Proceedings

of the

International Workshop

Chemical Data Analysis in the Large:
The Challenge of the Automation Age

May 22nd-26th, 2000, Bozen, Italy

Martin G. Hicks
(Editor)

Beilstein-Institut zur Förderung der Chemischen Wissenschaften

Trakehner Str. 7 – 9
60487 Frankfurt
Germany

Telephone: +49 (0) 69 7167 3211
Fax: +49 (0) 69 7167 3219
E-mail: info@beilstein-institut.de
Web-page: www.beilstein-institut.de

IMPRESSUM

Chemical Data Analysis in the Large: The Challenge of the Automation Age, Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd - 26th, 2000, Bozen, Italy.

Bibliographic Information of *Die Deutsche Bibliothek*

Die Deutsche Bibliothek notes this publication in the *Deutsche Nationalbibliographie*; details are available at http://dnb.ddb.de.

ISBN 3-89722-938-2

Printed by: Logos Verlag Berlin
Comeniushof, Gubener Str. 47
10243 Berlin
Tel.: +49 (0) 30 42 85 10 90
Fax: +49 (0) 30 42 85 10 92
Internet: http://www.logos-verlag.de

Published in „Chemical Data Analysis in the Large: The Challenge of the Automation Age“, Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/impressum/impressum.pdf

PREFACE

Managing and effectively utilizing large collections of highly diverse chemical data – especially when chemical structures are involved – is a major challenge for the chemical and pharmaceutical industries and universities, as well as for the producers of large publicly available databases. Automated techniques, such as, combinatorial chemistry coupled with high throughput screening result in the routine generation of enormous amounts of data. Methods of information handling such as knowledge discovery and data mining, machine learning, statistical analysis, and visualization, whose origins lie outside chemistry, are becoming more and more applicable in the area of chemical sciences. The aim of this workshop was to bring together experts from chemical and non-chemical fields to discuss new and better methods for handling and analyzing large amounts of data of a chemical nature.

The remote location of Schloss Korb – set on a hillside overlooking Bozen/Bolzano – provided the ideal venue for the participants to spend time discussing issues of interest and to make contact with scientists from different disciplines. The format of these workshops, with ample time for discussion between the lectures and afterwards at lunch and dinner, provided the participants with something rarely found at larger meetings – time to think and time to talk.

Over three days we heard a series of invited talks, which covered the following areas:

Knowledge Discovery and Data Mining
Information Extraction and Text Mining
Data Compression and Clustering of Large Data Sets
Chemical Structure Representations
Structure Browsing and Similarity Indexes
Virtual Screening and Library Design
Property Prediction
Visualization of Data and Physicochemical Properties

The scientific program was compiled by Martin Hicks (Beilstein-Institut), Gerald Maggiora (Pharmacia) and Peter Willett (University of Sheffield).

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/preface/preface.pdf

The Beilstein-Institut organizes and sponsors scientific meetings, workshops and seminars, with the aim of catalyzing advances in chemical science by facilitating the interdisciplinary exchange and communication of ideas amongst the participants. We were very pleased that speakers from both inside and outside the mainstream chemical community accepted invitations to speak. The resonance that we had both during and after this workshop clearly reflected the attractiveness of the scientific program and the format of the workshop.

We would like to thank particularly the authors who provided us with written versions of the papers that they presented. Special thanks go to all those involved with the preparation and organization of the workshop, as well as to the speakers and participants for their contributions in making this workshop a success.

Werner Brich

Martin G. Hicks

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/preface/preface.pdf

CONTENTS

WEB SEARCH, FILTERING, AND TEXT MINING: TECHNOLOGY FOR A NEW ERA OF INFORMATION ACCESS

BRUCE CROFT

NSF Center for Intelligent Information Retrieval, Computer Science Department, University of Massachusetts, Amherst, MA 01003-4610 USA
E-mail: croft@cs.umass.edu

Received: 30th May 2000 / Published 11th May 2001

ABSTRACT

Much of the information in science, engineering and business has been recorded in the form of text. Traditionally, this information would appear in journals or company reports, but increasingly it can be found online in the World-Wide Web. Tools to support information access and discovery on the Internet are proliferating at an astonishing rate. Some of this development reflects real progress but there are also many exaggerated claims. The focus of this presentation will be to review the important technologies for text-based information access on the Web and to describe the progress that is being made by researchers in these areas.

INTRODUCTION

Ten years ago, the primary technologies being used to construct large information systems were database systems, information retrieval systems, and information filtering systems. Database systems were used to handle large volumes of structured data and to provide guarantees of reliability and consistency despite systems failures and high volumes of update transactions. Information retrieval systems were used to search large databases of text, such as scientific abstracts, legal materials, or newspaper stories. Information filtering or "clipping" services provided periodic updates in the form of text stories, mostly in the business domain, based on user profiles.

In the relatively short period since, there have been many developments that have affected how information technology is talked about and used. The most important of these have been the growth of the Internet and the availability of cheap hardware. The technologies for the large information systems discussed today include the Internet (and intranets and extranets), Web search, portals, agents, collaborative filtering, XML and metadata, and data mining.

There are many opinions about the current technology for information systems, including that everything is different, everything is the same, and everything is a mess. What people generally do agree on is that there is much more data on-line, much of that data is unstructured (i.e. text, image, video), and that the data is much more distributed than in the past. This statement is usually applied to the Web in general, but it also applies, with some reservations, to scientific information.

This paper provides a brief description of some of the new technologies and reviews their current status and future research directions.

SEARCH ENGINES

One of the major tools for information access is the search engine. Most search engines use information

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/croft/croft.pdf

retrieval techniques to rank Web pages in presumed order of relevance based on a simple query. Compared to the bibliographic information retrieval systems of the 70's and 80's, the new search engines must deal with information that is much more heterogeneous, "messy", more varied in quality, and vastly more distributed or "linked".

In the current Web environment, queries tend to be short (1-2 words) and the potential database is very large and growing rapidly. Estimates of the size of the Web range from 500 million to a billion pages, with many of these pages being portals to other databases (the "hidden Web").

In response to this huge expansion of potential information sources, today's Web search engines have emphasized speed and coverage, with less importance attached to effectiveness. With the growing number of complaints about "information overload", however, this is beginning to change. Similarly, most Web search engines use a centralized architecture where "Web crawlers" gather Web pages and a single, very large index is created. An approach like this has inherent scalability problems.

There has been a growing awareness that effective information retrieval is a hard problem. Indeed, in a recent Turing Award lecture, it was identified as a software "grand challenge". To address this challenge, researchers in information retrieval and related areas of computer science are proposing new retrieval models and techniques to support distributed architectures, summarization, question answering, cross-lingual retrieval, better interfaces, and multimodal search.

Retrieval models provide the underlying framework for a search engine. In other words, they are the basis for the algorithms that score and rank the Web pages. Recent developments in this area include ranking algorithms based on link structure (e.g. www.google.com) and language modeling. The algorithms based on link structure analyze link patterns to identify sites that are highly linked. This is similar to the citation analysis techniques developed in the 1970s for scientific articles. Probabilistic techniques based on language modeling are the basis of effective algorithms for a variety of language tasks, such as speech recognition and machine translation, and are beginning to demonstrate effectiveness improvements in large-scale experiments. This work is also being used in the development of cross-lingual techniques, where queries are given in one language and the results are found across a variety of other languages.

There has also been considerably more work recently that is applying natural language processing techniques to the problem of information retrieval. Much of this work is being done under the title of "question answering". The goal of this type of information access is to produce a concise answer to well-formulated queries. In the case of simple queries such as "What is the boiling point of water?" both the answer and the task are well-defined. For other questions such as "What is the best drug for treating high blood pressure?" the answer is much less well-defined and will probably require combining data from a variety of sources. Techniques for distributed retrieval and summarization will be part of the solution.

Researchers in the area of distributed search are developing techniques for identifying relevant information sources, describing their contents, and combining results from multiple searches. Summarization researchers are looking at ways of generating a variety of different types of summary for single documents and groups of documents. The summary types include lists of keywords, extracted sentences, and generated text. Visualization techniques and techniques for automatically generating taxonomies are also important.

One of the key aspects of improving the effectiveness of Web search involves getting better descriptions of the user's information need. A short one or two word query is generally not descriptive of the actual information need and is not helpful to the search engine. Techniques such as automatic query expansion and machine learning through relevance feedback have been developed to address this problem. The growing ubiquity of wireless devices is also leading to a new interest in voice interfaces, which bring a variety of new challenges and opportunities to the designer of Web search engines, including dealing with longer queries.

There has also been considerable research on multimedia and multimodal retrieval. Multimedia retrieval involves algorithms for representing and comparing image and video data. A number of promising techniques have been developed, but large scale experimentation has not been done except for some specialized tasks such as face retrieval. Multimodal retrieval involves frameworks for combining evidence from multiple sources, such as image and text, into overall estimates of relevance for complex objects.

XML/METADATA

XML is a new standard developed for Web page markup or, more generally, for describing the structure of data that is more loosely formatted than a standard database schema instance. It is related to the older SGML and HTML markup standards. There has been a considerable amount of publicity about XML and an increasing number of compatible tools are becoming available. The XML standards activity has also expanded to include the definition of ontologies for the description of document content in addition to the structure. MPEG7 has related aims for video data.

There is no doubt that these efforts on standardizing format and content through XML and metadata will have a large impact on future information systems. There is, however, less reason to believe that this approach will solve the access problem. Manual indexing using controlled vocabularies is one of the oldest methods of text representation that has been used in information retrieval systems. There is abundant evidence that this approach does not scale and, in general, is of limited effectiveness when used as the only representation. The most effective applications of this type of indexing are in limited domains with a substantial investment in ontology development, such as medicine (MESH) and chemistry (Chemical Abstracts). Developing XML-based ontologies in such domains would be useful if done in conjunction with content-based searching and categorization. Categorization is a technique for automatically assigning labels (or controlled vocabulary terms) to new documents. A considerable amount of research has been done in this area and, with enough training data, good results can be achieved. As more XML data described using ontologies and metadata becomes available, categorization techniques will become viable.

INFORMATION FILTERING

Information filtering has been around for some time in the form of "current awareness" systems. A number of Web tools provide this functionality (often under the "agent" label). Most of the applications of this technology are in the business and news domains. Many of these systems use simple Boolean matching techniques, although there has been much research and a number of new companies applying machine learning techniques to this problem. Effective filtering is, however, as difficult as effective search, and the problems involved with proactively sending too much data that is not relevant to the users have resulted in varying levels of acceptance. Many of the

techniques being developed to improve search, however, will also result in more effective filtering so we can expect to see more applications involving this technique in the future.

Collaborative filtering is a complementary technique based on matching user preferences that has become popular in e-commerce applications. It remains to be seen whether the combination of content-based and collaborative filtering will improve information access in scientific and engineering contexts.

TEXT DATA MINING

A considerable amount of research is being carried out under the heading of text data mining. This includes a variety of techniques such as information extraction, clustering, and discovery of associations or "rules". All of these techniques combine statistical methods with some level of linguistic analysis. In contrast to data mining using relational database systems, where a number of commercial packages are available, text data mining is still an open research issue. Evaluation of research in this area is also difficult, since many of the results are presented with examples instead of statistical data.

Information extraction techniques are designed to extract "facts" from text. In many cases, this means very simple facts such as names of companies, people, and monetary amounts, but in general this technique can be used to extract more complex information, such as filling a database according to a template or schema. Extraction is a key component of text data mining since it provides the objects for the statistical analysis. Much of the research in this area has been done with newspaper text, but results with scientific text are beginning to be reported. There has also been recent work focusing on information extraction based on the structure of Web pages.

Clustering is used to group related information. This technique has been well-studied in information retrieval but has recently been the subject of a number of new papers. Information extraction and clustering can be used with other techniques to discover interesting associations in text databases. The applications of this type of discovery have been mostly based on business information, but it may also be useful in scientific and engineering contexts.

"Literature-based discovery" is an interesting area of research that has been underway for some time and is an early example of text data mining. By analyzing the literatures of related fields for topics that are related but not connected by direct reference, Swanson and his colleagues at the University of Chicago have found a number of connections in the medical literature (specifically, Medline abstracts) that have been the subject of follow-up scientific investigations.

CONCLUSIONS

The Web is a huge, relatively unstructured and sometimes unreliable source of information. The development of XML and ontology standards for metadata will promote sharing and introduce a limited amount of structure to the Web, but they are not the whole solution to the information problem. Many new tools are being developed to exploit unstructured information and to make it more useful to specific user communities such as scientists. These tools can also be used for information access and discovery with scientific literature and databases. Techniques such as text data mining, however, will require considerably more research and experimentation before their effectiveness can be established.

Papers describing research in a number of these areas and extensive references to other papers can be found in [1].

REFERENCES AND NOTES

[1] Croft, W.B. (editor), *Advances in Information Retrieval*, Kluwer Academic Publishers, Boston, **2000**.

INFORMATION EXTRACTION FROM BIOLOGICAL SCIENCE JOURNAL ARTICLES: ENZYME INTERACTIONS AND PROTEIN STRUCTURES

ROBERT GAIZAUSKAS,* KEVIN HUMPHREYS AND GEORGE DEMETRIOU

Department, of Computer Science, University of Sheffield, Regent Court, Portobello Street Sheffield, Sl 4DP UK
E-mail: robertg@shef.ac.uk ; kwh@shef.ac.uk ; demetri@shef.ac.uk

Received: 28th June 2000 / Published 11th May 2001

ABSTRACT

With the explosive growth of scientific literature in the area of molecular biology, the need to process and extract, information automatically from on-line text sources has become increasingly important. Information extraction technology, as defined and developed through the U.S. DARPA Message Understanding Conferences (MUCs), has proved successful at extracting information primarily from news-wire texts and primarily in domains concerned with human activity. In this paper we consider the application of this technology to the extraction of information from scientific journal papers in the area of molecular biology. We describe how an information extraction system designed to participate in the MUC exercises has been modified for two bioinformatics applications: EMPathIE, concerned with enzyme and metabolic pathways; and PASTA, concerned with protein structure. The progress so far provides convincing grounds for believing that IE techniques will deliver novel and effective ways for the extraction of information from unstructured text sources in the pursuit of knowledge in the biological domain.

INTRODUCTION

Information Extraction (IE) may be defined as the activity of extracting details of predefined classes of entities and relationships from natural language texts and placing this information into a structured representation called a *template.* [1, 2] The prototypical IE tasks are those defined by the U.S. DARPA-sponsored Message Understanding Conferences (MUCs), requiring the filling of a complex template from newswire texts on subjects such as joint venture announcements, management succession events, or rocket launchings. [3, 4] While the performance of current technology is not yet at human levels overall, it is approaching human levels for some component tasks (e.g. the recognition and classification of named entities in text) and is at a level at which comparable technologies, such as information retrieval and machine translation, have found useful application. IE is particularly relevant where large volumes of text make human analysis infeasible, where template-oriented information seeking is appropriate (i.e. where there is a relatively stable information need and a set of texts in a relatively narrow domain), where conventional information retrieval technology is inadequate, and where some error can be tolerated.

One area where we believe these criteria are met, and where IE techniques have as yet been applied only in a limited way (though see [5-7]), is the construction of databases of scientific information

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/gaizauskas/gaizauskas.pdf

from journal articles for use by researchers in molecular biology. The explosive growth of textual material in this area means that no one can keep up with what is being published. Conventional retrieval technology returns both too little, because of the complex, non-standardized terminology in the area, and too much, because what is sought is not whole texts in which key terms appear, but facts buried in the texts. Further, useful templates can be defined for some scientific tasks. For example, scientists working on drug discovery have an ongoing interest in reactions catalyzed by enzymes in metabolic pathways. These reactions may be viewed as a class of events, like corporate management succession events, in which various classes of entities (enzymes, compounds) with attributes (names, concentrations) are related by participating in the event in specific roles (substrate; catalyst; product). Finally, some error can be tolerated in these applications, because scientists can verify the information against the source texts -the technology serves to assist, not to replace, investigation.

Thus, we believe automatically extracting information from scientific journal papers is an important and feasible application of IE techniques. It is also interesting from the perspective of IE research because it extends IE to *domains* and to *text genres* where it has never been applied before. To date most IE applications have been to domains of human activity, predominately economic activity, and have involved newswire texts that have a characteristic lexis, structure and length. Applying IE to scientific journal papers in the area of molecular biology means a radical shift of subject domain away from the world of people, companies, products and places that have largely figured in previous applications. It also means dealing with a text genre in which there is a vast and complex technical vocabulary, where the texts are structured into subsections dealing with method, results, and discussion, and where the texts are much longer. These differences pose tough challenges for IE techniques as developed so far: can they be applied successfully in this area?

In this paper we describe the use of the technology developed through MUC evaluations in two bioinformatics applications. The next section describes the general functionality of an IE system. We then describe the two specific applications on which we are working: extraction of information about enzymes and metabolic pathways and extraction of information about protein structure, in both cases from scientific abstracts and journal papers. The following section describes the principle processing stages and techniques of our system, followed by a section that presents evaluations of the system's performance. While much further refinement of the system for both applications is possible, indications are that IE can indeed be successfully applied to the task of extracting information from scientific journal papers.

INFORMATION EXTRACTION TECHNOLOGY

The most recent MUC evaluation (MUC-7, [4]) specified five separate component tasks, which illustrate the main functional capabilities of current IE systems:

1. *Named Entity recognition* requires the recognition and classification of named entities such as organizations, persons, locations, dates and monetary amounts.
2. *Coreference resolution* requires the identification of expressions in the text that refer to the same object, set or activity. These include variant forms of name expression *(Ford Motor Company ... Ford),* definite

noun phrases and their antecedents *(Ford ... the American car manufacturer),* and pronouns and their antecedents *(President Clinton .. he).* Coreference relations are only marked between certain syntactic classes of expressions (noun phrases and pronouns) and a relatively constrained class of relationships to mark is specified, with clarifications provided with respect to bound anaphors, apposition, predicate nominals, types and tokens, functions and function values, and metonymy.

3. *Template Element filling* requires the filling of small scale templates (slot-filler structures) for specified classes of entity in the texts, such as organizations, persons, certain artifacts, and locations, with slots such as name (plus name variants), description as supplied in the text, and subtype.
4. *Template Relation filling* requires filling a two slot template representing a binary relation with pointers to template elements standing in the relation. For example, a template relation of `employee_of` containing slots for a person and organization is filled whenever a text makes clear that a particular person is employed by a particular organization. Other relations are `product_of` and `location_of`.
5. *Scenario Template filling* requires the detection of relations between template elements as participants in a particular type of event, or scenario (rocket launches for MUC-7), and the construction of an object-oriented structure recording the entities and various details of the relation.

Systems are evaluated on each of these tasks as follows. Each task is precisely specified by means of a task definition document. Human annotators are then given these definitions and use them to produce by hand the 'correct' results for each of the tasks - filled templates or texts tagged with name classes or coreference relations (these results are called *answer keys).* The participating systems are then run and their results, called *system responses,* are automatically scored against the answer keys. Chief metrics are *precision* - percentage of the system's output that is correct (i.e. occurs in the answer key) - and *recall* - percentage of the correct answer that occurs in the system's output.

State-of-the-art (MUC-7) results for these five tasks are as follows (in the form recall/precision): named entity - 92/95; coreference - 56/69; template element - 86/87; template relation - 67/86; scenario template 42/65.

TWO BIOINFORMATICS APPLICATIONS OF INFORMATION EXTRACTION

We are currently investigating the use of IE for two separate bioinformatics research projects. The Enzyme and Metabolic Pathways Information Extraction (EMPathIE) project aims to extract details of enzyme reactions from articles in the journals *Biochimica et Biophysica Acta* and *FEMS Microbiology Letters.* The utility for biological researchers of a database of enzyme reactions lies in the ability to search for potential sequences of reactions, where the products of one reaction match the requirements of another. Such sequences form metabolic pathways, the identification of which can suggest potential sites for the application of drugs to affect a particular end result. Typically, journal articles in this domain describe details of a single enzyme reaction, often with little indication of related reactions and which pathways the reaction may be part of. Only by combining details from several articles can potential pathways be identified.

