The Entrance of Informatics into Combinatorial Chemistry

Richard Pommier Swanson

Abstract

Combinatorial chemistry has risen to prominence as the pharmaceutical development strategy of choice in less than twenty years. In that time the combinatorial library has emerged as a central artifact. Libraries grew from random collections of a hundred molecules into immense arrays of hundreds of thousands of molecules. The definition of a "good" library has changed, from a large, maximally diverse set of compounds to a small set of diverse, "drug-like" molecules. The process of screening libraries to find new lead compounds now includes virtual screening, in an effort to reduce the workload and speed up the movement of a lead into the clinical testing phase of development. The success of all these aspects of the drug discovery enterprise has relied heavily on informatics techniques to index, archive, search and retrieve library data, and design, construct, analyze, and screen the libraries themselves.

The entrance of informatics into combinatorial chemistry took place roughly midway through its twenty-year existence. The combinatorial explosion provided a set of challenges for which informatics offered solutions. At first the challenge was to create and navigate through as large a solution space as possible. Later, drug researchers chose a more reserved approach, using genetic algorithms along with biological and chemical quantitative structure-activity relationship, or QSAR, data to reduce the search space. The success of combinatorial synthesis for all intents and purposes marks the end of the era of so-called Woodwardian synthesis, where, to borrow a phrase, compounds were synthesized "the old-fashioned way, one molecule at a time."

The last forty years have seen significant changes in the way that chemists practice their discipline. Perhaps the most significant change is the trend that has chemists moving away from the laboratory and into cyberspace to design and develop new molecules, particularly pharmaceuticals. A new approach to molecular design and synthesis planning has evolved, one in which the computer is the key player. Timothy Lenoir (1998) has already noted the reshaping of the field of biomedicine into an information science, and methods borrowed from the field of informatics have become part of the standard repertoire of strategies that chemists use.

In this paper I examine the role of informatics in one specific branch of chemistry, namely, combinatorial chemistry. I have chosen this field because it places particularly strong demands on the information sciences for its successful execution. As a relatively new practice—scarcely twenty years old—combinatorial chemistry, or combichem, did not usher in the age of chemical informatics, but it certainly grew up in the information age. As such, combichem makes for an interesting case study of the efforts required to merge informatics with chemistry. My goal is to identify the special challenges that combinatorial chemistry presented to chemists for which informatics provided solutions. I shall present some examples of how and when these solutions materialized in order to illustrate how informatics became integral to combinatorial chemistry.

What Is Combinatorial Chemistry?

Combinatorial chemistry is a laboratory technique whereby chemists synthesize a large number of molecules differing from each other in chemical make-up or structure. The phrase *a large number* means as many as tens of thousands of unique molecular variations; nowadays a single laboratory might prepare as many as a million new compounds each year. The point of doing so is to create a pool of molecules to be screened for properties that might make them suitable candidates for new

pharmaceuticals. This strategy is a change from the days when a chemist selected a single target molecule as a candidate, then developed a synthesis strategy to arrive at that molecule and only that one.

Combichem is rooted in a laboratory technique where chemical reactions (linking amino acids into new sequences, as one example) are carried out on a solid substrate rather than in solution, as so much chemistry usually is done. Solid-phase chemical synthesis, as it is known, dates back to the mid-1960s and the Nobel Prize–winning work of R. Bruce Merrifield, a chemist at Rockefeller University of New York (Merrifield, 1965). It was not until early in the 1980s when a novel innovation on solid-phase synthesis changed the nature of the practice irrevocably, leading to the technique now called combinatorial chemistry.

The announcement of this innovation first appeared in the pages of an obscure Hungarian paper dated 1982, where chemist Arpad Furka (2002) described his first attempt at a combinatorial approach to solid-phase synthesis (Seneci, 2000). Imagine five small "building block" molecules (generically, A through E) and all possible three-member sequences of those molecules (e.g., A-A-A or C-E-B). Furka's approach was to synthesize all possible sequences, not just some preselected examples, using the solid-phase synthesis technique. The details of how this was done do not concern us here, but it is important to recognize how quickly the technique caught on and grew in scope. To illustrate the growth, an early combinatorial synthesis scheme in the mid-1980s might have produced a collection of forty or even a hundred molecules. Less than a decade later Kit S. Lam's group reported generating an array of 2.5 million compounds (Lam et al., 1991). These arrays were called "libraries" of molecules, and navigating through them became one of the primary informatics tasks combinatorial chemists faced. The combinatorial library is the primary object around which informatics techniques are centered.

