Journal of Cheminformatics

COMMENT

Open Access

Searching chemical databases in the pre-history of cheminformatics

Peter Willett[1*]

Abstract

This article highlights research from the last century that has provided the basis for the searching techniques that are used in present-day cheminformatics systems, and thus provides an acknowledgement of the contributions made by early pioneers in the field.

Keywords Searching chemical databases, Early research articles, History of science

Introduction

This year sees the 15th anniversary of the founding of this journal in 2009, but many of the algorithms and methods that underlie modern cheminformatics date from very much earlier. This commentary seeks to highlight some of the more important contributions from the past that have provided the basis for techniques that are used today. The term 'cheminformatics' (and the alternative 'chemoinformatics') first appeared at the end of the 1990s [1, 2], with the first university MSc programmes appearing in 2000 [3]. It hence seems appropriate for the 'pre-history' discussed here to be based on articles that appeared in the research literature prior to the start of the present century, a rich literature that may by now be often forgotten. Specifically, a brief overview is provided of research that focused on ways of searching the files of machine-readable chemical structures that started to become available in the Fifties and Sixties, with Chen [4] and Willett [5] providing more extensive accounts of the historical development of cheminformatics.

*Correspondence:
Peter Willett
pwillettbjmt@gmail.com
[1] Sheffield, UK

Main text

Early chemical database systems normally represented their constituent structures by linear notations of various sorts but these were increasingly replaced by connection tables. Ray and Kirsch [6] recognised that the latter could be regarded as labelled graphs and that substructure searching could hence be implemented by means of a subgraph isomorphism algorithm. This approach was taken up by workers at Chemical Abstracts Service (CAS) [7] who were then working on what became the first version of the CAS Registry System [8]. A vital component of this project was the concept of extended connectivity as described in Morgan's much cited article [9]. The basic algorithm was devised to provide a unique molecular code but it has since contributed to a range of graph-based procedures in both cheminformatics and other fields [10].

Subgraph isomorphism algorithms provided effective means for conducting substructure searches but even efficient implementations [11, 12] proved to be far too slow in operation for use with large databases, and the searching of such files only became possible with the development of screening systems in the Seventies. These encoded substructural fragments in binary-vector fingerprints, the matching of which drastically reduced the numbers of structures that needed to undergo the detailed, but time-consuming, subgraph isomorphism search. Systematic procedures for the design of such

fragment-based systems were described by Adamson et al. [13] and by Hodes [14], with both in-house and public substructure searching systems well established by the start of the Eighties. Some years, however, were to pass before it became possible to augment these systems with facilities for similarity searching that enabled the identification of those database molecules that were structurally most similar to an input query molecule. That a similarity in structure often reflects a similarity in activities or properties, the so-called Similar Property Principle, had been known for many years [15, 16] so searching with a known active query could suggest new molecules for biological screening. Studies in the mid-Eighties [17, 18] demonstrated that substructural fragments such as those used for screening substructure searches could also be used to quantify the degree of resemblance between pairs of molecules, and operational systems were rapidly adopted as an effective and an efficient method for similarity-based virtual screening. Similarity measures such as those developed for searching purposes have subsequently found application in a wide range of cheminformatics applications, such as property prediction, computer-aided synthesis design, database clustering and molecular diversity analysis [19, 20].

More sophisticated search algorithms are required for substructure searching in databases of chemical patents, where molecules are defined by generic, or Markush, structures, each of which encompasses not just a single molecule but many, or extremely many, related variants. Early attempts to address this problem were based on fragment codes of various types [21] but these were supplanted by graph-based approaches that started to appear in the late Eighties from Télésystèmes-Questel and Derwent [22] and from CAS [23], with the most extensive descriptions coming from a large-scale research programme conducted by Lynch et al. in collaboration with both Derwent and CAS *inter alia* that developed connection table, screening and isomorphism procedures that were able to encompass the much greater complexity of generic structures [24].