The Protein Active Site Template Acquisition

(PASTA) project aims to extract information concerning the roles of amino acids in protein molecules, and to create a database of protein active sites from both scientific journal abstracts and full articles. The motivation for the PASTA project stems from the need to extract and rationalize information in the protein structure literature. New protein structures are being reported at very high rates and the number of co-ordinate sets (currently about 12000) in the Protein Data Bank (PDB) [8] can be expected to increase ten-fold in the next five years. The full evaluation of the results of protein structure comparisons often requires the investigation of extensive literature references, to determine, for instance, whether an amino acid has been reported as present in a particular region of a protein, whether it is highly conserved, implicated in catalysis, and so on. When working with several different structures, it is frequently necessary to go through a large number of scientific articles in order to discover any functional or structural equivalences between residues or groups of residues. Computational methods that can extract information directly from these articles would be very useful to biologists in comparison classification work and to those engaged in modeling studies.

The following section describes the EMPathIE and PASTA tasks, including the intended extraction results from documents containing text such as that shown in Figure 1.

EMPathIE

One of the inspirations for the Enzyme and Metabolic Pathways application was the existence of a manually constructed database for the same application. The EMP database [9] contains over 20,000 records of enzyme reactions, collected from journal articles published since 1964. That such a database has been constructed and is widely used demonstrates the utility of the application. EMPathie aims to extract only a key subset of the fields found in the EMP database records.

Results: We have determined the crystal structure of a triacylglycerol lipase from Pseudomonas cepacia (Pet) in the absence of a bound inhibitor using X-ray crystallography. The structure shows the lipase to contain an alpha/betahydrolase fold and a catalytic triad comprising of residues Ser87, His286 and Asp264. The enzyme shares several structural features with homologous lipases from Pseudomonas glumae (PgL) and Chromobacterium viscosum (CvL), including a calcium-binding site. The present structure of Pet reveals a highly open conformation with a solvent-accessible active site. This is in contrast to the structures of PgL and Pet in which the active site is buried under a closed or partially opened 'lid', respectively.

Figure 1: Sample text fragment from a scientific paper in Molecular Biology

The main fields required in a record of an enzyme reaction are: the enzyme name, with an enzyme classification (EC) number, if available, the organism from which the enzyme was extracted, any known pathway in which the reaction occurs, compounds involved in the reaction, with their roles classified as either substrate (input), product (output), activator, inhibitor, cofactor or buffer, and any compounds known not to be involved in the reaction, with their roles classified as either non-substrate or non-product.

The template definitions include three Template Elements: *enzyme, organism* and *compound,* a single Template Relation: *source,* relating *enzyme* and *organism* elements, and a Scenario Template for the specific metabolic pathway task. The Scenario Template describes a pathway involving one or more interactions, each of which is a reaction between an enzyme and one or more participants, possibly under certain constraints. A manually produced sample Scenario Template is shown here, taken from an article on *isocitrate*

lyase activity in FEMS Microbiology Letters.

```
<ENZYME-1> :=
   NAME:  isocitrate lyase
   EC.CODE:  4.1.3.1

<ORGANISM-1> :=
   NAME: Haloferax volcanii
   STRAIN: ATCC 29605
   GENUS: halophilic Archaea

<COMPOUND-1> :=
   NAME: phenylhydrazone

<COMPOUND-2> :=
   NAME: KCl

<SOURCE-1> :=
   ENZYME: <ENZYME-1>
   ORGANISM: <ORGANISM-1>

<PATHWAY-1> :=
   NAME: glyoxylate cycle
   INTERACTION: <INTERACTION->

<INTERACTION-1> :=
   ENZYME: <ENZYME-1>
   PARTICIPANTS:  <PARTICIPANT-1>
                  <PARTICIPANT-2>

<PARTICIPANT-1> :=
  COMPOUND: <COMPOUND-1>
  TYPE: Product
  TEMPERATURE: 35C

<PARTICIPANT-2> :=
  COMPOUND: <COMPOUND-2>
  TYPE: Activator
  CONCENTRATION: 1.75 M
```

This template describes a single interaction found to be part of the metabolic pathway known as the glyoxylate cycle, where the interaction is between the enzyme isocitrate lyase and two other participants. The first participant is the compound glyoxylate phenylhydrazone, which has the role of a product of the interaction at a temperature of 35C. The second is the compound KCl, which has the role of an activator at a concentration of 1.75M.

The template design follows closely the MUC-style IE template, and is richer than the EMP database record format in terms of making relationships between entities explicit.

PASTA

The entities to be identified for the PASTA task include proteins, amino acid residues, species, types of structural characteristics (secondary structure, quaternary structure), active sites, other (probably less important) regions, chains and interactions (hydrogen bonds, disulphide bonds etc.) In collaboration with molecular biologists we have designed a template to capture protein structure information, a fragment of which, filled with information extracted from the text in Figure 1, is shown below:

```
<RESIDUE-str97-521>:=
  TYPE:                 SERINE
  NUMBER:               "87"
  PROTEIN:              <PROTEIN-str97-521>
  SITE/FUNCTION         "active site"
                        "catalytic"
                        "interfacial activation"
                        "calcium-binding site"
  SECOND.STRUCT:        alpha-helix
  REGION:               'lid'
  ARTICLE:              <ARTICLE-str97-521>
<PROTEIN-str97-521>:=
  NAME:                 "Triacylglycerol lipase"
  PDB_CODE:             1LGY
  SPECIES:              <SPECIES-str97-521>
<SPECIES-str97-521>:=
  NAME:                 "Pseudomonas cepacia"
```

The residue information contains slots that describe the structural characteristics of the particular protein (e.g. SECONDARY structure, REGION) and the importance of the residue in the structure (e.g. SITE/FUNCTION). Other slots serve as pointers, linking different template objects together to represent relational information between entities (e.g. the PROTEIN and SPECIES slots). Further Template Relations can also be defined to link proteins or residues with structural equivalence.

THE EMPATHIE AND PASTA SYSTEMS

The IE systems developed to carry out the EMPathIE and PASTA tasks are both derived from the Large Scale Information Extraction (LaSIE) system, a general purpose IE system. under development at Sheffield since 1994. [10, 11] One of several dozen systems designed to take part in the MUC evaluations over the years, the LaSIE system more or less fits the description of a generic IE system. [12]

LaSIE is neither as 'deep' as some earlier IE systems that attempted full syntactic, semantic and discourse processing [13] nor as 'shallow' as some recent systems that use finite state pattern matching techniques to map directly from source texts to target templates. [14] The processing modules that make up the EMPathIE system are shown in figure 2, within the GATE development environment. [15] The PASTA system is similar and reuses several

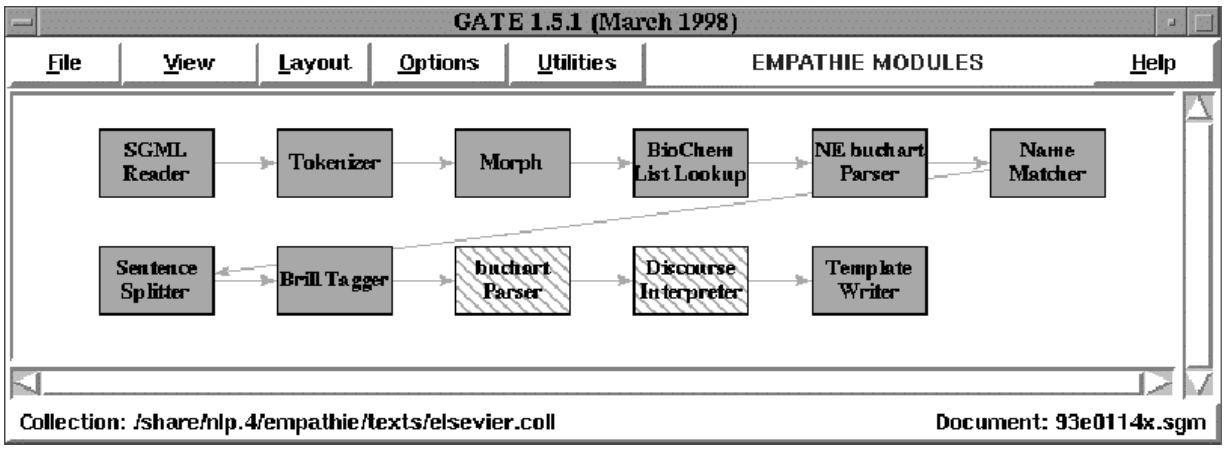


Figure2: EMPathIE system modules within GATE

modules, within the same environment. The architecture of the original LaSIE system has been substantially rearranged for its use in the biochemical domain, mainly to allow the reuse of general English processing modules, such as the part-of-speech tagger and the phrasal parser, without special retraining or adaptation to allow for the domain-specific terminology. This has resulted in an independent terminology identification subsystem, postponing general syntactic analysis until an attempt to identify terms has been made. In general, the original LaSIE system modules, developed for news-wire applications, have been reused, but with various modifications resulting from specific features of the texts, as described in the following. Both systems have a pipeline architecture consisting of four principal stages, described in the following sections: *text preprocessing* (SGML/structure analysis, tokenization), *lexical and terminological processing* (terminology lexicons; morphological analysis, terminology grammars), *parsing and semantic interpretation* (sentence boundary detection; part-of-speech tagging, phrasal grammars, semantic interpretation), and *discourse interpretation* (coreference resolution, domain modeling).

Text Preprocessing

Scientific articles typically have a rigid structure, including abstract, introduction, method and materials, results, and discussion sections; and for particular applications certain sections can be targeted for detailed analysis while others can be skipped completely. Where articles are available in SGML with a DTD, an initial module is used to identify particular markup, specified in a configuration file, for use by subsequent modules. Where articles arc in plain text, an initial module called 'sectionizer' is used to identify and classify significant sections using sets of regular expressions. Both the SGML and sectionizer modules may specify that certain text regions are to be excluded from any subsequent processing; avoiding detailed processing of apparently irrelevant text, especially within the discourse interpretation stage where coreference resolution is

a relatively expensive operation.

The tokenization of the input needs to identify tokens within compound names, such as abbreviations like *NaCl,* where *Na* and *Cl* need to be matched separately in the lexical lookup stage to avoid listing all possible sequences explicitly. The tokenization module must therefore make as few assumptions as possible about the input, proposing minimal tokens that may be recombined in subsequent stages.

Lexical and Terminological Processing

The main information sources used for terminology identification in the biochemical domain are: case-insensitive terminology lexicons, listing component terms of various categories; morphological cues, mainly standard biochemical suffixes; and hand-constructed grammar rules for each terminology class. For example, the enzyme name *mannitol-l-phosphate 5-dehydrogenase* would be recognised firstly by the classification of *mannitol* as a potential compound modifier, and *phosphate* as a compound, both by being matched in the terminology lexicon. Morphological analysis would then suggest *dehydrogenase* as a potential enzyme head, due to its suffix *-ase,* and then grammar rules would apply to combine the enzyme head with a known compound and modifier that can play the role of enzyme modifier.

The biochemical terminology lexicons, acquired from various publicly available resources, have been structured to distinguish various term components, rather than complete terms; which are then assembled by grammar rules. Resources such as the lexicon of enzyme names were manually split into separate lists of component terms, based purely on their apparent syntactic structure rather than any expert knowledge of whatever semantic structure the names reflect. Corresponding grammar rules were then added to recombine the components. Of course, lists of complete multi-word terms can also be used directly in the lexicons, but the rule-based approach has the advantage of being able to recognise novel combinations, not explicitly present in the term lists, and avoids reliance on the accuracy and completeness of available terminology resources. Component terms may also play multiple roles in different terminology classes, for instance amino acid names may be components of both protein and enzyme names, as well as terms in their own right, but the rule-based approach to terminology recognition means they only need to be listed in a single terminology category. The total number of terminology lexicon entries for the biochemical terms is thus comparable to other domains, with approximately 25,000 component terms in about 50 categories for each system at present.

Parsing and Semantic Interpretation

The syntactic processing modules treat any terms recognized in the previous stage as non-decomposable units, with a syntactic role of proper noun. The sentence splitting module cannot therefore propose sentence boundaries within a preclassified term. Similarly, the part-of-speech tagger only attempts to assign tags to tokens which are not part of proposed terms, and the phrasal parser treats terms as preparsed noun phrases. Of course, this approach does not necessarily assume the terminology recognition subsystem to be fully complete and correct, and subsequent syntactic or semantic context can still be used to reclassify or remove proposed terms. In particular, tokens which are constituents of terms proposed but not classified by the NE subsystem, i.e. potential but unknown NEs, are passed to the tagger and phrasal parser as normal, but the potential term is passed to the parser in addition, as a proper name, to allow the

phrasal grammar to determine the best analysis. If the unclassified NE is retained after phrasal parsing, it may be classified within the discourse interpreter using its semantic context or as a result of being coreferred with an entity of a known class.

The phrasal grammar includes compositional semantic rules, which are used to construct a semantic representation of the 'best', possibly partial, parse of each sentence. This predicate logic-like representation is passed on as input to the discourse interpretation stage.

Discourse Interpretation

The discourse interpreter adds the semantic representation of each sentence to a predefined *domain model,* made up of an ontology, or concept hierarchy, plus inheritable properties and inference rules associated with concepts. The domain model is gradually populated with instances of concepts from the text to become a *discourse model.* A powerful coreference mechanism attempts to merge each newly introduced instance with an existing one, subject to various syntactic and semantic constraints. Inference rules of particular instance types may then fire to hypothesize the existence of instances required to fill a template (e.g. an organism with a *source* relation to an enzyme), and the coreference mechanism will then attempt to resolve the hypothesized instances with actual instances from the text.

The template writer module reads off the required information from the final discourse model and formats it as in the template specification.

Initial domain models for the EMPathIE and PASTA tasks have been manually constructed directly from the template definition. This involves the addition of concept nodes to the system's semantic network for each of the entities required in the template, with subhierarchies for possible subtypes, as required. Property types are added for each of the template slots (e.g. *concentration, temperature),* and consequence rules added to hypothesize instances for each slot of a template entity; from an appropriate textual trigger. The Discourse Interpreter's general coreference mechanism is then used to attempt to resolve hypothesized instances with instances mentioned in the text. Subsequent refinement of these models will involve extending the concept subhierarchies and the addition of coreference constraints on the hypothesized instances, based on available training data

RESULTS AND EVALUATION

Evaluation

So far, a complete EMPathIE system exists which has been developed by concentrating on the full texts of six journal papers (the *development* corpus) and evaluated against a corpus of a further seven journal papers (the *evaluation* corpus). Filled templates for all thirteen of these journal papers were produced by trained biochemists highlighting key entities on paper copies of the texts and adding marginal notes where necessary to specify compound roles in interactions and any additional slot values such as concentration, temperature, etc. The annotations were translated to template format by the system developer (with the system frozen before evaluation texts were seen), but some degree of subjective interpretation was required in this process. The annotation would therefore probably be difficult to reproduce without a detailed task specification document, which would be aided by inter-annotator agreement studies to highlight areas of ambiguity in the task definition. However, the current templates at least have the advantage of being produced with some degree of consistency by the developer alone, and so do allow a useful measure of the system's accuracy.

Overall template-filling results are shown in Table 1. The columns show: the number of items the system correctly identified (CORrect), the number of items where the system response and the answer key differed (INCorrect), the number of items the system missed (MISsing), the number the system spuriously proposed (SPUrious) and the standard metrics of RECall and PREcision; discussed in section 2 above. Here "items" refers to filled slot occurrences in the templates. Scoring proceeds by first aligning template objects in the system response with objects in the answer key and then counting the number of matching slot fills in the aligned objects (see [4] for details).

Test Set	COR	INC	MIS	SPU	REC	PRE
Dev	150	121	330	61	25	45
Eval	213	193	518	93	23	43

Table 1: Initial Template results for EMPathIE

In addition to evaluating the template filling capabilities of the prototype we have evaluated its performance at correctly identifying and classifying term classes in the texts (this corresponds to the MUC named entity task). To do this six of the seven evaluation corpus articles were manually annotated for eleven terminology or named entity classes. The results are shown in Table 2.¶

Name_Type	COR	INC	MIS	SPU	REC	PRE
compound	538	27	553	39	48	89
element	24	0	19	14	56	63
enzyme	612	0	12	23	98	96
genus	15	0	18	11	45	58
location	33	1	15	24	67	57
measure	566	0	120	81	83	87
organism	188	9	53	64	75	72
organization	35	6	31	8	49	71
pathway	0	0	15	4	0	0
person	17	1	58	9	22	63
TOTALS	2028	44	894	277	68	86

Table 2: Initial Named Entity results for EMPathIE

The development of the PASTA system has reached the stage where a prototype system exists which can produce templates as described above. A corpus of 52 abstracts of journal articles has been manually annotated with terminology classes, by the system developer with the assistance of a molecular biologist, to allow an automatic evaluation of the PASTA terminology system using the MUC scoring software. Table 3 shows some preliminary results for the main terminology classes.

Name_Type	COR	INC	MIS	SPU	REC	PRE
protein	358	0	52	12	87	97
species	111	0	22	3	83	97
residue	175	0	4	13	98	93
site	53	0	34	10	61	84
region	19	0	24	0	44	100
2_struct	78	0	1	1	99	99
sup_struct	84	0	0	5	100	94
4_struct	115	0	5	3	96	97
chain	27	0	12	0	69	100
base	38	0	0	1	100	97
atom	42	0	2	10	95	81
nonprotein	107	0	0	21	100	84
interaction	10	0	3	1	77	91
TOTALS	1217	0	159	80	88	94

Table 3: Initial Named Entity results for PASTA

Discussion

It should be stressed that these evaluation results are very preliminary, and we would expect them to improve substantially with further development.

The overall EMPathIE template filling precision scores for both the development and evaluation sets are very close to the score of the LaSIE system in the MUC-7 evaluation (42%). Recall is noticeably lower however (47% in MUC-7), but this is certainly affected by the limited amount of training data available, giving a much smaller set of key words and phrases to use as cues for template fills.

It is clear that the EMPathIE task requires much more specialist domain-specific knowledge than the MUC tasks, which typically require only general knowledge of companies and business procedures. The EMPathIE task, as the process of manually filling the templates has demonstrated, can only be performed with the use of detailed domain knowledge, very little of which has been incorporated into the system. For example, a single mention of *cyanide* in one of the evaluation texts causes its entry as an *inhibitor* in the manually filled template, though no explicit information in the text would allow it to be classified as such. Only domain-specific knowledge that *cyanide* is usually an inhibitor allows it to be classified in this case. Such cases are missed completely by the system because the specific knowledge required has not been entered, mainly due to the fact, that the developer is not an expert in the domain.

Further consultation with experts would allow more domain-specific information to be entered, improving recall in particular. With this, and a more extensive training set, it should be entirely possible for system performance on the EMPathIE task to equal the best MUC-7 scores (48% recall, 68% precision, from different systems).

The terminology recognition results are more encouraging, and compare favorably with MUC named entity results, particularly the PASTA results. It should be noted that both the EMPathIE and PASTA terminology recognition tasks require the recognition of a considerably broader class of terms than the MUC named entity task and that considerably smaller sets of training data were available. The discrepancy between the EMPathIE and PASTA results on this task can probably be explained by the fact that there was in fact no training data available specifically for the EMPathIE task before the evaluation was carried out, only the informal feedback of biologists looking at system output. Furthermore, the annotation of texts for the EMPathIE terminology task was carried out by a larger group of people than carried out the PASTA annotation task and without a formal annotation specification. Thus, this annotated data is almost certainly less consistently annotated and the results should therefore be interpreted with some caution.

CONCLUSION

Between these two projects much of the low-level work of moving IE systems into the new domain of molecular biology and the new text genre of journal papers has been carried out. We have generalized our software to cope with longer, multi-sectioned articles with embedded SGML; we have generalized tokenization routines to cope with scientific nomenclature and terminology recognition procedures to deal with a broad range of molecular biological terminology. All of this work is reusable by any IE application in the area of molecular biology.

In addition we have made good progress in designing template elements, template relations, and scenario templates whose utility is attested by working molecular biologists and in adapting our IE software to fill these templates. Preliminary evaluations demonstrate the difficulty of the task, but results are encouraging, and the steps to take to improve performance straightforward. Thus, we are optimistic that IE techniques will deliver novel and effective ways for scientists to make use of the core literature that defines their disciplines.

ACKNOWLDEGMENTS

EMPathIE is a 1.5 year research project in collaboration with, and funded by, GlaxoWellcome plc and Elsevier Science. The authors would like to thank Dr. Charlie Hodgman of GlaxoWellcome for

supplying domain expertise and Elsevier for supplying electronic copy of relevant journals. PASTA is funded under the UK BBSRCEPRSC Biolnformatics Programme (BIF08754) and is a collaboration between the Departments of Computer Science, Information Studies and Molecular Biology and Biotechnology at the University of Sheffield. The authors would like to thank Dr. Peter Artymiuk and Prof. Peter Willett of the University of Sheffield for supplying their expertise in the biochemistry domain.