What Is Chemical Informatics?

The coining of the term *chemoinformatics* has been attributed to Frank Brown (1998), who wrote about the emerging importance of information handling in the field of drug discovery. Brown did not provide a succinct definition of the term in his paper. I am not aware of any statements in print that provide a clear and universal definition of chemical informatics. Having such a definition before us is important for the clarity of this paper, so I will provide one for that purpose shortly. The field of bioinformatics has working definitions, such as the one proposed by the definition committee of the National Institutes of Health Biomedical Information Science and Technology Initiative Consortium in July 2000: "Bioinformatics is research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data" (National Institutes of Health, 2000).

I propose a working definition for chemical informatics similar to this, with appropriate substitution of chemical terms for the biological ones: chemical informatics is research, development, or application of computational tools and approaches for expanding the use of chemical data, including those to acquire, store, organize, archive, analyze, or visualize such data.

The central artifact in my paper is the combinatorial library, and so I am interested in how the above definition applies to it. A keeper of a chemical library must have a system for uniquely identifying and retrieving each member in the library's collection, just as a book librarian must have such a system. Establishing a clear and consistent nomenclature for all known (or proposed) chemical substances has been an enduring challenge to chemists for more than a century (Wiswesser, 1968). Some nomenclatures are readily mastered by chemists but are not always suited to manipulation by computers. The most suitable system for computers is a so-called line notation, where each molecule's structural identity is represented in a linear string of symbols. The International Union of Pure and Applied Chemistry (IUPAC) rules for naming compounds provide a systematic way of naming molecules, and experienced chemists use these rules comfortably. The IUPAC system is not a satisfactory line notation for use with computers. Finding a line notation suitable for computer manipulation was central to the successful implementation of chemical informatics in combinatorial chemistry.

Identifying the members of a library is one challenge, but determining what members should be in the collection is even more challenging. This task is known as library design. A primary decision to make in library design is whether the potential members of the library should be chemically similar or chemically diverse. Early pharmaceutical search strategies involved identifying molecules that were similar in structure and biological activity to known drugs. Similarity analysis was one kind of informatics tool that chemists used in library design.

However, once chemists analyzed the forty-five best-selling pharmaceuticals then on the market (Martin et al., 1995), a different structural feature was deemed important. These forty-five molecules demonstrated structural *diversity* more so than similarity. Combinatorial chemists sought a new ideal, a library of diverse molecules, anticipating that a new drug was more likely to come out of a diverse library than out of a library of closely related molecules.

Designing a diverse chemical library and then synthesizing its members still are not enough to ensure a fruitful search for a new drug molecule. To return to the analogy with books once more, now that one has filled a library with a large number of books, one must decide which books are worth reading after all. In chemical terms this means screening the members of the new library to find the ones that are most likely to exhibit the desired therapeutic activity. Chemists first screened candidates from the library using laboratory methods, but we will see how a type of analysis known as quantitative structure-activity relationship (QSAR) analysis led to the virtual screening of libraries.

Informatics in Chemical Line Notation

The first successful chemical naming system readily usable by computers was the Wiswesser line notation (WLN), developed by William J. Wiswesser in 1949. Among Wiswesser's concerns was developing a system that only used characters found on contemporary typewriter keyboards, avoiding visual subtleties like superscripts and subscripts. Adaptations for the limitations of tabulating machine keyboards and computer punched cards were available. Wiswesser identified the key features of a well-crafted line notation, citing the desiderata for any structure-describing notation given by the Coding Commission of IUPAC. Those desiderata were simplicity of usage; ease of printing, typewriting, and manipulation by machine methods; instant recognizability; conciseness; and uniqueness (Wiswesser, 1954). The WLN remains a viable line notation to this day, but it proved less satisfactory for combinatorial chemistry than a successor notation, known as SMILES (Simplified Molecular Identification and Line Entry System).