Specialized techniques are also required if searches are to be carried out on databases of chemical reactions. Here, there is the need to encode and to search not just the reactants and products of a reaction but also the reaction centres, i.e., the parts of the molecules where some sort of substructural change had taken place as a result of the reaction. The automatic identification of these changes was first suggested by Vleduts (in an article that also discussed how a computer could assist in the design of novel synthetic pathways) [25]. He subsequently described how the changes could be implemented by means of a maximum common subgraph isomorphism algorithm [26], an approach that formed the basis for the reaction searching systems that started to appear in the early Eighties [27].

The work discussed thus far had all involved 2D representations of molecular structure, but work by Gund in the early Seventies demonstrated that graphs based on inter-atomic distances could be searched for pharmacophoric patterns in just the same way as could patterns of atoms and bonds [28, 29]. At this point in the Seventies, the only 3D structures that were widely available were those in the Cambridge Structural Database and the matching operations in Gund's work were far too slow for database applications. The first limitation was overcome with the introduction in 1987 of CONCORD and CORINA, structure generation programs that were sufficiently rapid in operation to enable the conversion of existing 2D databases to 3D form [30, 31]. The second limitation was addressed by the extension of conventional screening and subgraph isomorphism procedures to enable the processing of distance-based graphs, hence forming the basis for the first operational 3D substructure searching systems at Pfizer and Lederle in the late Eighties [32, 33]. Subsequent work took account of the fact that many molecules are not rigid but contain rotatable bonds, hence enabling the development of a range of flexible searching systems [34]; in like vein, the initial work on ligand docking by Kuntz et al. in the early Eighties [35] was followed by extensions to encompass ligand flexibility in the following decade.

Conclusions

This commentary has considered early work in just one aspect of cheminformatics, but there was much important pre-historic work in other areas: for example, studies of pattern recognition in the Seventies [e.g., 36–38] can now be seen as foreshadowing the current intense interest in chemical applications of machine learning. It's hence hoped that this brief article will serve two functions. First, to acknowledge the contributions of early pioneers that have helped to lay the foundations of modern cheminformatics, the "standing-on-the-shoulders-of-giants" phenomenon commonly attributed to Newton [39]; and, second, to spur others to study and then to document the history not just of searching techniques but also of the many other aspects of cheminformatics as it comes to play an increasingly important role across the whole spectrum of chemical science.

Acknowledgements
Not applicable.

Author contributions
This submission is entirely my work.

Funding
Not applicable.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Competing interests
The authors declare no competing interests.

Received: 14 September 2024 Accepted: 20 October 2024
Published online: 04 November 2024