REFERENCES AND NOTES

[1] Cowie, J.; Lehnert, W. *Communications of the ACM,* **1996**, 39, 80.

[2] Gaizauskas, R.; Wilks, Y. *Journal of Documentation,* **1998**, 54, 70.

[3] DARPA: Defense Advanced Research Projects Agency. *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Morgan Kaufmann, **1995**.

[4] DARPA: Defense Advanced Research Projects Agency. *Proceedings of the Seventh Message Understanding Conference (MUC-7),* **1998**. Available at http://www.saic.com/.

[5] Fukuda, K.; Tsunoda, T.; Tamura, A.;. Takagi, T. Information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing '98 (PSB'98),* pages 707-718. Hawaii, **1998**.

[6] Rindflesh, T.; Tanabe, L.; Weinstein, J.; Hunter, L. Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of the Pacific Symposium on Biocomputing '2000 (PSB'2OOO),* pages 517-528. Hawaii, **2000**.

[7] Thomas, J.; Milward, D.; Ouzounis, C.; Pulman, S.; Carroll, M. Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Pacific Symposium on, Biocomputing '2000 (PSB'2000),* pages 541-551, Hawaii, **2000.**

[8] Bernstein, F.; Koetzle, J.; Williams, G. Meyer, E. J.; Brice, M.; Rodgers, J.; Kennard, O.; Shi-manouchi, M.; Tasumi, M. *Journal of Molecular Biology,* **1977**, 112, 535.

[9] E. Selkov, E.; Basmanova, S.; Gaasterland, T.; Goryanin, I.; Gretchkni, Y.; Meltsev, N.; Nenashev, V.; Overbeek, R.; Panyushkina, E.; Pronevitch, L.; Selkov, E.; Yunis, I. *Nucleic Acids Res.,* **1996**, 24, 26.

[10] R. Gaizauskas, R; Wakao, T.; Humphreys, K.; Cunningham, H.; Wilks, Y. Description of the LaSIE system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)* DARPA **1995**, 207.

[11] Humphreys, K.; Gaizauskas, R.; Azzam, S.; Huyck, G.; Mitchell, B.; Cunningham, H.; Wilks., Y. Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)* DARPA (**1998**). Available at http://www.saic.corn/.

[12] Hobbs, J. R. The generic information extraction system. In *Proceedings of the Fifth Message Understanding Conference (MUC-5),* Morgan Kaufman, **1993**, 87.

[13] Hobbs, J. R. Description of the TACITUS system as used for MUC-3. In *Proceedings of the Third Message Understanding Conference (MUC-3),* Morgan Kaufmann, **1991**, 200.

[14] Appelt, D., Hobbs, J.; Bear, J.; Israel, D.; Kameyama, M.; Kehler, A.; Martin, D.; Myers, K.; Tyson, M. SRI International FASTUS system: MUC-6 Test Results and Analysis. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)* DARPA (**1995**), pages 237-248.

[15] H. Cunningham, H.; Humphreys, K.; Gaizauskas, R.; Wilks, Y. Software Infrastructure for Natural Language Processing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97),* **1997**, 237. Available as http://xxx.lani.gov/ps/9702005.

¶ In calculating both EMPathie *and* PASTA terminology results we have used a weak criterion of correctness whereby a response is correct if its type matches the type of the answer key and its text extent matches a substring of the key's extent. Insisting on the stronger matching criterion of strict string identity lowers recall and precision scores by approximately 4 % overall

DESIGNING COMBINATORIAL LIBRARIES BY EXPLORING DRUG SPACE

VALERIE J. GILLET

University of Sheffield, Western Bank, Sheffield S10 1EP, UK.
E-mail: V.Gillet@sheffield.ac.uk

Received: 13th July 2000 / Published 11th May 2001

ABSTRACT

The techniques of combinatorial chemistry and high throughput screening are in widespread use in the pharmaceutical and agrochemical industries. During the last few years, many different computational approaches have been developed to select compounds for screening and to design combinatorial libraries. The main approaches are reviewed in the first half of this paper. In the second half, we describe how the library design program SELECT has be used to demonstrate that significant improvements in diversity can be achieved by basing library design in product space rather than in reactant space. A series of experiments are reported involving two combinatorial libraries, three different descriptors and three different diversity indices. Finally, a further significant advantage of performing library design in product space is the ability to optimise multiple properties simultaneously. Thus, SELECT can be used to design libraries that are both diverse and have drug-like physiochemical properties.

INTRODUCTION

Combinatorial chemistry is the process whereby large numbers of compounds are synthesized simultaneously in what are known as combinatorial libraries. The technique, together with the related technology of high-throughput screening, is now used routinely in programs for the discovery of novel bioactive compounds in the pharmaceutical and agrochemical industries. In contrast, traditional approaches to medicinal chemistry involved synthesizing one compound at a time, testing or screening that compound for activity, and then iteratively designing and testing new compounds based on the results. Using traditional methods, a medicinal chemist can synthesize approximately 50 compounds per year. The new technologies, which were introduced in the late eighties and early nineties, have vastly increased throughput so that tens of thousands of compounds can now be made in a single cycle.

Initially the belief was that simply making and testing large numbers of compounds would lead to increased chances of finding actives. However, it soon became apparent that it would not be possible to make all potential compounds due to the combinatorial effect, and nor, in fact, was this desirable. In a typical combinatorial reaction the number of suitable reactants that are available just from commercial sources would result in billions of potential products, which far exceeds the capacity of current combinatorial technologies. For example, Walters et al. [1] report that if all suitable reactants for constructing a benzodiazepine library are extracted from the Available Chemicals Directory the resulting library would consist of in the order of 10^9 compounds. Thus, there is a need to select the compounds that are actually made and tested. The libraries that could potentially be made using all available reactants are referred to as virtual libraries and virtual screening is the process of reducing a

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/gillet/gillet.pdf

virtual library to a practical size for combinatorial synthesis and high-throughput screening. Virtual screening techniques can also be used to select compounds for screening from in-house databases and to determine which compounds should be purchased from external suppliers in compound acquisition programs.

Diversity Analysis and Compound Selection Strategies

Virtual screening, or compound selection, techniques build on the similar property principle, which states that structurally similar molecules are likely to have similar properties. [2] The converse of this suggests that dissimilar molecules will tend to have different properties. This is illustrated in Figure 1, which represents a structure space based on two orthogonal properties. Assuming that the properties are relevant to biological activity, then molecules that are close in the space will tend to have similar biological activity. [3] In terms of a screening experiment, the molecules convey redundant SAR information: they have similar structure and similar activity. In library design and HTS, usually the aim is to screen compounds against a number of different biological targets and a subset of compounds that is evenly spread throughout structure space is likely to maximize coverage of biological activity space. Therefore, there has been a great deal of interest in selecting diverse subsets of compounds that cover as much of the structure space as possible without including redundancy.

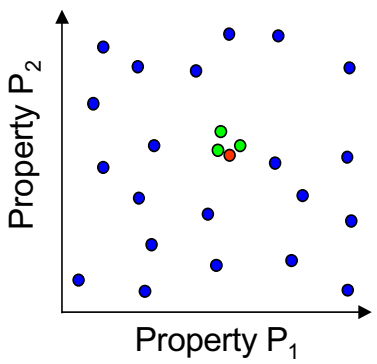


Figure 1: Given an active molecule (red) in a property space that is relevant to biological activity, then according to the similar property principle molecules that are close to it (green) are also likely to

Many diversity measures and subset selection procedures are based on calculations of similarities or dissimilarities between molecules. [4] Consequently, there has been a great deal of interest in measuring structural similarity and dissimilarity [5] and in applying the measures to analyze the diversity of sets compounds [6] and to design diverse combinatorial libraries. Measuring the similarity between two compounds requires that the molecules are described using some numerical descriptors and a coefficient that is used to quantify the degree of similarity between the two sets of descriptors associated with the molecules. [5] The design of diverse libraries requires three major components: the descriptors used to characterize the molecules; a subset selection procedure; and a diversity index that quantifies the degree of diversity in the resulting library.

Descriptors

Many different descriptors have been developed for both similarity and diversity analyses. They have been reviewed extensively (see, for example, [7]) and will be described briefly here. For use in library design, descriptors should have the following characteristics: they should be relevant, that is they should capture structural properties that influence the biological activity of interest; they should be rapid to calculate to enable them to be applied to large datasets; and ideally they should also be chemically interpretable. Descriptors can be categorized as one-, two- or three-dimensional. One-dimensional descriptors are single valued integers or real numbers and they include physicochemical properties, such as molecular

weight, molar refractivity etc., and topological indices, which are indices calculated from the 2D representation of a molecule as a graph. There can be a large number of such descriptors, for example, the Molconn-Z program [8] generates several hundreds of indices, and typically in diversity analyses the number of variables is reduced to a small number that contains most of the information using a technique such as principal components analysis. Two-dimensional fingerprints are probably the most commonly used descriptors in diversity analyses. Three- and four-point pharmacophores are represented as binary vectors or bitstrings, where each bit is set to 0 or 1, depending on whether a particular substructural feature is present or absent from a molecule. Examples include, UNITY [9] and Daylight [10] fingerprints and MACCS screens. [11] Three-dimensional descriptors are also used in diversity analyses, for example, three- and four-point pharmacophores. [12] A pharmacophore point is a substructural feature that is thought likely to influence binding to a receptor, for example, a hydrogen bond donor. They are represented as binary vectors with each bit corresponding to a particular arrangement of pharmacophore points in 3D space. Although they are appealing as descriptors of bioactivity since receptor binding is a 3D event, the calculation of the descriptors is a non-trivial task due to the fact that molecules in general are flexible.

Given a set of molecular descriptors, the similarity (and hence dissimilarity) between two molecules can be calculated using an association coefficient such as the Tanimoto coefficient, which is typically used with binary data such as fingerprints, or a distance measure such as Euclidean distance, which is typically used with physicochemical property data. Association coefficients and distance measures are reviewed by Willett *et al*. [5]

Currently, there is no clear picture as to which descriptors are best. A number of evaluation studies have been performed that tend to suggest that 2D fingerprints are most effective (see, for example, [13] and [14]), however, there is a need for further studies of this kind.

SUBSET SELECTION

Given a set of molecular descriptors and assuming we have some way of measuring diversity *via* a diversity index, then, in theory, the most diverse subset of a given size can be found by generating all possible subsets and calculating the diversity of each one. However, exploring all possible subsets of size n within a dataset of size N requires evaluation of:

$$\frac{N!}{n!(N-n)!}$$

subsets. This is not computationally feasible for typical values of n and N encountered in library design. Therefore, computationally efficient methods are required to find subsets that are approximate solutions. The subset selection methods that have been applied to selecting diverse subsets can be divided into four different categories: [15] dissimilarity-based compound selection methods; clustering; partitioning and optimization techniques. Each of these methods will be described briefly.

In Dissimilarity-Based Compound Selection (DBCS), [16] heuristics are used to provide good, although non-optimal, solutions to the subset problem. The basic algorithm involves initially selecting the first compound and placing it in the subset and then an iterative loop is entered where in each iteration the compound remaining in the dataset that is most dissimilar to those already in the subset is selected and added to the subset. The algorithm terminates when the required number of compounds has been selected. The various

algorithms developed for DBCS differ in the way the first compound is selected and in the way the dissimilarity between one compound and a group of compounds is measured. For example, the first compound may be selected at random, as the one that is in the center of the dataset, or as the one that is most dissimilar to all the others. The most common ways in which the dissimilarity between one compound and a set of compounds is measured are known as MaxMin and MaxSum. MaxMin selects the compound that has maximum distance to its closest neighbor and MaxSum selects the compound whose average distance to all the compounds in the subset is a maximum. The algorithms are typically used with 2D fingerprints or topological indices as descriptors and the distance coefficients for measuring pairwise dissimilarities are usually the complements of the Tanimoto coefficient and the cosine coefficient.

Clustering [17] is the process of dividing objects into groups, or clusters, so that the objects within a cluster are similar and objects from different clusters are dissimilar. A representative subset can then be chosen by selecting one or more compounds from each cluster. Thus, clustering is an indirect way of selecting a subset since the molecules must first be clustered. The clustering process itself is computationally expensive whereas the subsequent selection process is trivial. As with DBCS, clustering also requires the ability to measure similarities or distances between objects and in subset selection it is most often used with 2D fingerprints. There are many different approaches to clustering but the method that has been found to be most effective is Ward's clustering [13] which is a hierarchical agglomerative method. The computational expense of clustering means that the size of datasets that can be handled is limited.

Partitioning, or cell-based, approaches [12] to subset selection involve firstly defining a low dimensional chemistry space, for example, one that is based on molecular properties such as molecular weight, lipophilicity *etc*. The range of values associated with each property is then divided into a series of bins and the combinatorial product of all bins defines a set of cells that cover the entire space. The molecules are then positioned within the space according to their particular properties. A subset can be chosen by selecting one molecule from each cell. Partitioning methods are sometimes referred to as absolute diversity measures, rather than relative measures since the space is defined independently of the molecules that are positioned within it, unlike clustering, where the clusters are determined by the intermolecular distances themselves. This characteristic of partitioning methods means that it is easy to perform database comparisons, which can be a very useful procedure in library design. Partitioning schemes have been developed for low-dimensional data such as physicochemical properties and also for a new type of descriptor called BCUT descriptors. [12] Partitioning schemes are also used with the vector based, one-dimensional, three- and four-point pharmacophores described earlier. One difference with pharmacophore data compared to physicochemical property data is that a single molecule will typically occupy more than one cell and in some cases an individual molecule can occupy a large number of cells, such molecules are sometimes called promiscuous.

The final category of subset selection algorithms is that of optimization techniques. Several methods have been developed that fall into this category. The methods require definition of a function or diversity index that is to be optimized. Examples of algorithms that have been applied to subset selection include genetic algorithms [18,19,20], simulated annealing, [21] and experimental design techniques. [22] In these algorithms, the function is

calculated many times and hence the complexity of the calculation is restricted. Diversity indices that are used with optimization techniques include distance based indices such as the sum of pairwise dissimilarities [18] and the number of distinct pharmacophores [19] covered by a subset of compounds.

most combinatorial library design efforts to date have been based on reactant-based selection. The methods assume that by maximizing diversity in the reactants, maximum diversity in the product molecules will be achieved. However, recent evidence suggests that significantly more diverse libraries can be achieved if selection is performed in product space [18,23,24,25].

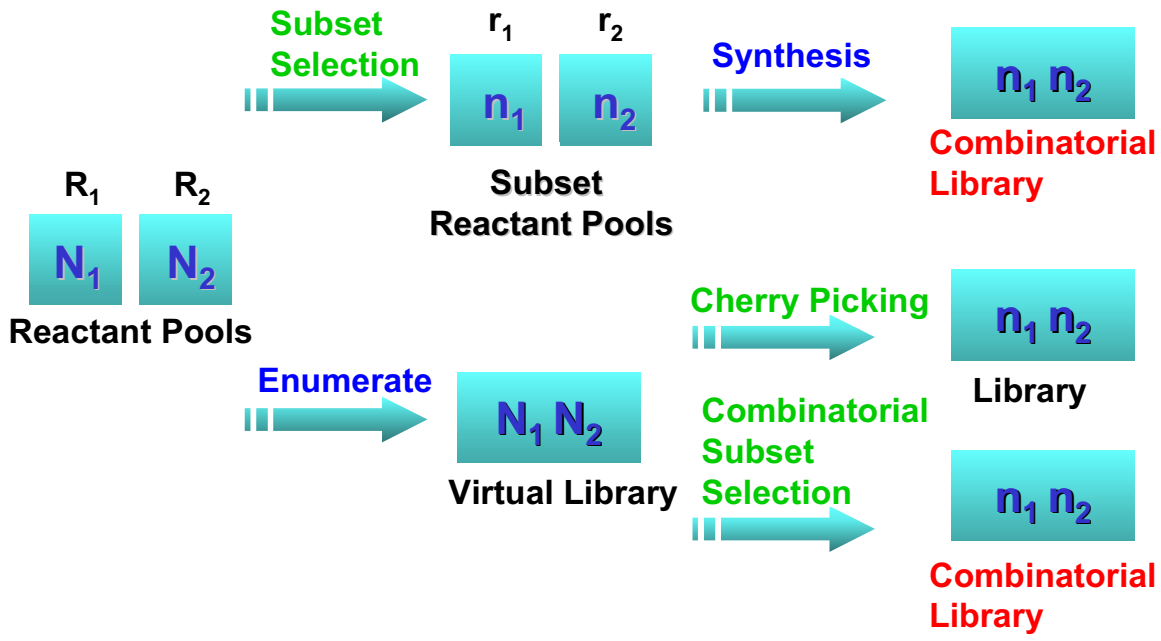


Figure 2: Three different strategies are available for designing diverse combinatorial libraries. They are reactant-based selection, shown in the top half of the figure; cherry-picking in product space; and combinatorial subset selection in product space, known as product-based selection.

COMBINATORIAL LIBRARY DESIGN

The discussion so far has concentrated on strategies for the selection of subsets of compounds with particular reference to the selection of HTS sets from, for example, in-house databases, and the selection of compounds to purchase from external suppliers. In combinatorial library design, any of the techniques already described can be applied directly to choose subsets of reactants from those that are available for use in a combinatorial synthesis. The subsets of reactants are then used to build a combinatorial library in a process known as *reactant-based selection*. This approach is shown schematically in the top-half of Figure 2. Indeed,

In product-based library design a virtual library is enumerated using all available reactants, as shown in the bottom-half of Figure 2. The simplest way of performing product-based selection is to apply any of the techniques described previously in a process known as *cherry-picking*. This approach, however, is synthetically inefficient as far as combinatorial chemistry is concerned since it does not take into account the combinatorial constraint and hence is highly unlikely to result in a combinatorial library. The synthetic inefficiency of cherry-picking is shown in Figure 3, where a two-component combinatorial reaction is represented by a two-dimensional array. The rows represent the reactants

	Y_1	Y_2	Y_3	Y_4	Y_5
X_1	X_1Y_1	X_1Y_2	X_1Y_3	X_1Y_4	X_1Y_5
X_2	X_2Y_1	X_2Y_2	X_2Y_3	X_2Y_4	X_2Y_5
X_3	X_3Y_1	X_3Y_2	X_3Y_3	X_3Y_4	X_3Y_5
X_4	X_4Y_1	X_4Y_2	X_4Y_3	X_4Y_4	X_4Y_5
X_5	X_5Y_1	X_5Y_2	X_5Y_3	X_5Y_4	X_5Y_5

Figure 3: A two component combinatorial library is represented as a two-dimensional array with the reactants in one pool represented by the rows and the reactants in the second pool represented by the columns. A cherry-picked library consisting of four molecules is shown highlighted in red. The 4 × 3 combinatorial library that contains these four molecules is shown in blue.

in one pool, the columns represent the reactants in the second pool and the elements represent the full virtual library. A cherry-picked subset is equivalent to picking compounds from anywhere in the array, for example the subset of four compounds that are highlighted. Synthesizing these four compounds combinatorially would require synthesis of 12 products in a 4 × 3 library.

A combinatorial subset can be selected by intersecting the rows and columns of the matrix, as shown in Figure 4 where the products of a 2 × 2 combinatorial subset library are highlighted. Generating all possible combinatorial subsets in order to find the most diverse is then equivalent to

	Y_1	Y_2	Y_3	Y_4	Y_5
X_1	X_1Y_1	X_1Y_2	X_1Y_3	X_1Y_4	X_1Y_5
X_2	X_2Y_1	X_2Y_2	X_2Y_3	X_2Y_4	X_2Y_5
X_3	X_3Y_1	X_3Y_2	X_3Y_3	X_3Y_4	X_3Y_5
X_4	X_4Y_1	X_4Y_2	X_4Y_3	X_4Y_4	X_4Y_5
X_5	X_5Y_1	X_5Y_2	X_5Y_3	X_5Y_4	X_5Y_5

Figure 4: A combinatorial subset can be selected by intersecting the rows and columns of the matrix.

permuting the rows and columns of the matrix in all possible ways. However, matrix manipulation of this sort represents an enormous search space and, in practice, investigating all possible combinatorial subsets is infeasible for real library design problems. Once again, an approximate solution can be found by using an optimization technique. We have implemented a genetic algorithm (GA) that is able to select a combinatorial subset from a full virtual library of products, within the program called SELECT. [24] In SELECT, each chromosome of the GA encodes a combinatorial subset in the form of lists of the reactants that make up the library. The fitness function of the GA involves enumerating the sub-library and measuring its diversity. The GA iterates through generations using crossover, mutation and roulette wheel selection until it converges on an optimally diverse library. Diversity can be measured using a number of different molecular descriptors and diversity indices, for example, Daylight fingerprints and the sum-of-pairwise dissimilarities using the cosine coefficient.