SMILES is both a notation and a software package that was developed in 1987 by the team of David Weininger, Gilman Veith, and Eric Anderson. Veith and Weininger worked together at the Environmental Protection Agency's Environmental Research QSAR Laboratory in Duluth, Minnesota, before Weininger took an academic position at Pomona College in Claremont, California. In their initial publication announcing SMILES, Anderson, Veith, and Weininger (1987) stated that the main advantage of SMILES over the WLN was that it was easier for nonchemists to use and it required a minimum of computer processing time. Whereas the WLN had at least eighteen hierarchical rules to follow when developing a molecular line representation, SMILES had just five. Ultimately, the SMILES software included an algorithm to create a unique SMILES notation from the user's input notation of choice, easing the user's burden even further.

The theoretical basis for SMILES was molecular graph theory (Weininger, 1988). Briefly described, molecular graph theory views molecules by their atoms and the connectivity between them, that is, their bonds. In graph theory one can classify and order the nodes and edges of a graph; nodes and edges are analogous to atoms and bonds. Weininger, Veith, and Anderson used the analytical methods found in graph theory to generate linear strings of symbols representing the three-dimensional structures of molecules, that is, the arrangement of atoms and bonds. Grounded in information theory, SMILES in turn influenced the development of new informatics tools and techniques.

There are abundant references in the chemical literature illustrating the use of SMILES notation to carry out analyses of chemical data sets. Two examples will serve to illustrate the point. Less than a year after the announcement of the SMILES notation David Weininger left his position at Pomona College to form Daylight Chemical Information (DCI) Systems, Inc., the company that now owns and markets the SMILES software package. DCI collaborated with Parke-Davis Pharmaceuticals to develop a method for seeking structural commonalities in diverse data sets (Shemetulskis, Weininger, Blankley, Yang, & Humblet, 1996). A computer algorithm dubbed Stigmata performed the analysis, seeking commonalities among publicly available chemical databases. Stigmata took SMILES notations for the molecules in the database, then developed so-called fingerprints for the molecules in the set. A fingerprint here meant a 2,048-bit string representing a substructure of a molecule that could be common to a number of different molecules. Analysis of the frequency of different fingerprint patterns indicated the degree of commonality among the members of the data set. The hope was that this type of fingerprint analysis could be used effectively for designing combinatorial libraries.

Another example of how SMILES notation stimulated new chemical informatics methods is the collaboration between the Krebs Institute for Biomolecular Research and the GlaxoWellcome research group in Britain. The research group led by Valerie Gillet developed software to guide the design of a combinatorial library (Gillet, Willett, Bradshaw, & Green, 1999). A computer program called SELECT employed a genetic algorithm to make decisions about how best to design a diverse library (see the next section and "Conclusions about Informatics in Combinatorial Chemistry" for more on SELECT). SMILES-derived DCI fingerprints were the molecular descriptors on which the diversity of the library was evaluated.

Informatics in Library Design

Journal articles on the subject of library design did not appear prior to 1995. The literature in combinatorial chemistry beginning with Furka's 1982 paper and continuing on to 1995 primarily dealt with the details of chemical synthesis, enhancing reaction yields and such. Chemists reported ever more new ways to generate increasingly large libraries. It became evident that designing a library, not just generating it, would render the search through combinatorial space more productive than it was at the time. After sorting through the laboratory issues of synthesis, chemists gave attention to the question of library design and the informatics issues that accompany it.

Approaches to designing chemical libraries varied. No matter what the particular approach, informatics guided the choices chemists made for designing the library. Combinatorial libraries were challenging to design because they potentially encompassed tens of thousands, even millions, of molecules. By the 1990s, when the laboratory issues associated with combinatorial chemistry were largely resolved, chemists began to consider all *possible* molecules in their search domain, not just all *existing* molecules. A domain as vast as that required some selection criteria to narrow the search field and thus avoid wasting time and resources fruitlessly synthesizing and analyzing molecules with no biological or therapeutic activity.

