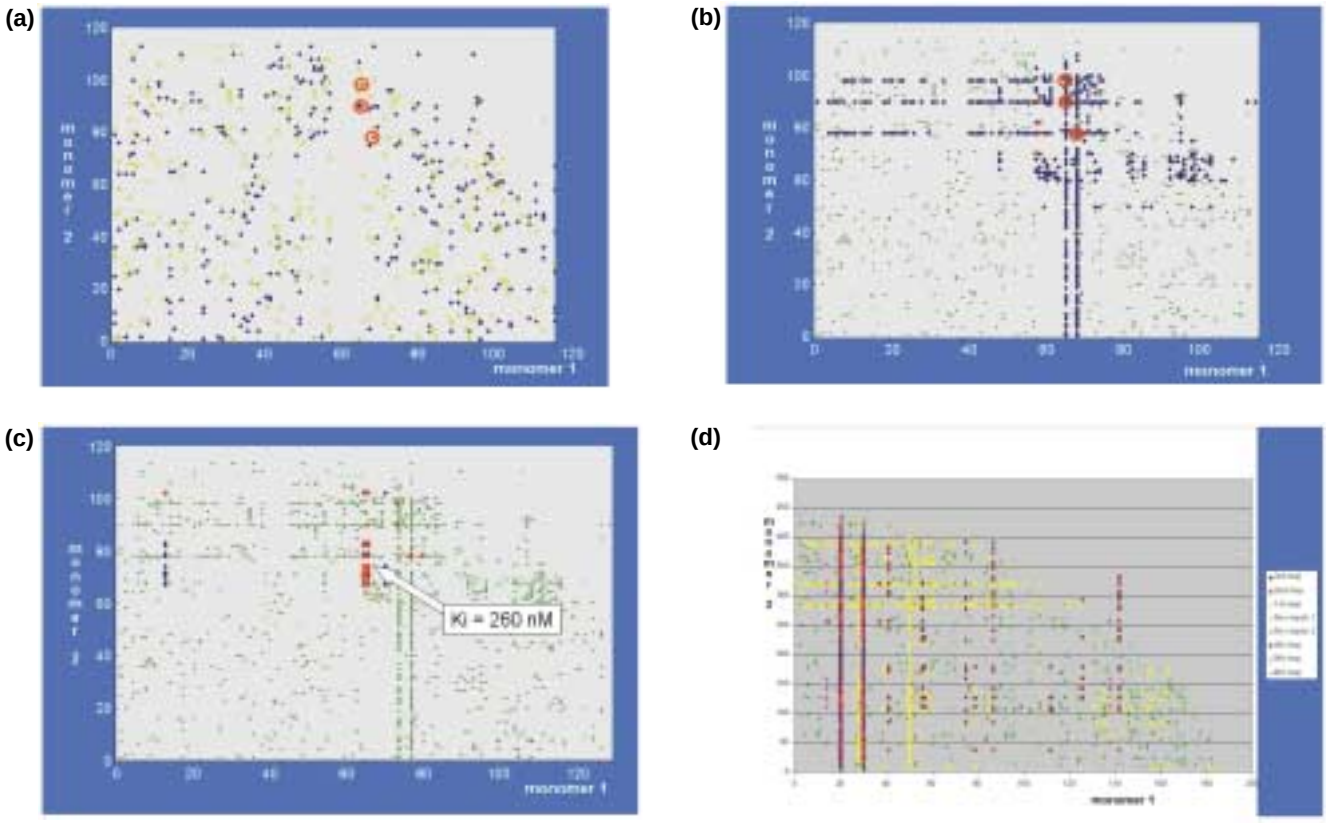

Figure 6. (a) Non-combinatorial selection of 700 compounds from a virtual library of 300,000 in two templates: yellow and dark-blue diamonds. The three active compounds are circled in red. Each axis of the chart consists of a list of the amine monomers used in each position around the aromatic core. The *x*-axis corresponds to the amines used to displace the most reactive chlorine, the *y*-axis to the amines used in the displacement of the second, less reactive chlorine. The amines were ordered by similarity (according to a judgement made by medicinal chemists; see Fig. 7 for design outline). **(b)** A corresponding view of the 693 compounds made in the first closed loop in response to the three hits found in Fig. 8a. The original 700 compounds screened are represented as green dashes; the 693 new compounds in the first loop are shown as blue diamonds (inactive) and red circles (active); the three original hits upon which the SAR was based are still shown circled in red. In this display, some of the compounds synthesized sit on top of one another in the chart, as multiple templates sometimes shared the same monomer pairings; this third dimension of the data is missing from this compressed 2D view. **(c)** Corresponding figure for the second closed-loop; all previous compounds are now shown as green dashes, and the second loop compounds as blue diamonds (inactive) or red circles (active); again, this is a compressed 2D view of a 3D dataset. **(d)** A compressed 2D representation of all six closed-loops (CLs). The compounds from the six loops are represented with the following symbols: CL1, yellow triangles; CL2, magenta squares; CL3, blue diamonds; CL4, brown circles; CL5, open red diamonds; CL6, blue dashes; two groups of file compounds are shown as crosses and as crossed crosses.

Acknowledgements

We are grateful for the expert assistance provided in hardware development, automation, information systems, chemistry and screening by the following Pfizer colleagues: Mark Kemp, John Taylor, Pam Greengrass, Duncan Armour and their teams; Ruth Tucker, Mark Lord, Ian Parsons, Jason Robson, Derek Hearn and Wilma Keighley and her team; James Merson and his team; and George Perkins, Chris Selway, Mike Stace, Mark Taylor and Adrian Wright. We are also grateful to Andy Kemp of Aitken Scientific, Ian Hibbard and Grant Cameron of Anachem and Jeremy Batt of Micromass for their productive collaboration on the autopurification system, and to David Webb of PA Consulting for project management of LiCRA and S1D.

References

1 Labaudiniere, R.F. (1998) RPR's approach to high-speed parallel synthesis for lead generation. *Drug Discov. Today* 3, 511–515
2 Hird, N.W. (1999) Automated synthesis: new tools for the organic chemist. *Drug Discov. Today* 4, 265–274
3 Dolle, R.E. (2000) Comprehensive survey of combinatorial library synthesis: 1999. *J. Comb. Chem.* 2, 383–433
4 Coates, W.J. *et al.* (2000) Successful implementation of automation in medicinal chemistry. *Drug Discov. Today* 5, 521–527
5 Terrett, N.K. *et al.* (1995) Combinatorial synthesis – the design of compound libraries and their application to drug discovery. *Tetrahedron* 51, 8135–8173
6 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25
7 Ripka, W.C. *et al.* (2001) High-throughput purification of compound libraries. *Drug Discov. Today* 6, 471–477
8 Fang, L. *et al.* (2000) Evaluation of evaporative light-scattering detector for combinatorial library quantitation by reversed phase HPLC. *J. Comb. Chem.* 2, 254–257