SELECT has been used to compare the diversity that can be achieved with reactant-based selection relative to product-based selection. [18,23] The libraries that were examined are a two-component amide library (Figure 5) where the virtual library of 10,000 products is built from 100 amides and 100 carboxylic acids, and a three-component thiazoline-2-imine library (Figure 6), also of 10,000 products, which is built from 10 isothiocyanates, 40 amines and 25 haloketones.

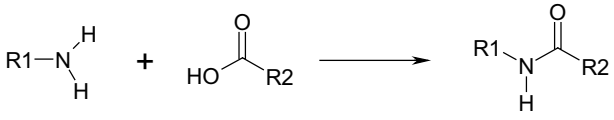


Figure 5: The amide library.

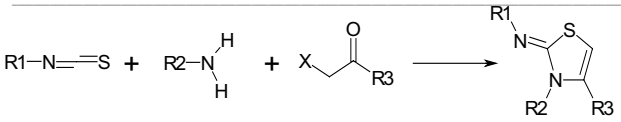


Figure 6: The thiazoline-2-imine library.

Index	Descriptors	Reactants	Products	Min	%Δ
SUM_{COS}	Daylight	0.565 (0.002)	0.586 (0.002)	0.356	9.4
SUM_{TAN}		0.715 (0.002)	0.744 (0.002)	0.522	12.5
NN		0.253 (0.003)	0.305 (0.001)	0.045	20.1
SUM_{COS}	UNITY	0.552 (0.002)	0.566 (0.002)	0.339	5.9
SUM_{TAN}		0.715	0.727	0.507	5.5
NN		0.243	0.294	0.045	20.5
SUM_{COS}	Molconn-Z	0.278 (0.001)	0.288 (0.000)	0.121	6.5
SUM_{TAN}		0.451	0.470	0.217	7.5
NN		0.107	0.150	0.036	37.7

Table 1: Reactant-based versus product-based diversities for 30 × 30 amide libraries selected from a full virtual library of 100 × 100. The column headed *Min* gives the diversity calculated when SELECT was run to find combinatorial subsets with minimum diversity. The final column, %Δ, gives the percentage difference in diversity between product-based and reactant-based selection relative to the range of values possible (calculated by subtracting the *Min* diversity from the Product diversity).

In both cases the reactants were selected at random from the SPRESI database. [26] The experiments were performed for three different types of descriptors, namely 1024 bit Daylight fingerprints, 992 bit UNITY fingerprints, and 538 Molconn-Z parameters that were standardized in the range 0-1. Three different diversity indices were used, namely, the sum-of-pairwise dissimilarities using the cosine coefficient, the sum-of-pairwise dissimilarities using the Tanimoto coefficient, and the average nearest neighbor distance using the Tanimoto coefficient. The results are shown in Tables 1 and 2 for the amide library and the thiazoline-2-imine library, respectively. In all cases it can be seen that product-based designs result in more diverse libraries than do reactant-based designs.

The effect is more pronounced over all the descriptors and metrics for the three-component thiazoline-2-imine library. Unlike reactant-based selection, product-based selection takes into account the relationships between reactants in different pools and hence it is reasonable to expect that the relative effectiveness of product-based selection should increase with the number of reactant pools.

Index	Descriptor	% Δ
SUM_{COS}	Daylight	24.8
SUM_{TAN}		22.3
NN		34.6
SUM_{COS}	UNITY	12.9
SUM_{TAN}		8.0
NN		35.6
SUM_{COS}	Molconn-Z	12.6
SUM_{TAN}		11.4
NN		49.2

Table 2: The percentage difference between reactant-based and product-based diversities is reported for 6 × 10 × 15 thiazoline-2-imine libraries selected from a full virtual library of 10 × 40 × 25.

The effectiveness of product-based selection versus reactant-based selection using SUM_{COS} or SUM_{TAN} as the diversity index is more pronounced for Daylight fingerprints than for UNITY fingerprints or Molconn-Z parameters. Daylight fingerprints are based on calculating the paths of up to 7 atoms within a molecule. When they are calculated for a product molecule there are likely to be several paths that span reactants that originate in different pools, thus there will be parts of the fingerprint that are unique to the product molecule and that are not found in its constituent reactants. This is especially the case for the three-component library. Thus it is not surprising that better results can be achieved by

performing the analysis in product-space. UNITY fingerprints, however, also include some structural keys that record the presence or absence of particular fragments. The structural keys tend to be more localized than the path-based fragments and hence there will be fewer bits that arise in the product molecules only. It is more difficult to explain the performance seen with the Molconn-Z parameters since they encompass a huge range of types of molecular descriptor.

The difference between product-based and reactant-based selection is most marked for the NN diversity index. Combinatorial libraries tend to contain clusters of closely related compounds since each reactant in a reactant pool exists in a product molecule with every reactant in the other pools. The presence of closely related compounds can still result in relatively high diversity values using both the SUM_{COS} and SUM_{TAN} indices. [27] However, the presence of clusters of compounds will result in low diversity values according to the NN index, which prefers an even distribution of compounds. Thus, maximizing the NN index in product space is likely to produce a better spread of compounds throughout the space than can be achieved by just considering the reactants alone.

DRUG-SPACE

Early libraries designed on the basis of diversity did show increased rates of finding hits. However, inspection of the hits revealed that they tended to have undesirable properties as far as potential drug candidates are concerned. For example, they tended to have high molecular weights, to be too lipophilic, or to be insoluble. [28] In fact, it appears that maximizing diversity tends to bias the molecules in libraries away from desired ranges of these properties. Thus, the emphasis in library design has now shifted towards designing libraries that, while still diverse, contain compounds constrained to have drug-like physicochemical properties.

One way in which this is attempted is by applying preliminary filters to eliminate non-drug-like molecules from the reactant pools, for example, removing toxic and reactive groups; compounds

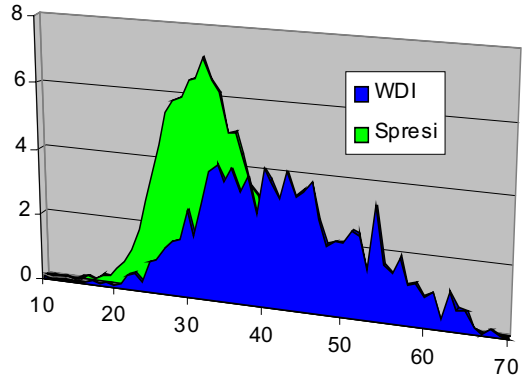


Figure 7: Drug-like weights are used to discriminate between compounds in WDI and compounds in the SPRESI database.

with a large number of rotatable bonds; and compounds with high molecular weights. Some subset selection techniques now use additional properties within the design, for example, tailored D-optimal design [28] and the use of secondary properties to select compounds from cells in a

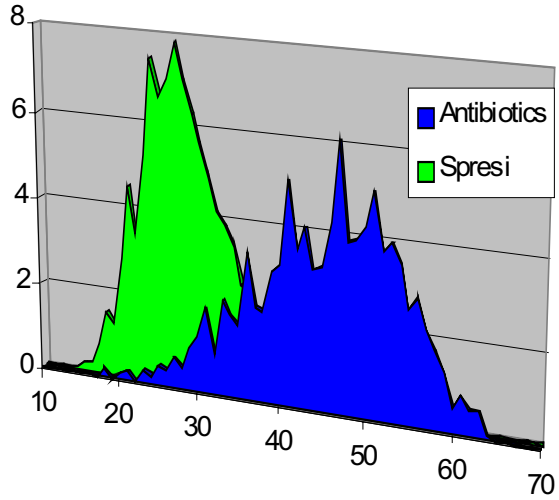


Figure 8: Weights have been derived to discriminate between compounds having antibiotic activity from non-drug-like molecules as found in SPRESI

partitioning procedure.

Recently, several more sophisticated approaches have been described that attempt to predict drug-likeness. [29-32] These methods have their basis in the fact that physicochemical properties are distributed differently in databases of drug-like molecules relative to non-drug-like molecules. [29] An example of this type of approach is the bioactivity profiles approach we have developed, [29] where a series of weights is derived that represent different values of physicochemical properties. The weights can be used to score and rank molecules according to their ability to discriminate between active and inactive compounds. Optimum weights are found using a GA. Figure 7 shows the results of applying drug-like weights to discriminate between molecules in the World Drugs Index, [33] which represents drug-like molecules, from molecules in SPRESI which represents non-drug-like molecules. The method

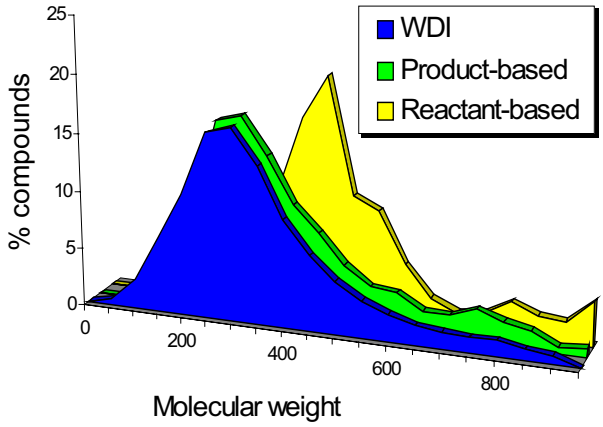


Figure 9: The molecular weight profiles of amide libraries designed using reactant-based selection (in yellow) are compared with libraries that are optimized in product-space (green) and the profile of molecular weights found in WDI (blue).

can be tailored for different classes of activity, for example, in Figure 8, weights have been derived to discriminate antibiotics from non-drugs. This method, and similar methods, can be used to rank datasets for screening so that the compounds that are predicted to be drug-like are screened first, and they can also be used to choose compounds from external suppliers in compound acquisition programs. In the library design context they could be used to choose drug-like reactants, however, they are less suited to product-based design since they do not take account of the combinatorial constraint.

Designing Drug-Like Combinatorial Libraries

We have extended the SELECT program to perform multi-objective optimization in product-space in order that libraries can be designed on multiple properties simultaneously. The fitness function of the GA now consists of a weighted sum as shown:

$$f(n) = w_1.diversity + w_2.complementarity + w_3.cost + w_4.property1 + w_5.property2 \ldots$$

The objectives on library design would typically include diversity along with a number of other properties. The complementarity term can be used to design a library that is complementary to an existing library by maximizing the diversity that would result if the two libraries were merged. The third term represents the cost of synthesizing a library which can be estimated from the cost of the individual reactants that constitute the library. The remaining terms can be used to tailor the physicochemical property profiles of a library. The properties of a library are optimized by comparing the distribution of the property within a library with the distribution of the same property in some reference collection, for example, this could be a collection of drug-like molecules such as those found in the WDI. The weights are user-definable and are usually set to maximize diversity and complementarity while minimizing normalized values of cost and the RMSD between the profile of the properties within the library and the reference profiles.

The effect of multi-component optimization can be seen in Figure 9 where the molecular weight profile of an amide library selected by performing reactant-based selection on diversity alone is show in yellow. The profile of a library selected by performing product-based selection based on diversity and molecular weight simultaneously is shown in green. The molecular weight profile is optimized relative to the profile of molecular weight found in WDI, which is shown in blue. It can be seen that reactant-based selection often results in libraries with poor physicochemical properties. The product-based selection, conversely, has enabled the design of libraries with profiles that are much more WDI-like and that are thus more likely to contain bioactive compounds.

CONCLUSIONS

Many different approaches to designing diverse libraries have been developed, involving a variety of different subset selection techniques and molecular descriptors. We have shown that product-based selection results in libraries that are more diverse than if selection is performed at the reactant-level. Experience has shown that libraries designed on diversity alone have a tendency to contain non-drug-like molecules and it is now apparent that other criteria should also be taken into account. Product-based designs such as that developed in the SELECT program allow for multiple properties to be optimized simultaneously.

ACKNOWLEDGEMENTS

Thanks are due to John Bradshaw, Darren Green, Orazio Nicolotti, Peter Willett and David Wilton for their contributions to SELECT and the bioactivity profiling work, which was funded by GlaxoWellcome Research and Development with software support provided by Daylight Chemical Information Systems and Tripos Inc.

REFERENCES AND NOTES

[1] Walters, W.P.; Stahl, M.; Murcko, M.A. Virtual Screening – An Overview. *Drug Discovery Today*, **1998**, *3*, 160.
[2] Johnson, M.A.; Maggiora, G.M., Eds. Concepts and Applications of Molecular Similarity, John Wiley, New York **1990**.
[3] Patterson, D.E.; Cramer, R.D.; Ferguson, A.M.; Clark, R.D.; Weinberger, L.E. Neighbourhood Behaviour: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem*. **1996**, *39*, 3049.
[4] Gillet, V.J. Background Theory of Molecular Diversity. In *Molecular Diversity in Drug Design*; Dean, P.M.; Lewis, R.A., Eds., Kluwer, Dordrecht **1999**.
[5] Willett, P.; Barnard, J.M.; Downs, G.M. Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983.
[6] Willett P. Computational Tools for the Analysis of Molecular Diversity. *Perspect. Drug Disc. Design*, 1997;7/8, 1.
[7] Brown, R.D. Descriptors for Diversity Analysis. *Perspect. Drug Discov. Design.*, **1997**, *7/8*, 31.
[8] The MOLCONN-Z software is available from eduSoft LC at URL http://www.eslc.vabiotech.com/
[9] UNITY Chemcial Information Software. Tripos Inc., 1699 Hanley Rd., St. Louis, MO 63144.
[10] Daylight Chemical Information Systems, Inc., Mission Viejo, CA, USA.
[11] MACCS II. Molecular Design Ltd., San Leandro, CA.
[12] Mason, J.S.; Pickett, S.D. Partition-Based Selection. *Perspect. Drug Disc. Design*, **1997**, *7/8*, 85.
[13] Brown, R.D.; Martin Y.C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand Binding. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 1.
[14] Matter H. Selecting Optimally Diverse Compounds from Structural Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem*. **1997**, *40*, 1219.
[15] Willett P. Subset-Selection Methods for Chemical Databases. In *Molecular Diversity in Drug Design*, Dean, P.M.; Lewis, R.A., Eds., Kluwer, Dordrecht **1999**.
[16] Lajiness, M.S. Dissimilarity-Based Compound Selection Techniques. *Perspect. Drug Disc. Design*, **1997**, *7/8*, 65.
[17] Dunbar, J.B. Cluster-Based Selection. *Perspect. Drug Disc. Design*, **1997**, *7/8*, 51.
[18] Gillet, V. J.; Willett, P.; Bradshaw, J. The

Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci*, **1997**, *37*, 731.

[19] Good, A.C.; Lewis, R.A. New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick, *J. Med. Chem*., **1997**, *40*, 3926.

[20] Brown, R.D.; Martin, Y.C. Designing Combinatorial Library Mixtures Using a Genetic Algorithm. *J. Med. Chem*., **1997**, *40*, 2304.

[21] Agrafiotis, D.K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 841.

[22] Martin, E.J.; Blaney, J.M.; Siani, M.S.; Spellmeyer, D.C.; Wong, A.K.; Moos, W.H. Measuring Diversity - Experimental Design of Combinatorial Libraries for drug Discovery. *J. Med. Chem*., **1995**, *38*, 1431.

[23] Gillet, V.J., Nicolotti, O. New Algorithms for Compound Selection and Library Design, *Perspect. Drug Disc. Design*, in the press.

[24] Gillet, V.J.; Willett, P.; Bradshaw, J.; Green D.V.S. Selecting Combinatorial Libraries to Optimise Diversity and Physical Properties *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 169.

[25] Jamois, E.A.; Hassan, M.; Waldman, M. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets. *J. Chem. Inf. Comput. Sci*., **2000**, *40*, 63.

[26] The SPRESI database is available from Daylight Chemical Information Systems, Inc., Mission Viejo, CA, USA.

[27] Snarey, M;, Terrett, N.K.; Willett, P.; Wilton D.J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graphics Modelling*, 1997, **15**, 372.

[28] Martin, E.J.; Crichlow, R.W. Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J. Comb. Chem*. **1999**, *1*, 32.

[29] Gillet, V.J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis And Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165.

[30] Ajay; Walters, W.P.; Murcko, M. Can We Learn to Distinguish Between “Drug-like” and “Nondrug-like” Molecules? *J. Med. Chem.*, **1998**, *41*, 3314.

[31] Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.*, **1998**, *41*, 3325.

[32] Wagner, M. van Geerstein, V.J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280.

[33] The World Drugs Index is available from Derwent Information at URL http://www.derwent.com/

CLUSTERING IN MASSIVE DATA SETS

FIONN MURTAGH

School of Computer Science, The Queen's University of Belfast, Belfast BT7 1NN, Northern Ireland.
E-mail: f.murtagh@qub.ac.uk

Received: 25th May 2000 / Published 11th May 2001

ABSTRACT

We review the time and storage costs of search and clustering algorithms. We exemplify these, based on case studies in astronomy, information retrieval, visual user interfaces, chemical databases, and other areas. First we describe nearest neighbor searching, an elemental form of clustering, and a basis for clustering algorithms to follow. Next we review a number of families of clustering algorithms. Finally we discuss visual or image representations of data sets, from which a number of interesting algorithmic developments arise.

'Now', said Rabbit, 'this is a search, and I've organised it - '
'Done what to it?' said Pooh.
'Organised it. Which means - well, it's what you do to a Search,
when you don't all look in the same place at once.'
A.A. Milne, The House at Pooh Corner (1928) –M. S. Zakaria

INTRODUCTION

Nearest neighbor searching is considered first for one main reason: its utility for the clustering algorithms reviewed later. They are the building blocks for the most efficient implementations of hierarchical clustering algorithms, and they can be used to speed up other families of clustering algorithms. We will then deal with facets of visual or image representations of data sets.

The best match or nearest neighbor problem is important in many disciplines. In statistics, *k*-nearest neighbors, where *k* can be 1, 2, *etc.*, is a method of non-parametric discriminant analysis. In pattern recognition, this is a widely used method for unsupervised classification (see [1]).

Nearest neighbor algorithms are the building block of clustering algorithms based on nearest neighbor chains; or of effective heuristic solutions for combinatorial optimization algorithms such as the traveling salesman problem, which is a paradigmatic problem in many areas. In the database and more particularly data mining fields, NN searching is called similarity query, or similarity join. [2]

In the next section, we begin with data structures where the objective is to break the O(*n*) barrier for determining the nearest neighbor (NN) of a point. A database record or tuple may be taken as a point in a space of dimensionality *m*, the latter being the associated number of fields or attributes. These approaches have been very successful, but they are restricted to low dimensional NN-searching. For higher dimensional data, a wide range of bounding approaches have been proposed, which remain O(*n*) algorithms but with a low constant of proportionality.

We assume familiarity with basic notions of similarity and distance, the triangular inequality, ultrametric spaces, Jaccard and other coefficients, normalization and standardization. For an implicit treatment of data theory and data coding, see [3]. Useful background reading can be found in [4]. In

particular output representational models include discrete structures, *e.g.* rooted labeled trees or dendrograms, and spatial structures, [5] with many hybrids.

BINNING OR BUCKETING

In this approach to NN-searching, a preprocessing stage precedes the searching stage. All points are mapped onto indexed cellular regions of space, so that NNs are found in the same or in closely adjacent cells. Taking the plane as as example, and considering points (x_i, y_i), the maximum and minimum values on all coordinates are obtained (*e.g.* (x_j^{min}, y_j^{min})). Consider the mapping (Fig. 1)

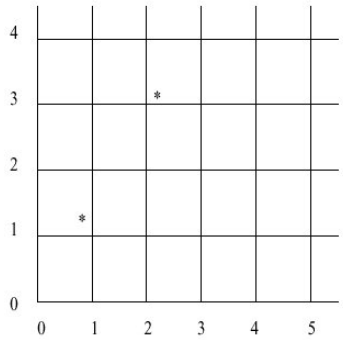
$$x_i \rightarrow \left\{ \left\lfloor \left(x_{ij} - x_j^{min} \right) \Big/ r \right\rfloor \right\}$$

where constant r is chosen in terms of the number of equally spaced categories into which the interval $[x_j^{min}, x_j^{max}]$ is to be divided. This gives to x_i an integer value between 0 and $\lfloor (x_{ij}^{max} - x_{ij}^{min})/r \rfloor$ for each attribute j. $O(nm)$ time is required to obtain the transformation of all n points, and the result may be stored as a linked list with a pointer from each cell identifier to the set of points mapped onto that cell.