Medicinal chemists have long recognized that beneficial drugs come from diverse sources. Modern-day researchers reemphasized the importance of molecular diversity in the combinatorial search for new drugs (Moos & Green, 1993). Diversity of molecules was determined by the conventional laboratory methods used to identify molecular structure. No formalized definition of diversity was apparent in the accompanying literature on molecular diversity. My own interpretation of the literature leads me to believe that generating molecular diversity simply meant to synthesize molecules with unique structural formulas, no matter how subtle the differences were. That being the case, diversity was simply a characteristic of molecules. An entire body of literature already existed discussing the use of informatics to gauge molecular diversity, based on either structural diversity or property diversity. Chemists used QSAR and quantitative structure-property relationship (QSPR) analyses to judge both similarity and diversity of molecules in a collection or library. It was not until 1995 when a shift in the literature appeared, with the term *diversity* applied to libraries of molecules rather than individual molecules. This shift marked the first application of informatics to library design.

A number of library design strategies have evolved in the last decade, each with strengths suited for a specific combinatorial application. One of the first to appear in the literature came out of the Chiron Corporation research group of Emeryville, California (Martin et al., 1995). The group's goal was to design a library of biologically active large molecules known as N-substituted glycine peptoids, or NSGs, which were thought to be promising antibiotics. NSGs have a three-part structure: a backbone, a side chain, and an end cap. Chiron's aim was to design a diverse library of NSGs using 721 different molecules for possible side chains and 1,133 other molecules for end caps, on as many as 160,000 different backbone molecules. The number of all possible permutations of these building blocks went well into the trillions, an impractical number to synthesize completely. They sought to select a subset of the entire set of possible NSGs, selecting a practical number of them to represent a maximally diverse group.

Chiron's chemists sought diversity among the proposed molecules in the library by analyzing the structural features of the reagent molecules. The details are well beyond the scope of this paper, but some key elements to the technique deserve attention. First, the team identified eighteen descriptors of the molecules used as side chains and end caps. These descriptors included such chemical properties as lipophilicity, molecular shape, molecular weight, and number of elements. Other descriptors included the DCI fingerprints described earlier, and the "receptor" qualities of the molecules, that is, the part of their chemical structure that allows them to bind with other molecules.

With descriptors chosen and numeric descriptor values identified for all the starting molecules, the diversity analysis began. The team used a type of statistical regression called multidimensional scaling (MDS) (Catterjee & Price, 1977), a similarity measure known as the Tanimoto coefficient, and a matrix-based method known as D-optimal design (Federov, 1972). None of these methods originated with chemical applications in mind, but chemists appropriated them to serve their needs. To visualize the results of the diversity analysis, the team used a special graphing technique known as a "flower pot representation" adapted from the field of molecular graphics (Connolly, 1993). The importance of the Chiron study lay in the fact that it demonstrated how mere visual inspection of different molecular structures by a chemist did not fully reveal the diversity of the molecules (diversity in the pre-1995 sense of the word). For example, two molecules with identical structures except for the absence of one atom in the second molecule were shown to differ significantly in six of the eighteen descriptors. Further, the study provided a technique for bringing the scope of the combinatorial search process down to a manageable scale; in the NSG example the team could expect to find maximum diversity within a practically sized subset of the original set of trillions of possible molecules. The Chiron team admitted that this initial design effort required validation but that these computation techniques borrowed from informatics would govern the future of library design (Martin et al., 1995).

An alternative approach to library design sought diversity in the product space in contrast to seeking it in the reactant space, as in the Chiron example. Consider the especially challenging nature of this task: one sought diversity in a set of hundreds of thousands of virtual molecules for which no laboratory reference data existed. Genetic algorithms were developed in the mid-1970s to deal with problems of this magnitude (Holland, 1975).