References

1. Warr WA (1999) What is cheminformatics? Paper presented at the 218th ACS National Meeting, New Orleans, 22–26 August 1999. At https://www.warr.com/warrzone2000.html (accessed 4 September 2024)
2. Hann M, Green R (1999) Chemoinformatics—a new name for an old problem? Curr Opinion Chem Biol 3:379–383
3. Wild DJ, Wiggins GD (2006) Challenges for chemoinformatics education in drug discovery. Drug Discov Today 11:436–439
4. Chen WL (2006) Chemoinformatics: past, present and future. J Chem Inf Model 46:2230–2255
5. Willett P (2008) From chemical documentation to chemoinformatics: fifty years of chemical information science. J Inf Sci 34:477–499
6. Ray LC, Kirsch RA (1957) Finding chemical records by digital computers. Science 126(378):814–819
7. Cossum WE, Krakiwsky LMF (1965) Advances in automatic chemical substructure searching. J Chem Doc 5:33–35
8. Leiter DP, Morgan HL, Stobaugh RE (1965) Installation and operation of a registry for chemical compounds. J Chem Doc 5:238–242
9. Morgan HL (1965) The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. J Chem Doc 5:107–113
10. Al Jishi R, Willett P (2010) The *Journal of Chemical Documentation* and the *Journal of Chemical Information and Computer Sciences*: Publication and citation statistics. J Chem Inf Model 50:1915–1923
11. Sussenguth EH (1965) A graph-theoretic algorithm for matching chemical structures. J Chem Doc 5:36–43
12. Ullmann JH (1976) An algorithm for subgraph isomorphism. J Assoc Comput Mach 23:31–42
13. Adamson GW, Cowell J, Lynch MF, McLure AHW, Town WG, Yapp AM (1973) Strategic considerations in the design of screening systems for substructure searches of chemical structure files. J Chem Doc 13:153–157
14. Hodes L (1976) Selection of descriptors according to discrimination and redundancy—application to chemical-structure searching. J Chem Inf Comput Sci 16:88–93
15. Adamson GW, Bush JA (1973) A method for the automatic classification of chemical structures. Inf Stor Retriev 9:561–568
16. Wilkins CL, Randic M (1980) A graph theoretical approach to structure–property and structure–activity correlation. Theoret Chimica Acta 58:45–68
17. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular-features in structure activity studies—definition and applications. J Chem Inf Comput Sci 25:64–73
18. Willett P, Winterman V, Bawden D (1986) Implementation of nearest-neighbour searching in an online chemical structure search system. J Chem Inf Comput Sci 26:36–41
19. Johnson MA, Maggiora GM (eds) (1990) Concepts and applications of molecular similarity. John Wiley, New York
20. Willett P (2016) Molecular similarity approaches in chemoinformatics: early history and literature status. ACS Symposium Ser 1222:67–89
21. Barnard JM (ed) (1984) Computer handling of generic chemical structures. Gower, Aldershot
22. Shenton KE, Norton P, Ferns EA (1988) Generic searching of patent information. In: Warr WE (ed) Chemical structures: the international language of chemistry. Springer, Berlin
23. Fisanick W (1990) The Chemical Abstracts Service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts. J Chem Inf Comput Sci 30:145–154
24. Lynch MF, Holliday JD (1996) The Sheffield Generic Structures project: a retrospective review. J Chem Inf Comput Sci 36:930–936
25. Vleduts GE (1963) Concerning one system of classification and codification of organic reactions. Inf Stor Retriev 1:117–146
26. Vleduts GE (1977) Development of a combined WLN/CTR multilevel approach to the algorithmical analysis of chemical reactions in view of their automatic indexing. British Library Research and Development Department Report, London
27. Willett P (ed) (1986) Modern approaches to chemical reaction searching. Gower, Aldershot
28. Gund P, Wipke WT, Langridge R (1974) Computer searching of a molecular structure file for pharmacophoric patterns. In: International Conference on Computers in Chemical Research and Education. Elsevier, Amsterdam
29. Gund P (1977) Three-dimensional pharmacophoric pattern searching. Progress Mol Subcellular Biol 5:117–143
30. Pearlman RS (1987) Rapid generation of high-quality approximate 3D molecular structures. Chem Design Automat News 2:1–7
31. Hiller C, Gasteiger J (1987) Ein automatisierter Molekülbaukasten. In: Gasteiger J (ed) Software Entwicklung in der Chemie. Springer, Berlin
32. Jakes SE, Watts N, Willett P, Bawden D, Fisher JD (1987) Pharmacophoric pattern matching in files of 3D chemical structures: evaluation of search performance. J Mol Graph 5:41–48
33. Sheridan RP, Nilakantan R, Rusinko A, Bauman N, Haraki KS, Venkataraghavan R (1989) 3DSEARCH: a system for three-dimensional substructure searching. J Chem Inf 29:255–260
34. Martin YC, Willett P (eds) (1998) Designing bioactive molecules: three-dimensional techniques and applications. American Chemical Society, Washington
35. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288
36. Kowalski BR, Bender CF (1972) Pattern recognition. A powerful approach to interpreting chemical data. J Am Chem Soc 94:5632–5639
37. Cramer RD, Redl G, Berkoff CE (1974) Substructural analysis. A novel approach to the problem of drug design. J Med Chem 17:533–535
38. Stuper AJ, Jurs PC (1975) Classification of psychotropic drugs as sedatives or tranquilizers using pattern recognition techniques. J Am Chem Soc 97:182–187
39. Wikipedia (2024) Standing on the shoulders of giants. At https://en.wikipedia.org/Wiki/Standing-on-the-shoulders-of-giants (accessed 4 September 2024)

Publisher's Note