NN-searching begins by finding the closest point among those that have been mapped onto the same grid cell as the target point. This gives a current NN point. A closer point may be mapped onto some other grid cell if the distance between target point and current NN point is greater than the distance between the target point and any of the boundaries of the cell containing it. Some further implementation details can be found in [6].

A powerful theoretical result regarding this approach is as follows. For uniformly distributed points, the NN of a point is found in $O(1)$, or constant, expected time (see [7] or [8] for proof). Therefore this approach will work well if approximate uniformity can be assumed or if the

$$x^{min}, y^{min} = 0,0, x^{max}, y^{max} = 50,40, r = 10$$



Point (22,32) is mapped onto cell (2,3); point (8,13) is mapped onto cell (0,1).

Figure 1: Example of simple binning in the plane.

data can be broken down into regions of approximately uniformly distributed points.

Simple Fortran code for this approach is listed, and discussed, in [9]. The search through adjacent cells requires time that increases exponentially with dimensionality (if it is assumed that the number of points assigned to each cell is approximately equal). As a result, this approach is suitable for low dimensions only. Rohlf [10] reports on work in dimensions 2, 3, and 4; and Murtagh [11] in the plane. Rohlf also mentions the use of the first 3 principal components to approximate a set of points in 15-dimensional space.

From the constant expected time NN search result, particular hierarchical agglomerative clustering methods can be shown to be of linear expected time, $O(n)$. [11] The expected time complexity for Ward's minimum variance method is given as $O(n \log n)$. Results on the hierarchical clustering of up to 12,000 points are discussed.

The limitation on these very appealing computational complexity results is that they are only really feasible for data in the plane. Bellman's curse of dimensionality manifests itself here as always. For dimensions greater than 2 or 3 we proceed to the situation where a binary search tree can provide us with a good preprocessing of our data.

Multidimensional Binary Search or kD Tree

A binary search tree preprocesses the data to be searched through by two-way subdivision, and subdivisions continue until some prespecified number of data points is arrived at. See example in Fig. 2. We associate with each node of the decision tree the definition of a subdivision of the data only, and we associate with each terminal node a pointer to the stored coordinates of the points. Using the approximate median of projections keeps the tree balanced, and consequently $O(\log n)$ levels, at each of which $O(n)$ processing is required. Hence the construction of the tree takes $O(n \log n)$ time.

The search for a NN then proceeds by a top-down traversal of the tree. The target point is transmitted through successive levels of the tree using the defined separation of the two child nodes at each node. On arrival at a terminal node, all associated points are examined and a current NN selected. The tree is then backtracked: if the points associated with any node could furnish a closer point, then subnodes must be checked out.

The approximately constant number of points associated with terminal nodes (hyper-rectangular cells in the space of points) should be greater than 1 in order that some NNs may be obtained without requiring a search of adjacent cells (other terminal nodes). Friedman *et al*. [12] suggest a value of the number of points per bin between 4 and 32 based on empirical study.

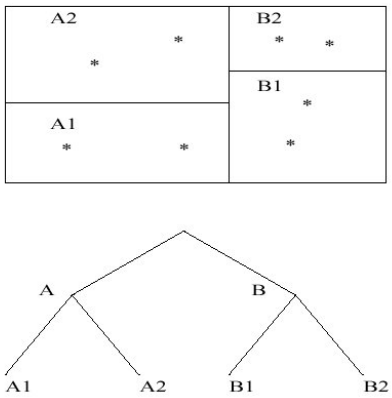


Figure 2: A MDBST using planar data

The MDBST approach only works well with small dimensions. To see this, consider each coordinate being used once and once only for the subdivision of points, *i.e.* each attribute is considered equally useful. Let there be p levels in the tree, *i.e.* 2^p terminal nodes. Each terminal node contains approximately c points by construction and so $c2^p = n$. Therefore $p = \log_2 n/c$. As sample values, if $n = 32768$; $c = 32$; then $p = 10$. That is in 10-dimensional space, using a large number of points associated with terminal nodes, more than 30000 points will need to be considered. For high dimensional spaces, two alternative MDBST specifications are as follows.

All attributes need not be considered for splitting the data if it is known that some are of greater interest than others. Linearity present in the data may manifest itself *via* the variance of projections of points on the coordinates; choosing the coordinate with greatest variance as the discriminator coordinate at each node may therefore allow repeated use of certain attributes. This has the added effect that the hyper-rectangular cells into which the terminal nodes divide the space will be approximately cubic in shape. In this case, Friedman *et al*. [12] show that search time is $O(\log n)$ on average for the finding of a NN. Results

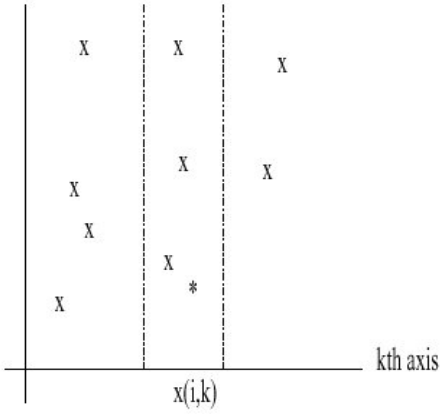


Figure 3: Two-dimensional example of projection-based bound. Points with projections within distance c of given point's (*) projection, alone, are searched. Distance c is defined with reference to a candidate or current nearest neighbor.

obtained for dimensionalities of between 2 and 8 are reported in [12], and in the application of this approach to minimal spanning tree construction in Bentley and Friedman. [13] LISP code for the MDBST is discussed in [14].

The MDBST has also been proposed for very high dimensionality spaces, *i.e.* where the dimensionality may be greater than the number of points, as could be the case in a keyword-based system. Keywords (coordinates) are batched, and the following decision rule is used: if some one of a given batch of node-defining discriminating attributes is present, then take the left subtree, else take the right subtree. Large n, well in excess of 1400, was stated as necessary for good results. [15, 16] General guidelines for the attributes that define the direction of search at each level are that they be related, and the number chosen should keep the tree balanced. On intuitive grounds, our opinion is that this approach will work well if the clusters of attributes, defining the tree nodes, are mutually well separated.

An MDBST approach is used by Moore [17] in the case of Gaussian mixture clustering. Over and above the search for nearest neighbors based on Euclidean distance, Moore allows for the Mahalanobis metric, *i.e.* distance to cluster centers that are "corrected" for the (Gaussian) spread or morphology of clusters. The information stored at each node of the tree includes covariances.

Moore [17] reports results on numbers of objects of around 160,000, dimensionalities of between 2 and 6, and speedups of 8-fold to 1000-fold. Pelleg and Moore [18] discuss results on some 430,000 two-dimensional objects from the Sloan Digital Sky Survey (see the section "k-Means and Family" below).

PROJECTIONS AND OTHER BOUNDS

Bounding using Projection or Properties of Metrics

Making use of bounds is a versatile approach, which may be less restricted by dimensionality. Some lower bound on the dissimilarity is efficiently calculated in order to dispense with the full calculation in many instances.

Using projections on a coordinate axis allows the exclusion of points in the search for the NN of point x_i. Points x_k, only, are considered such that $(x_{ij} - x_{kj})^2 \leq c^2$ where x_{ij} is the j^{th} coordinate of x_i, and where c is some prespecified distance (see Fig. 3).

Alternatively, more than one coordinate may be used. The prior sorting of coordinate values on the chosen axis or axes expedites the finding of points whose full distance calculation is necessitated. The preprocessing required with this approach involves the sorting of up to m sets of coordinates, *i.e.* $O(mn \log n)$ time.

Using one axis, it is evident that many points may be excluded if the dimensionality is very small, but that the approach will worsen as the latter grows. Friedman *et al.* [19] give the expected NN search time, under the assumption that the points are uniformly distributed, as $O(mn^{1-1/n})$. This approaches the brute force $O(nm)$ as n gets large. Reported empirical results are for dimensions 2 to 8.

Marimont and Shapiro [20] extend this approach by the use of projections in subspaces of dimension greater than 1 (usually about m=2 is suggested). This can be further improved if the subspace of the principal components is used. Dimensions up to 40 are examined. The Euclidean distance is very widely used. Two other members of a family of Minkowski metric measures require less computation time to calculate, and they can be used to provide bounds on the Euclidean distance. We have:

$$d_1(x,x') \geq d_2(x,x') \geq d_\infty(x,x')$$

where d_1 is the Hamming distance defined as $\Sigma_j |x_j - x'_j|$, the Euclidean distance is given by the square root of $\Sigma_j (x_j - x'_j)^2$; and the Chebyshev distance is defined as $\max_j |x_j - x'_j|$.

Kittler [21] makes use of the following bounding strategy: reject all points y such that $d_1(x,y) \geq \sqrt{(m)}\delta$ where δ is the current NN d_2-distance. The more efficiently calculated d_1-distance may thus allow the rejection of many points (90% in 10-dimensional space is reported by Kittler). Kittler's rule is obtained by noting that the greatest d_1-distance between x and x′ is attained when

$$|x_j - x'_j|^2 = d_2^2(x,x')/m$$

for all coordinates, j. Hence $d_1(x,x') = d_2(x,x') \geq \sqrt{m}$ is the greatest d_1-distance between x and x'. In the case of the rejection of point y, we then have:

$$d_1(x,y) \leq d_2(x,y)/\sqrt{m}$$

and since, by virtue of the rejection

$$d_1(x,y) \geq \sqrt{m}\delta$$

it follows that $\delta \leq d_2(x,y)$.

Yunck [22] presents a theoretical analysis for the similar use of the Chebyshev metric. Richetin *et al.* [23] propose the use of both bounds. Using uniformly distributed points in dimensions 2 to 5, the latter reference reports the best outcome when the rule: reject all y such that $d_\infty(x,y) \geq \delta$ precedes the rule based on the d_1-distance. Up to 80% reduction in CPU time is reported.

Bounding using the Triangular Inequality

The triangular inequality is satisfied by distances: $d(x,y) \leq d(x,z) + d(z,y)$, where x, y and z are any three points. The use of a reference point, z, allows a full distance calculation between point x, whose NN is sought, and y to be avoided if

$$|d(y,z) - d(x,z)| \geq \delta$$

where δ is the current NN distance. The set of all distances to the reference point are calculated and stored in a preprocessing step requiring $O(n)$ time and $O(n)$ space. The above cut-off rule is obtained by noting that if

$$d(x,y) \geq |d(x,y) - d(y,z)|$$

then, necessarily, $d(x,y) \geq \delta$. The former inequality above reduces to the triangular inequality irrespective of which of $d(y,z)$ or $d(x,z)$ is the greater.

The set of distances to the reference point, $\{d(x,z)|x\}$, may be sorted in the preprocessing stage. Since $d(x,z)$ is fixed during the search for the NN of x, it follows that the cut-off rule will not then need to be applied in all cases.

Shapiro [25] generalized the single reference point approach, due to Burkhard and Keller, [24] to multiple reference points. The sorted list of distances to the first reference point, $\{d(x,z_1)|x\}$, is used as described above as a preliminary bound. Then the subsequent bounds are similarly employed to further reduce the points requiring a full distance calculation. The number and the choice of reference

points to be used is dependent on the distributional characteristics of the data. Shapiro [25] finds that reference points ought to be located away from groups of points. In 10-dimensional simulations, it was found that at best only 20% of full distance calculations were required (although this was very dependent on the choice of reference points).

Hodgson [26] proposes the following bound, related to the training set of points, *y*, among which the NN of point *x* is sought. Determine in advance the NNs and their distances, d(*y*,NN(*y*)) for all points in the training set. For point y, then consider $\delta y = \frac{1}{2}$ d(*y*,NN(*y*)). In seeking NN(*x*), and having at some time in the processing a candidate NN, *y′*, we can exclude all y from consideration if we find that $d(x,y') \leq \delta y'$. In this case, we know that we are sufficiently close to *y′* that we cannot improve on it.

We return now to the choice of reference points: Vidal Ruiz [27] proposes the storing of inter-point distances between the members of the training set. Given *x*, whose NN we require, some member of the training set is used as a reference point. Using the bounding approach based on the triangular inequality, described above, allows other training set members to be excluded from any possibility of being NN(*x*). Micó *et al.* [28] and Ramasubramanian and Paliwal [29] discuss further enhancements to this approach, focused especially on the storage requirements. Fukunaga and Narendra [30] make use of both a hierarchical decomposition of the data set (they employ repeatedly the k-means partitioning technique), and bounds based on the triangular inequality. For each node in the decomposition tree, the center and maximum distance to the center of associated points (the "radius") are determined. For 1000 points, 3 levels were used, with a division into 3 classes at each node. All points associated with a non-terminal node can be rejected in the search for the NN of point *x* if the following rule (Rule 1) is not verified:

$$d(x,g) - r_g < \delta$$

where δ is the current NN distance, *g* is the center of the cluster of points associated with the node, and r_g is the radius of this cluster. For a terminal node, which cannot be rejected on the basis of this rule, each associated point, *y*, can be tested for rejection using the following rule (Rule 2):

$$|d(x,g) - d(y,g)| \geq \delta.$$

These two rules are direct consequences of the triangular inequality.

A branch and bound algorithm can be implemented using these two rules. This involves determining some current NN (the bound) and subsequently branching out of a traversal path whenever the current NN cannot be bettered. Not being inherently limited by dimensionality, this approach appears particularly attractive for general purpose applications.

Other rejection rules are considered by Kamgar-Parsi and Kanal. [31] A simpler form of clustering is used in the variant of this algorithm proposed by Niemann and Goppert. [32] A shallow MDBST is used, followed by a variant on the branching and bounding described above.

Bennett *et al.* [2] use the nearest neighbor problem as a means towards solving the Gaussian distribution mixture problem. They consider a preprocessing approach similar to Fukunaga and Narendra [30] but with an important difference: to take better account of cluster structure in the data, the clusters are multivariate normal but not necessarily of diagonal covariance structure. Therefore very elliptical clusters are allowed. This in turn implies that a cluster radius is not of great benefit for establishing a bound on whether or not distances need to be calculated Bennett *et al.* [2] address this problem by seeking a stochastic

guarantee on whether or not calculations can be excluded. Technically, however, such stochastic bounds are not easy to determine in a high dimensional space.

An interesting issue raised in Beyer *et al.* [33] is also discussed by Bennett *et al.* [2] if the ratio of the nearest and furthest neighbor distances converges in probability to 1 as the dimensionality increases, then is it meaningful to search for nearest neighbors? This issue is not all that different from saying that neighbors in an increasingly high dimensional space tend towards being equidistant. In section 5, we will look at approaches for handling particular classes of data of this type.

Fast Approximate Nearest Neighbor Finding

Kushilevitz *et al.*, [34] working in Euclidean and L_1 spaces, propose fast approximate nearest neighbor searching, on the grounds that in systems for content-based image retrieval, approximate results are adequate. Projections are used to bound the search. Probability of successfully finding the nearest neighbor is traded off against time and space requirements.

The Special Case of Sparse Binary Data

"High-dimensional", "sparse" and "binary" are the characteristics of keyword-based bibliographic data, with values possibly in excess of 10000 for both n and m. Such data is usually stored as list data structures, representing the mapping of documents onto index terms, or *vice versa*. Commercial document collections are usually searched using a Boolean search environment. Documents associated with particular terms are retrieved, and the intersection (AND), union (OR) or other operations on such sets of documents are obtained. For efficiency, an *inverted file,* which maps terms onto documents, must be available for Boolean retrieval. The efficient NN algorithms, to be discussed, make use of both the document-term and the term-document files.

The usual algorithm for NN-searching considers each document in turn, calculates the distance with the given document, and updates the NN if appropriate. This algorithm is shown schematically in Fig. 4 (top). The inner loop is simply an expression of the fact that the distance or similarity will, in general, require O(m) calculation: examples of commonly used coefficients are the Jaccard similarity, and the Hamming (L_1 Minkowski) distance.

If $\overline{m}$ and $\overline{n}$ are, respectively, the average numbers of terms associated with a document, and the average number of documents associated with a term, then an average complexity measure, over n searches, of this usual algorithm is O($n\overline{m}$). It is assumed that advantage is taken of some packed form of storage in the inner loop (*e.g.* using linked lists).

Croft's algorithm (see [35] and Fig. 4) is of worst case complexity O(nm^2). However, the number of terms associated with the document whose NN is required will often be quite small. The National Physical Laboratory test collection, for example, which was used by Murtagh [36] has the following characteristics: n = 11429, m = 7491, $\overline{m}$ = 19.9, and $\overline{n}$ = 30.4. The outermost and innermost loops in Croft's algorithm use the document-term file. The center loop uses the term-document inverted file. An average complexity measure (more strictly, the time taken for best match search based on an average document with associated average terms) is seen to be O($\overline{nm}^2$).

```
Usual algorithm:
      Initialize current NN
      For all documents in turn do:
      ... For all terms associated with the document do:
      ... ... Determine (dis)similarity
      ... Endfor
      ... Test against current NN
      Endfor

Croft's algorithm:
      Initialize current NN
      For all terms associated with the given document do:
      ... For all documents associated with each term do:
      ... ... For all terms associated with a document do:
      ... ... ... Determine (dis)similarity
      ... ... Endfor
      ... ... Test against current NN
      ... Endfor
      Endfor

Perry-Willett algorithm:
      Initialize current NN
      For all terms associated with the given document, i, do:
      ... For all documents, i', associated with each term, do:
      ... ... Increment location i' of counter vector
      ... Endfor
      Endfor
```

Figure 4: Algorithms for NN-searching using high-dimensional sparse binary data.

In the outermost loop of Croft's algorithm there will eventually come about a situation where – if a document has not been thus far examined - the number of terms remaining for the given document do not permit the current NN document to be bettered. In this case we can cut short the iterations of the outermost loop. The calculation of a bound using the greatest possible number of terms that could be shared with a so-far unexamined document has been exploited by Smeaton and van Rijsbergen [37] and by Murtagh [36] in successive improvements on Croft's algorithm.

The complexity of all the above algorithms has been measured in terms of operations to be performed. In practice, however, the actual accessing of term or document information may be of far greater cost. The document-term and term-document files are ordinarily stored on direct access file storage because of their large sizes. The strategy used in Croft's algorithm, and in improvements on it, does not allow any viable approaches to batching together the records which are to be read successively, in order to improve accessing-related performance.

The Perry-Willett algorithm (see Perry and Willett, [38]) presents a simple but effective solution to the problem of costly I/O. It focuses on the calculation of the number of terms common to the given document x and each other document, y, in the document collection. This set of values is built up in a computationally efficient fashion. O(n) operations are subsequently required to determine the (dis)similarity, using another vector comprising the total numbers of terms associated with each document. Computation time (the same "average" measure as that used above) is O($\overline{nm}+n$). We now turn our attention to numbers of direct-access reads required.

In Croft's algorithm, all terms associated with the document whose NN is desired may be read in one

read operation. Subsequently, we require $\overline{nm}$ reads, giving in all $1+\overline{nm}$. In the Perry-Willett algorithm, the outer loop again pertains to the one (given) document, and so all terms associated with this document can be read and stored. Subsequently, $\overline{m}$ reads, *i.e.* the average number of terms, each of which demands a read of a set of documents, are required. This gives, in all, $1 + \overline{m}$. Since these reads are very much the costliest operation in practice, the Perry-Willett algorithm can be recommended for large values of n and m. Its general characteristics are that it requires, (i) as do all the algorithms discussed in this section, the availability of the inverted term-document file; and (ii) in-memory storage of two vectors containing *n* integer values.

HIERARCHICAL AGGLOMERATIVE CLUSTERING

The algorithms discussed in this section can be characterized as greedy. [39] A sequence of irreversible algorithm steps is used to construct the desired data structure.