A genetic algorithm is a stochastic search process patterned after Darwinian evolution. The computational technique has the data equivalents of chromosomes, genes, such genetic operators as crossover and mutation, and a fitness function to evaluate which members of the population "survive" to the next generation. The algorithm takes a starting population of molecules through several generations of "evolution," matching the new offspring for "fitness" against a target compound (perhaps a known biologically active molecule), and weeding out those genetic variations that stray from the target. The program runs through the process iteratively until it converges on a population of molecules carrying the properties of the target molecule. While the final population ought to share similarities with a target molecule, they ought to be diverse with respect to each other.

Chemists first started using genetic algorithms to solve large-scale search problems in the early 1990s (Ugi, Fontain, & Bauer, 1990; Fontain, 1992). These first applications sought reaction mechanisms to achieve molecular diversity but did not design libraries to be diverse. Genetic algorithms for designing diverse libraries first appeared around 1995, in the work of Robert Sheridan and Simon Kearsley of Merck Research Laboratories in Rahway, New Jersey, and Lutz Weber's group at Hoffmann–La Roche in Basel (Sheridan & Kearsley, 1995; Weber, Wallbaum, Broger, & Gubernator, 1995).

After these initial attempts at applying genetic algorithms to the problem of library design, a number of software projects appeared that refined the technique. Among them were the GALOPED project of Robert Brown and Yvonne Martin at Abbott Labs in Illinois; HARPick, developed by Andrew Good and Richard Lewis of Rhone-Poulenc Rorer in Essex, England; and the SELECT program of Valerie Gillet's group at the Krebs Institute for Biomolecular Research at the University of Sheffield, England (Brown & Martin, 1997; Good & Lewis, 1997; Gillet et al., 1999).

Each of these design methods employed a similarity analysis measure to evaluate the prospective library's diversity, based on the known or calculated properties and biological activity of either the reactants or the products. Correlating the biological activity of a molecule to its chemical structure is known as quantitative structure-activity relationship analysis, or QSAR. As an informatics tool QSAR played a part both in library design and in the next phase of the drug discovery process, screening the library for likely drug candidates.

Informatics in Combinatorial Library Screening

After a research team designed and synthesized a combinatorial library, all the members of the library still had to be screened to identify "hits" and "leads." A hit is a molecule that has desirable structural features but has not yet demonstrated the desired biological activity. A hit becomes a lead if further chemical and biological analysis reveals that it possesses suitable therapeutic characteristics. Screening became a bottleneck in the drug discovery process as chemists synthesized increasingly

large libraries; screening techniques were suited to small numbers of samples at a time. In large part advances in robotics and automation relieved the bottleneck through the advent of high-throughput screening (HTS) and ultrahigh-throughput screening (uHTS). These laboratory methods made it plausible to screen both chemically and biologically hundreds of thousands of molecules in a library in a matter of days. As an alternative virtual screening techniques have arisen, where the members of the library are screened *in silico* rather than in the well plate.

Virtual screening techniques take several forms, one of which is to pass the library through a computational "filter," which reduces the number of compounds to be screened. A first example was the REOS method (rapid elimination of swill) developed by a team at Vertex Corporation in Cambridge, Massachusetts (Walters, Stahl, & Murcko, 1998). Certain chemical functional groups were known to be correlated with toxicity in humans, and so it was prudent to remove these from consideration in the screening process. The REOS program searched a library whose members were indexed in SMILES format and filtered them based on five chemical functional groups. When used in conjunction with a second set of filtering criteria and applied to some large libraries, the pool of hits shrank to fewer than one half of one percent of the original library (Walters & Murcko, 2000).

A second application of informatics to library screening is a technique we have already seen—genetic algorithms. Just as these algorithms were applied to the design of prospective libraries, they were also employed to search for leads in existing libraries or simply to identify a manageable subset of a proposed library for screening. One of the first cases reported was the work of Jasbir Singh's group at the Sterling Winthrop Pharmaceutical Research Division of Eastman Kodak, now part of Sanofi Winthrop, in Collegeville, Pennsylvania (Singh et al., 1996).

Cluster analysis is another example of an information-based approach to library screening. It is based on the representation of molecules in a multidimensional space, as many dimensions as the number of properties selected to characterize the molecules under study. When the members of a library are "plotted" in the multidimensional space, clusters are sought with "coordinates" in the space that are similar to the coordinates of known active drugs. Those members of the library clustered near a target coordinate are selected for further analysis and optimization. Clustering as a tool for similarity analysis was discussed by R. Jarvis and E. Patrick in the early 1970s, with Peter Willett adapting it for chemical analysis in the 1980s (Willett, 1987).

Conclusions about Informatics in Combinatorial Chemistry

What can we learn from examining the roles that informatics played in the field of combinatorial chemistry? What do these examples say about the practice of chemistry generally and its progressive transformation into an information-based science? I will answer these questions by discussing the relationship of informatics to conventional laboratory-based methods and by looking at the importance of collaborations among professionals of heterogeneous backgrounds in the birth of chemical informatics.

It is evident that combinatorial chemistry began as a laboratory science, not an information science, where the phrase *laboratory science* implies that the chemist spent his or her time in a traditional lab setting, working with real chemicals, reaction vessels, and chemical analysis technologies. I indicated earlier that the combichem literature is dominated early on by reports of improving the efficacy of combinatorial reaction schemes or finding new ones suited to specific targets of research. Virtually no mention is made of the use of or need for informatics tools in combichem prior to 1992, a decade after the field is said to have been invented. The examples of SMILES line notation and QSAR methods, developed before 1992, were not expressly discussed in the literature with combichem until that year. Such technological developments as automation of combinatorial synthesis, HTS in its many forms, and fluorescent- or radiolabeled identification preceded the adoption of informatics techniques by combinatorial chemists.

Assured that combinatorial synthesis was plausible, chemists began generating immense libraries, only to be confronted with new difficulties. Through the early 1990s library design and construction suffered from a lack of direction; "synthesis for its own sake" seemed to be the motto. In the pharmaceutical industry especially chemists faced a dilemma. The promise of combinatorial chemistry—which opened up the pharmaceutical search space far beyond the number of naturally occurring molecules—was not being fulfilled. The number of new drugs entering the market did not increase relative to the pre-combichem era. John Chabala (1998) of Pharmacopeia Inc. reported that in spite of a tenfold increase in pharmaceutical R&D spending from 1976 to 1994, the number of new pharmaceuticals entering the market annually had not increased since the advent of

combichem and HTS. There were a number of contributing factors here (e.g., bottlenecks in clinical trials and Food and Drug Administration regulations), but clearly combinatorial libraries in and of themselves did not provide an automatic increase in productivity and profit for the pharmaceutical industry. Thus we saw changing strategies in drug discovery, from randomly generated libraries to rationally designed ones, and more recently a movement from large, maximally diverse libraries to smaller "drug-like" libraries. Informatics guided each of these shifts.

I want to expand on this last remark by noting a change in the way that designers of chemical libraries selected design criteria. In early examples of library design the criteria chosen for QSAR studies of similarity were either chemical or biological in nature. Research groups tended to fall into one of these divisions. From the chemistry side the tendency was to develop large libraries, since the possible chemical variations were bountiful in number. However, an important piece of the puzzle was neglected in this approach, namely, that such biological considerations as oral bioavailability and toxicity were not included in the design. From a medicinal biologist's point of view strictly chemical QSAR was not as important as biological and therapeutic activity. In the early years of combinatorial chemistry these interests were not taken into account together. The example of the REOS project at Vertex represents a turning point in drug screening, where biological and chemical influences were given simultaneous consideration.

The difficulty of merging chemical and biological considerations in combinatorial library design and screening was discussed in a recent article by Corwin Hansch and his colleagues. Hansch is recognized as a founder of QSAR analysis. He reflected on the difficulties in information management that presented themselves in 1962 and that linger today:

> It has been a struggle to understand how to commence the development of a science of chemical-biological interactions. . . . A start on this problem has been made by creating a database of over 17000 QSAR of which 8500 pertain to biological systems and 8600 are from mechanistic organic chemistry. This has not been an easy task, even for the development of simple QSAR from mechanistic organic chemistry, since there is no simplified method to collect such data! This illustrates the crux of the problem facing information science. (Hansch, Hoekman, Leo, Weininger, & Selassie, 2002, p. 783)

The crux of the problem is rooted in the fact that these two aspects of scientific research have distinct and separate histories and practices and that a clear and uniform communication stream between them has not existed. Hansch's goal of trying to compile a comprehensive database of chemical and biological QSAR is one of the most daunting informatics tasks that chemists and biochemists face.