We will not review hierarchical agglomerative clustering here. For essential background, the reader is referred to Murtagh and Heck, [3] Gordon, [40] or Jain and Dubes. [41] This section borrows on Murtagh. [42]

One could practically say that Sibson [43] and Defays [44] are part of the prehistory of clustering. Their $O(n^2)$ implementations of the single link method and of a (non-unique) complete link method, respectively, have been widely cited.

In the early 1980s a range of significant improvements were made to the Lance-Williams, or related, dissimilarity update schema, [45, 46] which had been in wide use since the mid-1960s. Murtagh [47] presents a survey of these algorithmic improvements. We will briefly describe them here. The new algorithms, which have the potential for exactly replicating results found in the classical but more computationally expensive approach, are based on the construction of nearest neighbor chains and reciprocal or mutual NNs (NN-chains and RNNs).

A NN-chain consists of an arbitrary point (a in Fig. 5); followed by its NN (b in Fig. 5); followed by the NN from among the remaining points (c, d, and e in Fig. 5) of this second point; and so on until we necessarily have some pair of points which can be termed reciprocal or mutual NNs. (Such a pair of RNNs may be the first two points in the chain; we have assumed that no two dissimilarities are equal.)

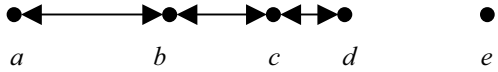


Figure 5: Five points, showing NNs and RNNs.

In constructing a NN-chain, irrespective of the starting point, we may agglomerate a pair of RNNs as soon as they are found. What guarantees that we can arrive at the same hierarchy as we would if we used traditional "stored dissimilarities" or "stored data" algorithms? Essentially this is the same condition as that under which no inversions or reversals are produced by the clustering method. Fig. 6 gives an example of this, where s is agglomerated at a lower criterion value (*i.e.* dissimilarity) than was the case at the previous agglomeration between *q* and *r*.

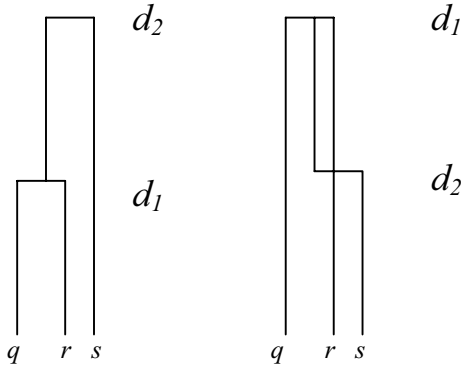


Figure 6: Alternative representations of a hierarchy with an inversion. Assuming dissimilarities, as we go vertically up, criterion values (d_1, d_2) decrease. But here, undesirably, $d_2 > d_1$.

Our ambient space has thus contracted because of the agglomeration. This is due to the algorithm used - in particular the agglomeration criterion - and it is something we would normally wish to avoid.

This is formulated as:

Inversion impossible if

$$d(i,j) < d(i,k) \text{ or } d(j,k) \Rightarrow d(i,j) < d(i \cup j, k)$$

This is essentially Bruynooghe's reducibility property [48] (see also [49]). Using the Lance-Williams dissimilarity update formula, it can be shown that the minimum variance method does not give rise to inversions; neither do the linkage methods; but the median and centroid methods cannot be guaranteed not to have inversions.

To return to Fig. 5, if we are dealing with a clustering criterion that precludes inversions, then c and d can justifiably be agglomerated, since no other point (for example, b or e) could have been agglomerated to either of these.

The processing required, following an agglomeration, is to update the NNs of points such as b in Fig. 5 (and on account of such points, this algorithm was dubbed *algorithme des célibataires* in [45]). The following is a summary of the algorithm:

NN-chain algorithm

Step 1 Select a point arbitrarily.

Step 2 Grow the NN-chain from this point until a pair of RNNs is obtained.

Step 3 Agglomerate these points (replacing with a cluster point, or updating the dissimilarity matrix).

Step 4 From the point which preceded the RNNs (or from any other arbitrary point if the first two points chosen in Steps 1 and 2 constituted a pair of RNNs), return to Step 2 until only one point remains.

In Murtagh [11, 47, 49] and Day and Edelsbrunner, [50] one finds discussions of O(n^2) time and O(n) space implementations of Ward's minimum variance (or error sum of squares) method and of the centroid and median methods. The latter two methods are termed the UPGMC and WPGMC criteria by Sneath and Sokal. [51] Now, a problem with the cluster criteria used by these latter two methods is that the reducibility property is not satisfied by them. This means that the hierarchy constructed may not be unique as a result of inversions or reversals (non-monotonic variation) in the clustering criterion value determined in the sequence of agglomerations. Murtagh [49] describes O(n^2) time and space implementations for the single link method, the complete link method and for the weighted and unweighted group average methods (WPGMA and UPGMA). This approach is quite general *vis á vis* the dissimilarity used and can also be used for hierarchical clustering methods other than those mentioned.

Day and Edelsbrunner [50] prove the exact O(n^2) time complexity of the centroid and median methods using an argument related to the combinatorial problem of optimally packing hyperspheres into an m-dimensional volume. They also address the question of metrics: results are valid in a wide class of distances including those associated with the Minkowski metrics.

The construction and maintenance of the nearest neighbor chain as well as the carrying out of agglomerations whenever reciprocal nearest neighbors meet, both offer possibilities for parallelization. Willet described implementations on an SIMD machine. [52]

Evidently both coordinate data and graph (*e.g.*, dissimilarity) data can be input to these agglomerative methods. Gillet et. al. [53] in the context of clustering chemical structure databases refer to the common use of the Ward method, based on the reciprocal nearest neighbors algorithm, on

data sets of a few hundred thousand molecules.

Applications of hierarchical clustering to bibliographic information retrieval are assessed in Griffiths *et al.* [54] Ward's minimum variance criterion is favored.

From details in White and McCain, [55] the Institute of Scientific Information (ISI) clusters citations (science, and social science) by first clustering highly cited documents based on a single linkage criterion, and then four more passes are made through the data to create a subset of a single linkage hierarchical clustering.

GRAPH CLUSTERING

Hierarchical clustering methods are closely related to graph-based clustering. Firstly, a dendrogram is a rooted labeled tree. Secondly, and more importantly, some methods like the single and complete link methods can be displayed as graphs, and are very closely related to mainstream graph data structures.

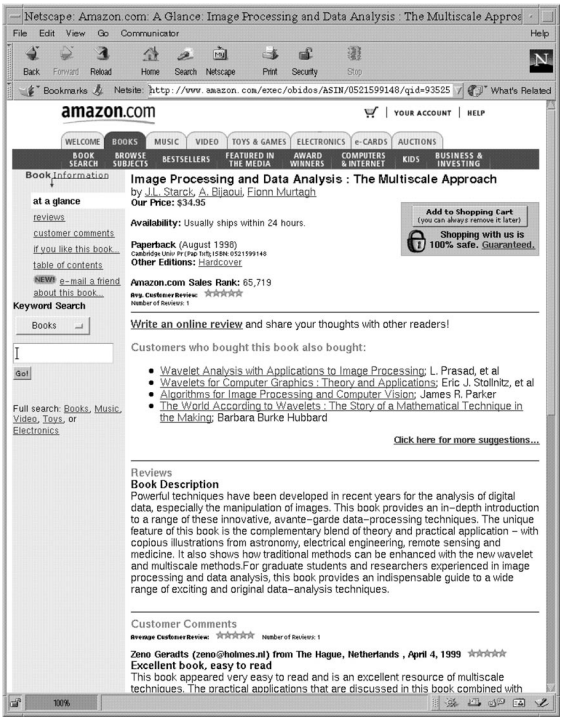


Figure 7: Example of a graph clustering in a data mining perspective at Amazon.com: "Customers who bought this book also bought …"

An example of the increasing prevalence of graph clustering in the context of data mining on the web is presented in Fig. 7: Amazon.com provides information on what other books were purchased by like-minded individuals.

The single link method was referred to in the previous section, as a widely used agglomerative, hence hierarchical, clustering method. Rohlf [56] reviews algorithms for the single link method with complexities ranging from $O(n \log n)$ to $O(n^5)$. The criterion used by the single link method for cluster formation is weak, meaning that noisy data in particular give rise to results that are not robust.

The minimal spanning tree (MST) and the single link agglomerative clustering method are closely related: the MST can be transformed irreversibly into the single link hierarchy. [57] The MST is defined as of minimal total weight, it spans all nodes (vertices) and is an unrooted tree. The MST has been a method of choice for at least four decades now either in its own right for data analysis, [58] as a data structure to be approximated (*e.g.* using shortest spanning paths, see Murtagh, [47], p. 96), or as a basis for clustering. We will look at some fast algorithms for the MST in the remainder of this section.

Perhaps the most basic MST algorithm, due to Prim and Dijkstra, grows a single fragment through n-1 steps. We find the closest vertex to an arbitrary vertex, calling these a fragment of the MST. We determine the closest vertex, not in the fragment, to any vertex in the fragment, and add this new vertex into the fragment. While there are fewer than n vertices in the fragment, we continue to grow it. This algorithm leads to a unique solution. A default $O(n^3)$ implementation is clear, and $O(n^2)$ computational cost is possible ([47], p. 98).

Sollin's algorithm constructs the fragments in parallel. For each fragment in turn, at any stage of the construction of the MST, determine its closest

fragment. Merge these fragments, and update the list of fragments. A tree can be guaranteed in this algorithm (although care must be taken in cases of equal similarity) and our other requirements (all vertices included, minimal total edge weight) are very straightforward. Given the potential for roughly halving the data remaining to be processed at each step, not surprisingly the computational cost reduces from O(n^3) to O(n^2 log n).

The real interest of Sollin's algorithm arises when we are clustering on a graph and do not have all $n(n-1)/2$ edges present. Sollin's algorithm can be shown to have computational cost m log n, where m is the number of edges. When m « $n(n-1)/2$ then we have the potential for appreciable gains.

The MST in feature spaces can of course make use of the fast nearest neighbor finding methods studied earlier in this article. See [47], (section 4.4) for various examples.

Other graph data structures that have been proposed for data analysis are related to the MST. We know, for example, that the following subset relationship holds:

$$MST \subseteq RNG \subseteq GG \subseteq GT$$

where RNG is the relative neighborhood graph, GG is the Gabriel graph, and DT is the Delaunay triangulation. The latter, in the form of its dual, the Voronoi diagram, has been used for analyzing the clustering of galaxy locations. References to these and related methods can be found in Murtagh. [59]

NEAREST NEIGHBOR FINDING ON GRAPHS

Clustering on graphs may be required because we are working with (perhaps complex non-Euclidean) dissimilarities. In such cases where we must take into account an edge between each and every pair of vertices, we will generally have an O(m) computational cost where m is the number of edges.

In a metric space we have seen that we can look for various possible ways to expedite the nearest neighbor search. An approach based on visualization - turning our data into an image - will be looked at below. However, there is another aspect of our similarity (or other) graph that we may be able to turn to our advantage. Efficient algorithms for sparse graphs are available. Sparsity can be arranged - we can threshold our edges if the sparsity does not suggest itself more naturally.

A special type of sparse graph is a planar graph, *i.e.* a graph capable of being represented in the plane without any crossovers of edges. For sparse graphs, algorithms with O(m log log n) computational cost were described by Yao [60] and Cheriton and Tarjan. [61] A short algorithmic description can be found in Murtagh [47] (pp. 107-108) and we refer in particular to the latter. The basic idea is to preprocess the graph, in order to expedite the sorting of edge weights (why sorting? - simply because we must repeatedly find smallest links, and maintaining a sorted list of edges is a good basis for doing this). If we were to sort all edges, the computational requirement would be O(m log m). Instead of doing that, we take the edge set associated with each and every vertex. We divide each such edge set into groups of size k. (The fact that the last such group will usually be of size < k is taken into account when programming.)

Let n_v be the number of incident edges at vertex v, such that $\Sigma_v\, n_v = 2m$. The sorting operation for each vertex now takes O(k log k) operations for each group, and we have n_v/k groups. For all vertices the sorting requires a number of operations which is of the order of $\Sigma_v\, n_v \log k = 2m \log k$. This looks like a questionable - or small - improvement over O(m log m). Determining the lightest edge incident on a vertex requires O(n_v/k) comparisons since we have to check all groups. Therefore the lightest edges incident on all vertices are found with O(m/k)

operations.

When two vertices, and later fragments, are merged, their associated groups of edges are simply collected together, therefore keeping the total number of groups of edges that we started out with.

We will bypass the issue of edges which, over time, are to be avoided because they connect vertices in the same fragment: given the fact that we are building an MST, the total number of such edges-to-be-avoided cannot surpass 2m. To find what to merge next, again O(m/k) processing is required. Using Sollin's algorithm, the total processing required in finding what to merge next is O(m/k log n). The total processing required for grouping the edges, and sorting within the edge-groups, is O(m log k), *i.e.* it is one-off and accomplished at the start of the MST-building process.

The total time is O(m/k log n)+O(m log k). If we fix k = log n, the second term dominates and gives overall computational complexity as O(m log log n). This result has been further improved to near linearity in m by Gabow *et al.*, [62] who develop an algorithm with complexity O(m log log log … n) where the number of iterated log terms is bounded by m/n.

Motwani and Raghavan [63] (chapter 10) base a stochastic O(m) algorithm for the MST on random sampling to identify and eliminate edges that are guaranteed not to belong to the MST.

Let us turn our attention now to the case of a planar graph. For a planar graph we know that $m \leq 3n$-6 for m >1. (For proof, see for example Tucker, [64] or any book on graph theory).

Referring to Sollin's algorithm, described above, O(n) operations are needed to establish a least cost edge from each vertex, since there are only O(n) edges present. On the next round, following fragment-creation, there will be at most ceil(n/2) new vertices, implying of the order of n/2 processing to find the least cost edge. The total computational cost is seen to be proportional to: n + n/2 + n/4 +…=O(n).

So determining the MST of a planar graph is linear in numbers of either vertices or edges. Before ending this review of very efficient clustering algorithms for graphs, we note that algorithms discussed so far have assumed that the similarity graph was undirected. For modeling transport flows, or economic transfers, the graph could well be directed. Components can be defined, generalizing the clusters of the single link method, or the complete link method. [65] provides an algorithm for the latter agglomerative criterion which is of computational cost O(m log n).

K-MEANS AND FAMILY

The non-technical person more often than not understands clustering as a partition. K-means looked at in this section, or the distribution mixture approach looked at in the section on fast model-based clustering, provide solutions. A mathematical definition of a partition implies no multiple assignments of observations to clusters, *i.e.* no overlapping clusters. Overlapping clusters may be faster to determine in practice, and a case in point is the one-pass algorithm described in Salton and McGill. [66] The general principle followed is: make one pass through the data, assigning each object to the first cluster which is close enough, and making a new cluster for objects that are not close enough to any existing cluster.

Broder *et al.* [67] use this algorithm for clustering the web. A feature vector is determined for each HTML document considered, based on sequences of words. Similarity between documents is based on an inverted list, using an approach like those described for the special case of binary data above. The similarity graph is thresholded, and components sought.

Broder [68] solves the same clustering objective

using a thresholding and overlapping clustering method similar to the Salton and McGill one. The application described is that of clustering the Altavista repository in April 1996, consisting of 30 million HTML and text documents, comprising 150 GBytes of data. The number of serviceable clusters found was 1.5 million, containing 7 million documents. Processing time was about 10.5 days. An analysis of the clustering algorithm used by Broder can be found in Borodin *et al.*, [69] who also consider the use of approximate minimal spanning trees.

The threshold-based pass of the data, in its basic state, is susceptible to lack of robustness. A bad choice of threshold leads to too many clusters or two few. To remedy this, we can work on a well-defined data structure such as the minimal spanning tree. Or, alternatively, we can iteratively refine the clustering. Partitioning methods, such as k-means, use iterative improvement of an initial estimation of a targeted clustering.

A very widely used family of methods for inducing a partition on a data set is called k-means, c-means (in the fuzzy case), Isodata, competitive learning, vector quantization and other more general names (non-overlapping non-hierarchical clustering) or more specific names (minimal distance or exchange algorithms).

The usual criterion to be optimized is:

$$\frac{1}{|I|}\sum_{q\in Q}\sum_{i\in q}\left\|\vec{i}-\vec{q}\right\|^2$$

where I is the object set, $|.|$ denotes cardinality, q is some cluster, Q is the partition, and q denotes a set in the summation, whereas $\vec{q}$ denotes some associated vector in the error term, or metric norm.

This criterion ensures that clusters found are compact, and therefore assumed homogeneous. The optimization criterion, by a small abuse of terminology, is more often referred to as a minimum variance one. A necessary condition that this criterion be optimized is that vector $\vec{q}$ be a cluster mean, which for the Euclidean metric case is:

$$\vec{q}=\frac{1}{|q|}\sum_{i\in q}\vec{i}$$

A batch update algorithm, due to Lloyd, [70] Forgy, [71] and others, makes assignments to a set of initially randomly chosen vectors, $\vec{q}$, as step 1. Step 2 updates the cluster vectors, $\vec{q}$. This is iterated. The distortion error, equation 1, is non-increasing, and a local minimum is achieved in a finite number of iterations.

An online update algorithm is due to MacQueen. [72] After each presentation of an observation vector, $\vec{i}$, the closest cluster vector, $\vec{q}$, is updated to take account of it. Such an approach is well-suited for a continuous input data stream (implying "online" learning of cluster vectors).

Both algorithms are gradient descent ones. In the online case, much attention has been devoted to best learning rate schedules in the neural network (competitive learning) literature: Darken and Moody [73, 74], Darken *et al.*, [75] Fritzke. [76]

A difficulty, less controllable in the case of the batch algorithm, is that clusters may become (and stay) empty. This may be acceptable, but also may be in breach of our original problem formulation. An alternative to the batch update algorithm is Späth's exchange algorithm. [77] Each observation is considered for possible assignment into any of the other clusters. Späth gives updating and "downdating" formulae. This exchange algorithm is stated to be faster to converge and to produce better (smaller) values of the objective function. Over decades of use, we have also verified that it is a superior algorithm to the minimal distance one.

K-means is very closely related to Voronoi

(Dirichlet) tesselations, to Kohonen self-organizing feature-maps, and various other methods. The batch-learning algorithm above may be viewed as

1. An assignment step, which we will term the E (estimation) step: estimate the posteriors,

   P(observations | cluster centers)

2. A cluster update step, the M (maximization) step, which maximizes a cluster center likelihood.

Neal and Hinton [78] cast the k-means optimization problem in such away that the both E- and M-steps monotonically increase the maximand's values. The EM algorithm may, too, be enhanced to allow for online as well as batch learning. [79]

In Thiesson *et al*., [80] k-means is implemented (i) by traversing blocks of data, cyclically, and incrementally updating the sufficient statistics and parameters, and (ii) instead of cyclic traversal, sampling from subsets of the data is used. Such an approach is admirably suited for very large data sets, where in-memory storage is not feasible. Examples used by Thiesson *et al*. [80] include the clustering of a half million 300-dimensional records.

FAST MODEL BASED CLUSTERING

It is traditional to note that models and (computational) speed do not mix. We review recent progress in this section.

Modeling of Signal and Noise

A simple and applicable model is a distribution mixture, with the signal modeled by Gaussians, in the presence of Poisson background noise.

Consider data which are generated by a mixture of (G-1) bivariate Gaussian densities, $f_k(x;\theta)\sim N(\mu_k;\Sigma_k)$, for clusters $k = 2; \ldots ;$ G, and with Poisson background noise corresponding to $k = 1$. The overall population thus has the mixture density

$$f(x;\theta) = \sum_{k=1}^{G} \pi_k f_k(x;\theta)$$

where the mixing or prior probabilities, π_k, sum to 1, and $f_1(x;\theta) = A^{-1}$, where A is the area of the data region. This is the basis for model-based clustering. [81-84]

The parameters, θ and π, can be estimated efficiently by maximizing the mixture likelihood

$$L(\theta,\pi) = \prod_{i=1}^{n} f(x_i;\theta),$$

with respect to θ and π, where x_i is the i^{th} observation.

Now let us assume the presence of two clusters, one of which is Poisson noise, the other Gaussian. This yields the mixture likelihood

$$L(\theta,\pi) = \prod_{i=1}^{n}\left[\pi_1 A^{-1} + \pi_2 \frac{1}{2\pi\sqrt{|\Sigma|}}\exp\left\{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right\}\right],$$

where $\pi_1 + \pi_2 = 1$.