The issue of collaboration among researchers from heterogeneous disciplines is a larger theme that comes out of my study. Each case I cited above illustrated how chemists collaborated with software engineers, biochemists, or information scientists to make their respective contribution to combinatorial chemistry. For example, the developers of SMILES included a water chemist (Weininger), a computational QSAR expert (Veith), and a software engineer (Anderson). The British team that developed the SELECT algorithm included information scientists from the University of Sheffield (Valerie Gillet and Peter Willett) and a computational chemist from GlaxoWellcome (David V. S. Green). Collaborations like these are not uncommon in the literature and are indicative of the hybrid nature of combinatorial chemistry.

These examples highlight a multifaceted relationship between academia, government, and industry in this period. In some cases developments in chemical informatics came solely out of one domain or the other, while in other cases the developments came out of collaborations among these domains. In nearly all cases industry provided the resources for the advancement or promotion of the software development or combinatorial strategy. Recall, for example, that the SMILES software was created out of research performed by David Weininger and others at the EPA's research station in Duluth, Minnesota. When Weininger founded DCI, the company's marketing efforts offered SMILES wider use and acceptance in the chemistry community. Through its collaboration with other companies such as Parke-Davis, DCI developed new software that complemented and enhanced the functionality of SMILES. It is not evident that the SMILES software would have found such widespread use had it remained the property of its cocreators in Duluth. Pharmaceutical companies relied on their academic partners for specialized expertise in areas in which they did not have established R&D departments for informatics. Later, established pharmaceutical companies often added such departments, either hiring their personnel from academia or acquiring existing software firms, while start-up combichem

companies included informatics research divisions from the outset. In such cases as Chiron the company's cofounders, Pablo Valenzuela, William Rutter, and Edward Penhoet, came directly from academic posts. The 1980s and much of the 1990s were a time of immense entrepreneurial growth in the pharmaceutical and related industries. It was in this time frame that entrepreneurs such as Weininger and the three cofounders of Chiron built their companies into successful and competitive enterprises.

Acknowledgments

The author gratefully acknowledges the Chemical Heritage Foundation and the Eugene Garfield Foundation for their support and thanks the anonymous referees for their remarks on earlier versions of this paper.

References

Anderson, E., Veith, G. D., & Weininger, D. (1987). *SMILES: A line notation and computerized interpreter for chemical structures.* U.S. Environmental Protection Agency, Environmental Research Brief EPA/600/M-87/021. Duluth, Minnesota: U.S. Environmental Protection Agency.

Brown, F.. (1998). Chemoinformatics: What is it and how does it impact drug discovery? *Annual Reports in Medicinal Chemistry, 33,* 375–384.

Brown, R. D., & Martin, Y. C. (1997). Designing combinatorial library mixtures using a genetic algorithm. *Journal of Medicinal Chemistry, 40,* 2,304–2,313.

Catterjee, S., & Price, B. (1977). *Regression analysis by example.* New York: John Wiley and Sons.

Chabala, J. C. (1998). Historical overview of the developing field of molecular diversity. In E. M. Gordon & J. F. Kerwin, Jr. (Eds.), *Combinatorial chemistry and molecular diversity in drug discovery* (pp. 3–15). New York: Wiley-Liss.

Connolly, M. L. (1993). The molecular surface package. *Journal of Molecular Graphics, 11,* 139–141.

Federov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press.

Fontain, E. (1992). The problem of atom-to-atom mapping: An application of genetic algorithms. *Analytica Chimica Acta, 265,* 227–232.

Furka, A. (1982). Study on possibilities of systematic searching for pharmaceutically useful peptides. Unpublished paper, notarized 15 June 1982. Available in English translation: http://szerves.chem.elte.hu/Furka/82Eng.htm (accessed March 2004).

Furka, A. (2002). Combinatorial chemistry page of Arpad Furka. Available: http://szerves.chem.elte.hu/Furka/ (accessed June 2002).

Gillet, V. J., Willett, P., Bradshaw, J., & Green, D. V. S. (1999). Selecting combinatorial libraries to optimize diversity and physical properties. *Journal of Chemical Information and Computer Science, 39,* 169–177.

Good, A. C., & Lewis, R. A. (1997). New methodology for profiling combinatorial libraries and screening sets. *Journal of Medicinal Chemistry, 40,* 3,926–3,936.

Hansch, C., Hoekman, D., Leo, A., Weininger, D., & Selassie, C. D. (2002). Chem-bioinformatics: Comparative QSAR at the interface between chemistry and biology. *Chemical Reviews, 102,* 783–812.

Holland, J. H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor: University of Michigan Press.

Lam, K. S., Salmon, S. E., Hersh, E. M., Hruby, V. J., Kazmierski, W. M., & Knapp, R. J. (1991). A new type of synthetic peptide library for identifying ligand-binding activity. *Nature, 354,* 82–84.

Lenoir, T. (1999). Shaping biomedicine as an information science. In M. E. Bowden, T. B. Hahn, & R. V. Williams (Eds.), *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems* (pp. 27–45). Medford, NJ: Information Today.

Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., & Moos, W. H. (1995). Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *Journal of Medicinal Chemistry, 38,* 1431–1436.

Merrifield, R. B. (1965). Automated synthesis of peptides. *Science, 150,* 178–185.

Moos, W. H., & Green, G. D. (1993). Recent advances in the generation of molecular diversity. *Annual Reports in Medicinal Chemistry, 28,* 315–324.

National Institutes of Health. (2000). Biomedical Information Science and Technology Initiative (BISTI). Available: http://www.bisti.nih.gov (accessed March 2004).

Seneci, P. (2000). *Solid phase synthesis and combinatorial technologies*. New York: John Wiley & Sons.

Shemetulskis, N. E., Weininger, D., Blankley, C. J., Yang, J. J., & Humblet, C. (1996). Stigmata: An algorithm to determine structural commonalities in diverse datasets. *Journal of Chemical Information and Computer Science, 36,* 862–871.

Sheridan, R. P., & Kearsley, S. K. (1995). Using a genetic algorithm to suggest combinatorial libraries. *Journal of Chemical Information and Computer Science, 35,* 310–320.

Singh, J., et al. (1996). Application of genetic algorithms to combinatorial synthesis: A computational approach to lead identification and lead optimization. *Journal of the American Chemical Society, 118,* 1669–1676.

Ugi, I., Fontain, E., & Bauer, J. (1990). Transparent formal methods for reducing the combinatorial abundance of conceivable solutions to a chemical problem. *Analytica Chimica Acta, 235,* 155–161.

Walters, W. P., & Murcko, M. A. (2000). Library filtering systems and prediction of drug-like properties. In H.-J. Boehm & G. Schneider (Eds.), *Virtual screening for bioactive molecules* (pp. 15–32). New York:Wiley-VCH.

Walters, W. P., Stahl, M. T., & Murcko, M. A. (1998). Virtual screening—an overview. *Drug Discovery Today, 3,* 160–178.

Weber, L., Wallbaum, S., Broger, C., & Gubernator, K. (1995). Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. *Angewandte Chemie (International Edition in English), 34,* 2,280–2,282.

Weininger, D. J. (1988). SMILES, a chemical language and information system. *Journal of Chemical Information and Computer Science, 28,* 31–36.

Willett, P. (1987). *Similarity and clustering methods in chemical information systems.* Letchworth: Research Studies Press.

Wiswesser, W. J. (1954). *A line-formula chemical notation*. New York: Thomas Y. Crowell.

Wiswesser, W. J. (1968). 107 years of line-formula notations (1861–1968). *Journal of Chemical Documentation, 8,* 146–150.