An iterative solution is provided by the expectation-maximization (EM) algorithm of Dempster *et al*. [85] We have already noted this algorithm in informal terms in the last section, dealing with *k*-means. Let the "complete" (or "clean" or "output") data be $y_i = (x_i, z_i)$ with indicator set $z_i = (z_{i1}, z_{i2})$ given by (1,0) or (0,1). Vector z_i has a multinomial distribution with parameters $(1;\pi_1,\pi_2)$. This leads to the complete data log-likelihood:

$$l(y,z;\theta,\pi) = \sum_{i=1}^{n}\sum_{k=1}^{2} z_{ik}\left[\log\pi_k + \log f_k(x_k;\theta\right]$$

The E-step then computes $\hat{z}_{ik} = E(z_{ik}|x_1,\ldots,x_n,\theta)$, *i.e.* the posterior probability that x_i is in cluster k. The M-step involves maximization of the expected complete data log-likelihood:

$$l^*(y;\theta,\pi) = \sum_{i=1}^{n}\sum_{k=1}^{2} \hat{z}_{ik}\left[\log\pi_k + \log f_k(x_i;\theta)\right].$$

The E- and M-steps are iterated until convergence.

For the 2-class case (Poisson noise and a Gaussian cluster), the complete-data likelihood is

$$L(y,z;\theta,\pi) = \prod_{i=1}^{n}\left[\frac{\pi_1}{A}\right]^{zi1}\left[\frac{\pi_2}{2\pi\sqrt{|\Sigma|}}\exp\left\{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right\}\right]^{zi2}$$

The corresponding expected log-likelihood is then used in the EM algorithm. This formulation of the problem generalizes to the case of G clusters, of arbitrary distributions and dimensions.

Fraley [86] discusses implementation of model-based clustering, including publicly available software.

In order to assess the evidence for the presence of a signal-cluster, we use the Bayes factor for the mixture model, M_2 that includes a Gaussian density as well as background noise, against the "null" model, M_1, that contains only background noise. The Bayes factor is the posterior odds for the mixture model against the pure noise model, when neither is favored *a priori*. It is defined as $B = p(x|M_2)/p(x|M_1)$, where $p(x|M_2)$ is the integrated likelihood of the mixture model M_2, obtained by integrating over the parameter space. For a general review of Bayes factors, their use in applied statistics, and how to approximate and compute them, see Kass and Raftery. [87]

We approximate the Bayes factor using the Bayesian Information Criterion (BIC). [88] For a Gaussian cluster and Poisson noise, this takes the form:

$$2\log B \approx BIC = 2\log L(\hat{\theta},\hat{\pi}) + 2n\log A - 6\log n,$$

where $\hat{\theta}$ and $\hat{\pi}$ are the maximum likelihood estimators of θ and π, and $L(\hat{\theta},\hat{\pi})$ is the maximized mixture likelihood.

A review of the use of the BIC criterion for model selection - and more specifically for choosing the number of clusters in a data set - can be found in Fraley and Raftery. [89]

An application of mixture modeling and the BIC criterion to gamma-ray burst data can be found in Mukherjee *et al.* [90] So far around 800 observations have been assessed, but as greater numbers become available we will find the inherent number of clusters in a similar way, in order to try to understand more about the complex phenomenon of gamma-ray bursts.

Application to Thresholding

Consider an image or a planar or 3-dimensional set of object positions. For simplicity we consider the case of setting a single threshold in the image intensities, or the point set's spatial density.

We deal with a combined mixture density of two univariate Gaussian distributions $f_k(x,\theta) \sim N(\mu_k,\sigma_k)$. The overall population thus has the mixture density

$$f(x;\theta) = \sum_{k=1}^{2}\pi_k f_k(x;\theta)$$

where the mixing or prior probabilities, π_k, sum to 1.

When the mixing proportions are assumed equal, the log-likelihood takes the form

$$l(\theta) = \sum_{i=1}^{n}\ln\left[\sum_{k=1}^{2}\frac{1}{2\pi\sqrt{|\sigma_k|}}\exp\left\{-\frac{1}{2\sigma_k}(x_i-\mu_k)^2\right\}\right]$$

The EM algorithm is then used to iteratively solve this (see Celeux and Govaert, [91]). This method is used for appraisals of textile (jeans and other fabrics) fault detection in Campbell *et al.* [92]. Industrial vision inspection systems potentially produce large data streams, and fault detection can be a good application for fast clustering methods. We are currently using a mixture model of this sort on SEM (scanning electron microscope) images of cross-sections of concrete to allow for subsequent characterization of physical properties.

Image segmentation, *per se*, is a relatively straightforward application, but there are novel and interesting aspects to the two studies mentioned. In the textile case, the faults are very often perceptual and relative, rather than "absolute" or capable of

being analyzed in isolation. In the SEM imaging case, a first phase of processing is applied to de-speckle the images, using multiple resolution noise filtering.

Turning from concrete to cosmology, the Sloan Digital Sky Survey [92] is producing a sky map of more than 100 million objects, together with 3-dimensional information (redshifts) for a million galaxies. Pelleg and Moore [18] describe mixture modeling, using a k-D tree preprocessing to expedite the finding of the class (mixture) parameters, *e.g.* means, covariances.

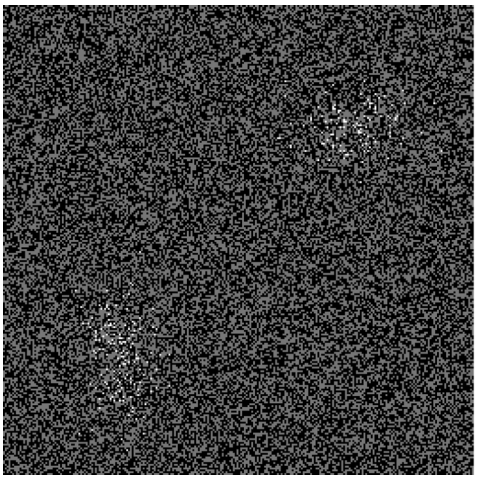


Figure 8: Data in the plane. The 256 × 256 image shows 550 "signal" points – two Gaussian-shaped clusters in the lower left and in the upper right – with in addition 40,000 Poisson noise points added. Details of recovery of the clusters are discussed in [95].

NOISE MODELING

In Starck *et al*. [93] and in a wide range of papers, we have pursued an approach for the noise modeling of observed data. A multiple resolution scale vision model or data generation process is used, to allow for the phenomenon being observed on different scales. In addition, a wide range of options are permitted for the data generation transfer path, including additive and multiplicative, stationary and non-stationary, Gaussian ("read out" noise), Poisson (random shot noise), and so on.

Given point pattern clustering in two- or three-dimensional spaces, we will limit our overview here to the Poisson noise case.

Poisson Noise with Few Events Using the à trous Transform

If a wavelet coefficient $w_j(x,y)$ is due to noise, it can be considered as a realization of the sum $\sum_{k\in K} n_k$ of independent random variables with the same distribution as that of the wavelet function (n_k being the number of events used for the calculation of $w_j(x,y)$). This allows comparison of the wavelet coefficients of the data with the values that can be taken by the sum of n independent variables.

The distribution of one event in wavelet space is then directly given by the histogram H_1 of the wavelet ψ. As we consider independent events, the distribution of a coefficient w_n (note the changed subscripting for w, for convenience) related to n events is given by n autoconvolutions of H_1 :

$$H_n = H_1 \otimes H_1 \otimes ... \otimes H_1$$

For a large number of events, H_n converges to a Gaussian. Fig. 8 shows an example of where point pattern clusters - density bumps in this case – are sought, with a great amount of background clutter. Murtagh and Starck [94] refer to the fact that there is no computational dependence on the number of points (signal or noise) in such a problem, when using a wavelet transform with noise modeling.

Some other alternative approaches will be briefly noted. The Haar transform presents the advantage of its simplicity for modeling Poisson noise. Analytic formulae for wavelet coefficient distributions have been derived by Kolaczyk, [96] and Jammal and Bijaoui. [97] Using a new wavelet transform, the Haar *à trous* transform, Zheng *et al.* [98] appraise a denoising approach for financial data streams, - an important preliminary step for

subsequent clustering, forecasting, or other processing.

Poisson Noise with Nearest Neighbor Clutter Removal

The wavelet approach is certainly appropriate when the wavelet function reflects the type of object sought (*e.g.* isotropic), and when superimposed point patterns are to be analyzed. However, non-superimposed point patterns of complex shape are very well treated by the approach described in Byers and Raftery. [99] Using a homogeneous Poisson noise model, they derive the distribution of the distance of a point to its k^{th} nearest neighbor.

Next, Byers and Raftery [99] consider the case of a Poisson process which is signal, superimposed on a Poisson process which is clutter. The k^{th} nearest neighbor distances are modeled as a mixture distribution: a histogram of these, for given k, will yield a bimodal distribution if our assumption is correct. This mixture distribution problem is solved using the EM algorithm. Generalization to higher dimensions, *e.g.* 10, is also discussed.

Similar data were analyzed by noise modeling and a Voronoi tesselation preprocessing of the data in Allard and Fraley. [100] It is pointed out there how this can be a very useful approach with the Voronoi tiles have meaning in relation to the morphology of the point patterns. However, it does not scale well to higher dimensions, and the statistical noise modeling is approximate.

Ebeling and Wiedenmann, [101] reproduced in Dobrzycki *et al.*, [102] propose the use of a Voronoi tesselation for astronomical X-ray object detection and characterization.

CLUSTER-BASED USER INTERFACES

Doyle first described Information retrieval by means of "semantic road maps" in detail. [103] The spatial metaphor is a powerful one in human information processing. The spatial metaphor also lends itself well to modern distributed computing environments such as the web. The Kohonen self-organizing feature map (SOM) method is an effective means towards this end of a visual information retrieval user interface. We will also provide an illustration of web-based semantic maps based on hyperlink clustering.

The Kohonen map is, at heart, k-means clustering with the additional constraint that cluster centers be located on a regular grid (or some other topographic structure) and furthermore their location on the grid be monotonically related to pairwise proximity. [104] The nice thing about a regular grid output representation space is that it lends itself well as a visual user interface.

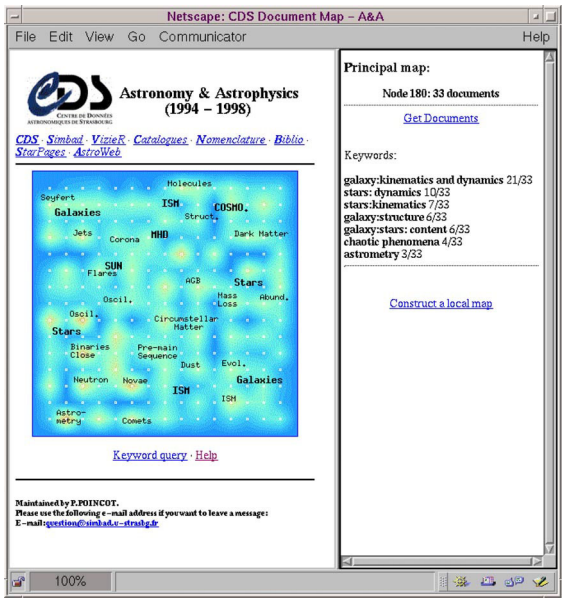


Figure 9: Visual interactive user interface to the journal *Astronomy and Astrophysics*.

Fig. 9 shows a visual and interactive user interface map, using a Kohonen self-organizing feature map (SOM). Color is related to density of document clusters located at regularly-spaced nodes of the map, and some of these nodes/clusters are annotated. The map is installed as a clickable

image-map, with CGI programs accessing lists of documents and - through further links - in many cases, the full documents. In the example shown, the user has queried a node and results are seen in the right-hand panel. Such maps are maintained for (currently) 12000 articles from the Astrophysical Journal, 7000 from Astronomy and Astrophysics, over 2000 astronomical catalogs, and other data holdings. More information on the design of this visual interface and user assessment can be found in Poinçot *et al*. [105, 106]

Guillaume [107] developed a Java-based visualization tool for hyperlink-based data, consisting of astronomers' names, astronomical object names, article titles, and with the possibility of other objects (images, tables, *etc.*). Through weighting, the various types of links could be prioritized. An iterative refinement algorithm was developed to map the nodes (objects) to a regular grid of cells, which as for the Kohonen SOM map, are clickable and provide access to the data represented by the cluster. Fig. 10 shows an example for an astronomer (Prof. Jean Heyvaerts, Strasbourg Astronomical Observatory).

These new cluster-based visual user interfaces are not computationally demanding. They are not however, scalable in their current implementation. Document management (see *e.g.* Cartia, [108]) is not so much the motivation, but rather the interactive user interface.

IMAGES FROM DATA

It is quite impressive how 2D (or 3D) image signals can handle with ease the scalability limitations of clustering and many other data processing operations. The contiguity imposed on adjacent pixels bypasses the need for nearest neighbor finding. It is very interesting therefore to consider the feasibility of taking problems of clustering massive data sets into the 2D image domain. We will look at a few recent examples of work in this direction.

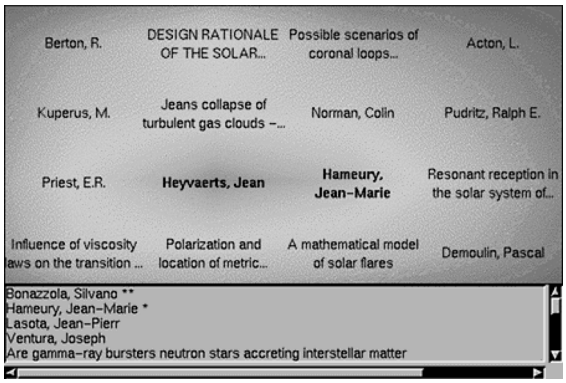


Figure 10: Visual interactive user interfaces, based on graph edges. Map for astronomer Jean Heyvaerts. Original in color.

Church and Helfman [109] address the problem of visualizing possibly millions of lines of computer program code, or text. They consider an approach borrowed from DNA sequence analysis. The data sequence is tokenized by splitting it into its atoms (line, word, character, *etc.*) and then placing a dot at position i,j if the i^{th} input token is the same as the j^{th}. The resulting dotplot, it is argued, is not limited by the available display screen space, and can lead to discovery of large-scale structure in the data.

When data do not have a sequence we have an invariance problem that can be resolved by finding some row and column permutation which pulls large array values together, and perhaps furthermore into proximity to an array diagonal. Berry *et al*. [110] have studied the case of large sparse arrays. Gathering larger (or nonzero) array elements to the diagonal can be viewed in terms of minimizing the envelope of nonzero values relative to the diagonal. This can be formulated and solved in purely symbolic terms by reordering vertices in a suitable graph representation of the matrix. A widely used method for symmetric sparse matrices is the Reverse Cuthill-McKee (RCM) method.

The complexity of the RCM method for ordering rows or columns is proportional to the product of the maximum degree of any vertex in the graph

representing the array values and the total number of edges (nonzeroes in the matrix). For hypertext matrices with small maximum degree, the method would be extremely fast. The strength of the method is its low time complexity but it does suffer from certain drawbacks. The heuristic for finding the starting vertex is influenced by the initial numbering of vertices and so the quality of the reordering can vary slightly for the same problem for different initial numberings. Next, the overall method does not accommodate dense rows (*e.g.*, a common link used in every document), and if a row has a significantly large number of nonzeroes it might be best to process it separately; *i.e.*, extract the dense rows, reorder the remaining matrix and augment fit by the dense rows (or common links) numbered last. Elapsed CPU times for a range of arrays and permuting methods are given in Berry *et al*., [110] and as an indication show performances between 0.025 to 3.18 seconds for permuting a 4000 x 400 array.

A review of public domain software for carrying out SVD and other linear algebra operations on large sparse data sets can be found in Berry *et al.* ([111], section 8.3).

Once we have a sequence-respecting array, we can immediately apply efficient visualization techniques from image analysis. Murtagh *et al*. [112] investigate the use of noise filtering (*i.e.* to remove less useful array entries) using a multiscale wavelet transform approach.

An example follows. From the Concise Columbia Encyclopedia (1989 2nd ed., online version) a set of data relating to 12025 encyclopedia entries and to 9778 cross-references or links was used.

Fig. 11 shows a 500 x 450 subarray, based on a correspondence analysis (*i.e.* ordering of projections on the first factor).

This part of the encyclopedia data was filtered using the wavelet and noise-modeling methodology described in Murtagh *et al*. [112] and the outcome is shown in Fig. 12. Overall the recovery of the more apparent alignments, and hence visually stronger clusters, is excellent. The first relatively long “horizontal bar” was selected - it corresponds to column index (link) 1733 = geological era.

The corresponding row indices (articles) are, in sequence:

SILURIAN PERIOD
PLEISTOCENE EPOCH
HOLOCENE EPOCH
PRECAMBRIAN TIME
CARBONIFEROUS PERIOD
OLIGOCENE EPOCH

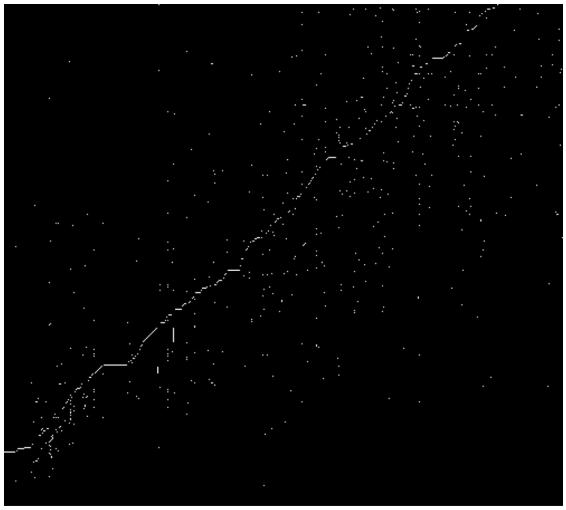


Figure 11: Part (500 × 450) of original encyclopaedia incidence data array.

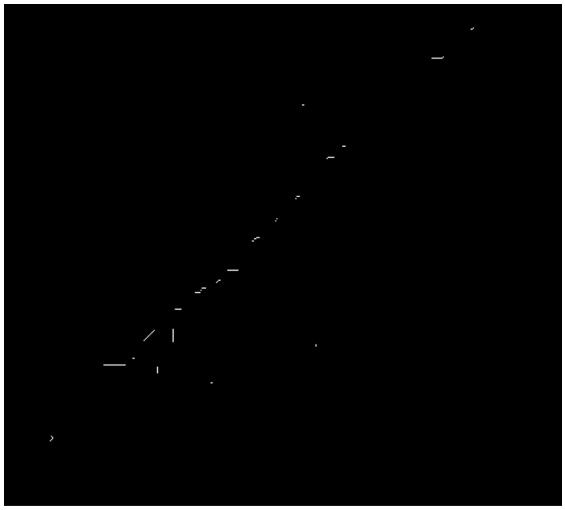


Figure 12: End-product of the filtering of the array shown in Figure 11.

ORDOVICIAN PERIOD
TRIASSIC PERIOD
CENOZOIC ERA
PALEOCENE EPOCH
MIOCENE EPOCH
DEVONIAN PERIOD
PALEOZOIC ERA
JURASSIC PERIOD
MESOZOIC ERA
CAMBRIAN PERIOD
PLIOCENE EPOCH
CRETACEOUS PERIOD

The work described here is based on a number of technologies: (i) data visualization techniques; (ii) the wavelet transform for data analysis; and (iii) data matrix permuting techniques. The wavelet transform has linear computational cost in terms of image row and column dimensions, and is independent of the pixel values.

CONCLUSIONS

Viewed from a commercial or managerial perspective, one could justifiably ask where we are now in our understanding of problems in this area relative to where we were back in the 1960s? Depending on our answer to this, we may well proceed to a second question: Why have all important problems not been solved by now in this area - are there major outstanding problems to be solved?

As described in this chapter, a solid body of experimental and theoretical results has been built up over the last few decades. Clustering remains a requirement that is a central infrastructural element of very many application fields.

There is continual renewal of the essential questions and problems of clustering, relating to new data, new information, and new environments. There is no logjam in clustering research and development simply because the rivers of problems continue to broaden and deepen. Clustering and classification remain quintessential issues in our computing and information technology environments. [113]

ACKNOWLEDGMENTS

Some of this work, in particular in the sections on Nearest Neighbor Finding on Graphs, K-Means and Family and Fast Model-Based Clustering, represents various collaborations with the following: J.L. Starck, CEA; A. Raftery, University of Washington; C. Fraley, University of Washington and MathSoft Inc.; D. Washington, MathSoft, Inc.; Ph. Poinçot and S. Lesteven, Strasbourg Observatory; D. Guillaume, Strasbourg Observatory, University of Illinois and NCSA.

LITERATURE AND NOTES

[1] Dasarathy, B.V., *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, IEEE Computer Society Press, New York, **1991**.

[2] Bennett, K.P., Fayyad, U. and Geiger, D., *Density-based indexing for approximate nearest neighbor queries*, Microsoft Research Technical Report MSR-TR-98-58, **1999**.

[3] Murtagh, F.; and Heck, A. *Multivariate Data Analysis*, Kluwer Academic, Dordrecht, **1987**.

[4] Arabie, P.; Hubert, L. J.; De Soete, G.; Eds., *Clustering and Classification*, World Scientific, Singapore, **1996**.

[5] Arabie, P.; Hubert, L. J. *An Overview of Combinatorial Data Analysis*, in Arabie, P.; Hubert, L. J.; De Soete, G. Eds, *Clustering and Classification*, World Scientific, Singapore, **1996**, 5.

[6] Murtagh, F. *Search Algorithms for Numeric and Quantitative Data* in Heck, A.; Murtagh, F. Eds, *Intelligent Information Retrieval: The Case of Astronomy and Related Space Sciences*, Kluwer Academic, Dordrecht, **1993**, 49.

[7] Delannoy, C. *RAIRO Informatique/Computer Science*, **1980**, 14, 275.

[8] Bentley, J. L.; Weide, B. W.; Yao, A. C. *ACM Transactions on Mathematical Software*, **1980**, 6, 563.

[9] Schreiber, T. *Efficient Search for Nearest Neighbors*, in Weigend, A. S.; Gershenfeld, N. A. Eds, *Predicting the Future and Understanding the Past: A Comparison of*

Approaches, Addison-Wesley, New York, **1993**.

[10] Rohlf, F. J. *Information Processing Letters*, **1978**, 7, 44.

[11] Murtagh, F. *Information Processing Letters*, **1983**, 16, 237.

[12] Friedman, J. H.; Bentley, J. L.; Finkel, R. A. *ACM Transactions on Mathematical Software*, **1977**, 3, 209.

[13] Bentley, J. L.; Friedman, J. H.*IEEE Transactions on Computers*, **1978**, C-27, 97.

[14] Broder, A. J. *Pattern Recognition*, **1990**, 23, 171.

[15] Weiss, S. F. *A Probabilistic Algorithm for Nearest Neighbor Searching*, in R.N. Oddy, R. N. *et al.*, Eds, *Information Retrieval Research,* Butterworths, London, **1981**, 325.

[16] Eastman, C. M.; Weiss, S. F. *Information Systems*, **1982**, 7, 115.

[17] Moore, A. *Advances in Neural Information Processing Systems*, 11, **1999**.

[18] Pelleg, D.; Moore, A. *Accelerating exact k-means algorithms with geometric reasoning*, Proceedings KDD-99, Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August, San Diego, **1999**.

[19] Friedman, J. H.; Baskett, F.; Shustek, L. J. *IEEE Transactions on Computers*, **1975**, C-24, 1000.

[20] Marimont, R. B.; Shapiro, M. B. *Journal of the Institute of Mathematics and its Applications*, **1979**, 24, 59.

[21] Kittler, J. *Kybernetes,* **1978**, 7, 313.

[22] Yunck, T. P. *IEEE Transactions on Systems, Man, and Cybernetics*, **1976**, SMC-6, 678.

[23] Richetin, M.; Rives, G.; Naranjo, M. *RAIRO Informatique/Computer Science*, **1980**, 14, 369.

[24] Burkhard, W. A.; Keller, R. M. *Communications of the ACM*, **1973**, 16, 230.

[25] Shapiro, M. *Communications of the ACM*, **1977**, 20, 339.

[26] Hodgson, M.E., *Remote Sensing of Environment*, **1988**, 25, 117.

[27] Vidal Ruiz, E. *Pattern Recognition Letters*, **1986**, 4, 145.

[28] Micó, L.; Oncina, J.; Vidal, E. An algorithm for finding nearest neighbors in constant average time with a linear space complexity, in *11th International Conference on Pattern Recognition, Volume II*, IEEE Computer Science Press, New York, **1992**, 557.

[29] Ramasubramanian, V.; Paliwal, K. K., *Pattern Recognition Letters*, **1992**, 13, 471.

[30] Fukunaga, K.; Narendra, P. M. *IEEE Transactions on Computers*, **1975**, C-24, 750.

[31] Kamgar-Parsi, B.; Kanal, L. N. *Pattern Recognition Letters*, **1985**, 3, 7.

[32] Niemann, H.; Goppert, R., *Pattern Recognition Letters*, **1988**, 7, 67.

[33] Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When is nearest neighbor meaningful?, in *Proceedings of the 7th International Conference on Database Theory (ICDT)*, Jerusalem, Israel, **1999**.

[34] Kushilevitz, E.; Ostrovsky, R.; Rabani, Y. Efficient search for approximate nearest neighbors in high-dimensional spaces", *Proc. of 30th ACM Symposium on Theory of Computing (STOC-30)*, **1998**.

[35] Croft, W. B. *Journal of the American Society for Information Science*, **1977**, 28, 341.

[36] Murtagh, F. *Information Processing Letters*, **1983**, 16, 237.

[37] Smeaton, A. F.; van Rijsbergen, C. J. *ACM SIGIR Forum*, **1981**, 16, 83.

[38] Perry, S. A.; Willett, P. *Journal of Information Science*, **1983**, 6, 59.

[39] Horowitz, E.; Sahni, S. *Fundamentals of Computer Algorithms*, Chapter 4 *The Greedy Method*, Pitman, London, **1979**.

[40] Gordon, A. D. *Classification*, 2nd ed., Champman and Hall, **1999**.

[41] Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, **1988**.

[42] Murtagh, F. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1992**, 14, 1056.

[43] Sibson, R. *Computer Journal*, **1973**, 16, 30.

[44] Defays, D. *Computer Journal*, **1977**, 20, 364.

[45] de Rham, C. *Les Cahiers de l'Analyse des Données*, **1980**, V, 135.

[46] Juan, J. *Les Cahiers de l'Analyse des Données*, 1982, VII, 219.

[47] Murtagh, F. *Multidimensional Clustering Algorithms*, Physica-Verlag, Würzburg, **1985**.

[48] Bruynooghe, M. Statistique et Analyse des Données, **1977**, no. 3, 24.

[49] Murtagh, F. *Computational Statistics Quarterly*, **1984**, 1, 101.

[50] Day, W.H.E.; Edelsbrunner, H. *Journal of Classification*, **1984**, 1, 7..

[51] Sneath, P.H.A.; Sokal, R.R., *Numerical Taxonomy*, W.H. Freeman, San Francisco, **1973**.

[52] Willett, P. *Journal of Documentation*, **1989**, 45, 1.

[53] Gillet, V. J.; Wild, D. J.; Willett, P.; Bradshaw, J. *The Computer Journal*, **1998**, 41, 547.

[54] Griffiths, A.; Robinson, L. A.; Willett, P. *Journal of Documentation*, **1984**, 40, 175.

[55] White, H. D.; McCain, K. W.; in M.E. Williams, Ed., *Annual Review of Information*

Science and Technology (ARIST), **1997**, Vol. 32, 99.

[56] Rohlf, F. J.; *Information Processing Letters*, **1978**, 7, 44.

[57] Rohlf, F. J.; *The Computer Journal*, **1973**, 16, 93.

[58] Zahn, C. T. *IEEE Transactions on Computers*, **1971**, C-20, 68.

[59] Murtagh, F. in Sandqvist, Aa.; Ray, T. P. Eds., *Central Activity in Galaxies: From Observational Data to Astrophysical Diagnostics*, Springer-Verlag, Berlin, **1993**, pp. 209-235.

[60] Yao, A. C. *Information Processing Letters*, **1975**, 4, 21.

[61] Cheriton, D.; Tarjan, D. E. *SIAM Journal on Computing*, **1976**, 5, 724.

[62] Gabow, H. N.; Galil, Z.; Spencer, T.; Tarjan, R. E. *Combinatorica*, **1986**, 6, 109.

[63] Motwani, R.; Raghavan, P. *Randomized Algorithms*, Cambridge University Press, **1995**.

[64] Tucker, A. *Applied Combinatorics*, Wiley, New York**, 1980**.

[65] Tarjan, R. E.*Information Processing Letters*, **1983**, 17, 37.

[66] Salton, G.; McGill, M. J. *Introduction to Modern Information Retrieva*l, McGraw-Hill, New York, **1983**.

[67] Broder, A. Z.; Glassman, S. C.; Manasse, M. S.; Zweig, G.;*Proc. Sixth International World Wide Web Conference*, **1997**, 391.

[68] Broder, A. Z. In Compression and Complexity of Sequences (SEQUENCES'97), pp. 21-29, IEEE Computer Society, **1998**.

[69] Borodin, A.; Ostrovsky, R.; Rabani, Y. "Subquadratic approximation algorithms for clustering problems in high dimensional spaces", *Proc. 31st ACM Symposium on Theory of Computing (STOC-99)*, **1999**.

[70] Lloyd, P. "Least squares quantization in PCM." Technical note, Bell Laboratories, **1957**. Published in *IEEE Transactions on Information Theory*, **1982**.

[71] Forgy, E. *Biometrics*, **1965**, 21, 768.

[72] MacQueen, J., Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281-297, Berkeley, University of California Press, **1976**.

[73] Darken, C.; Moody, J. "Note on learning rate schedules for stochastic optimization", *Advances in Neural Information Processing Systems 3*, Morgan Kaufmann, Palo Alto, **1991**.

[74] Darken, C.; Moody, J. "Towards faster stochastic gradient search", *Advances in Neural Information Processing Systems 4*, Morgan Kaufman, San Mateo, **1992**.

[75] Darken, C.; Chang, J.; Moody, J. "Learning rate schedules for faster stochastic gradient search", *Neural Networks for Signal Processing 2, Proceedings of the 1992 IEEE Worksho*p, IEEE Press, Piscataway, **1992**.

[76] Fritzke, B., "Some competitive learning methods", http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/JavaPaper

[77] Späth, H. *Cluster Dissection and Analysis: Theory, Fortran Programs, Examples*, Ellis Horwood, Chichester**, 1985**.

[78] Neal, R; Hinton, G. in M. Jordan, Ed., *Learning in Graphical Models*, Kluwer, Dordrecht, **1998**, pp. 355-371.

[79] Sato, M.; Ishii, S. in *Advances in Neural Information Processing Systems 11*, Kearns, M. S.; Solla, S. A.; Cohn, D. A. Eds., pp. 1052-1058, MIT Press, Cambridge, **1999**.

[80] Thiesson, B.; Meek, C.; Heckerman, D. "Accelerating EM for large databases", *Microsoft Research Technical Report MST-TR-99-31*, **1999**.

[81] Banfield, J. D.; Raftery, A. E. *Biometrics*, **1993**, 49, 803.

[82] Dasgupta, A.; Raftery, A.E. *Journal of the American Statistical Association*, **1998**, 93, 294.

[83] Murtagh, F.; Raftery, A. E. *Pattern Recognition*, **1984**, 17, 479.

[84] Banerjee, S.; Rosenfeld, A. *Pattern Recognition*, **1993**, 26, 963.

[85] Dempster, A. P.; Laird, N. M.; Rubin, D. B. *Journal of the Royal Statistical Society, Series, B* **1977**, 39, 1.

[86] Fraley, C. *SIAM Journal of Scientific Computing*, **1999**, 20, 270.

[87] Kass, R.E.; Raftery, A. E. *Journal of the American Statistical Association*, **1995**, 90, 773.

[88] Schwarz, G. *The Annals of Statistics*, **1978**, 6, 461.

[89] Fraley, C.; Raftery, A. E. *The Computer Journal*, **1998**, 41, 578.

[90] Mukherjee, S.; Feigelson, E. D.; Babu, G. J.; Murtagh, F.; Fraley, C.; Raftery, A. *The Astrophysical Journal*, **1998**, 508, 314.

[91] Celeux, G.; Govaert, G.; *Pattern Recognition*, **1995**, 28, 781.

[92] Sibson, R. *The Computer Journal*, **1973**, 16, 30.

[93] SDSS, Sloan Digital Sky Survey, http://www.sdss.org/

[94] Starck, J. L.; Murtagh, F.; Bijaoui, A. *Image and Data Analysis: The Multiscale Approach*, Cambridge University Press, New York, **1998**.

[95] Murtagh, F.; Starck, J. L.; *Pattern Recognition*, **1998**, 31, 847.
[96] Kolaczyk, E. D.; *Astrophysical Journal*, **1997**, 483, 340.
[97] Jammal, G.; Bijaoui, A. "Multiscale image restoration for photon imaging systems", *SPIE Conference on Signal and Image Processing: Wavelet Applications in Signal and Image Processing VII*, July **1999**.
[98] Zheng, G.; Starck, J. L.; Campbell , J. G.; Murtagh, F.; *Journal of Computational Intelligence in Finance*, 7, **1999**.
[99] Byers, S. D.; Raftery, A. E.; *Journal of the American Statistical Association*, **1998**, 93, 577.
[100] Allard, D.; Fraley, C.; *Journal of the American Statistical Association*, **1997**, 92, 1485.
[101] Ebeling, H.; Wiedenmann, G. *Physical Review E*, **1993**, 47, 704.
[102] Dobrzycki, A.; Ebeling, H.; Glotfelty, K.; Freeman, P.; Damiani, F; Elvis, M; Calderwood, T. *Chandra Detect 1.0 User Guide*, Chandra X-Ray Center, Smithsonian Astrophysical Observatory, Version 0.9, **1999**.
[103] Doyle, L. B. *Journal of the ACM*, **1961**, 8, 553.
[104] Murtagh, F. ; Hernández-Pajares, M., *Journal of Classification*, **1995**, 12, 165.
[105] Poinçot, Ph.; Lesteven, S.; Murtagh, F., *Astronomy and Astrophysics Supplement*, **1998**, 130, 183.
[106] Poinçot, Ph.,;Lesteven, S.; Murtagh, F. *Journal of the American Society for Information Science*, **2000**, 51,1081.
[107] Guillaume, D.; Murtagh, F. *Computer Physics Communications*, **2000**, 127, 215..
[108] Cartia, Inc., *Mapping the Information Landscape, client-server software system*, http://www.cartia.com/, **1999**.
[109] Church, K. W.; Helfman, J. I. *Journal of Computational and Graphical Statistics*, **1993**, 2, 153.
[110] Berry, M.W.; Hendrickson, B.; Raghavan, P. in Renegar, J., Shub, M. and Smale, S., Eds., *Lectures in Applied Mathematics (LAM) Vol. 32: The Mathematics of Numerical Analysis*, American Mathematical Society, **1996**, 99.
[111] Berry, M. W.; Drmač, Z.; Jessup, E. R. *SIAM Review*, **1999**, 41, 335.
[112] Murtagh, F.; Starck, J. L.; Berry M. *The Computer Journal*, **1999**, submitted.
[113] Murtagh, F. *The Computer Journal*, **1998**, 41, 517.

Beilstein-Institut Chemical Data Analysis in the Large, May 22nd – 26th 2000, Bozen, Italy

MODEL-BASED DATA COMPRESSION: FROM DATA COMPRESSION TO INFORMATION CONDENSATION

HOLGER WALLMEIER

Aventis Research & Technologies GmbH & Co KG, Core Technology Area Biomathematics, Industrial Park Hoechst, G 515 A, D-65926 Frankfurt am Main, Germany
E-mail: wallmeier@CRT.hoechst.com

Received: 23rd January 2001 / Published 11th May 2001

ABSTRACT

Support of industrial research and development activities by computing and information technologies today is coupled to huge amounts of data. Therefore, data management is a very crucial aspect of successful application of information technologies. Various strategies are used to handle the situation, each of which has its merits depending on the type of data, the context, and the usage.
Apart from the very straightforward approach to distribute data on appropriate storage media of sufficient volume, there are three different 'philosophies' of data compression.

1. Non-lossy data compression
2. Lossy data compression
3. Model-based data compression

Types 1 and 2 are probably the most widely used because they do not necessarily introduce a bias into the compressed data. There are a number of methods known today that are fully reversible, or at least reversible to a large extent.
This is different for model-based data compression. The idea is useful for data being produced by dynamic, deterministic systems. Important is the existence of a model with well-defined data scheme and data structure. These model features can be used to condense the corresponding original data. Two examples from industrial research are presented.
First example is the representation of computer simulations of molecular ensembles by correlation functions. The second example is the representation of microbiological studies on pathogenicity by kinetic constants. In both cases, the underlying model together with methods to generate compressed data representations allows efficient interpretation of simulations or experiments, respectively.
High levels of data condensation provide a variety of opportunities to link results from research and development to auxiliary information from many different sources. Thus, powerful infrastructures for decision support can be created.

INTRODUCTION

Production of Data

The typical scenario of data generation starts from a device or automaton producing data (production), which will be recorded using some representation characteristic for the data and, of course, characteristic for the production itself. Based on this representation, data will be processed to extract the related information. In addition, the data may be transferred into a repository for later use.

Whenever the original production is deterministic, a faithful model of the original data production can be found, at least in principle. In such cases there is a unique data scheme and, furthermore, a well-defined analytical data representation. Usually, such models are given by a system of differential equations, the solutions of which define the data scheme. The way in which these solutions are determined also defines options of data representation. Extraction of information is then straightforward (Figure 1).

The advantage of the correspondence between original production and model production is that data scheme and data representation of the model

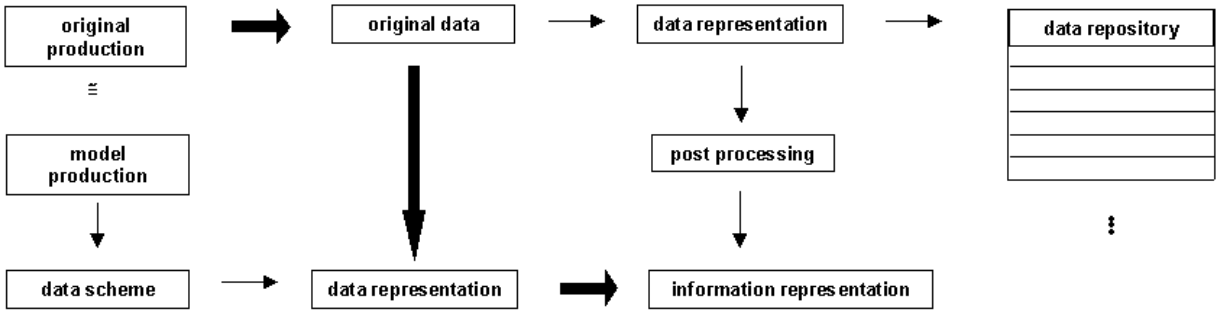


Figure 1: Data production and representation

can be used to generate a condensed representation of the original data. The information behind model data can usually be represented by few parameters.

The way in which this works will be shown by two examples. The first example is the computer simulation of molecular structures to analyze the stability of biomolecular complexes. In the second example, it will be shown how microbiological experiments with pathogenic bacteria can be analyzed in a very efficient way.

Representation and Condensation of Data

Extracting and condensing information from data means creating a specific representation of the information. Basically, there are two different approaches to representing information. On the one hand, information can be mapped using predefined descriptor sets, thus creating specific profiles. On the other hand, information can be mapped in terms of relationships of the given object to known objects, which results, at least, in a delimiting view of the information. Genealogic aspects can be taken into account quite easily on a class and instance basis.

Both approaches offer several ways to condense information. Descriptor sets and profiles, for example can be handled using statistical methods such as clustering and classification, which also suggest strategies of visualization familiar from statistics and data mining.

Representing information by specifying relationships is first of all a simple and direct way of classifying objects based on similarity. In addition, this concept directly leads into the world of semantic networks.

An Example from Molecular Modeling

Molecular modeling can be very useful to assess questions regarding, very generally speaking, stability and affinity of molecular systems. A very powerful, even though 'expensive' tool is the simulation of the dynamics of molecular systems. The underlying paradigm is based on perturbation theory. Simulations can be considered as computer experiments that allow the study of the response of a given system (molecular model) to some defined perturbation. The perturbation applied most frequently is just the kinetic energy of the N particles associated with a given temperature T according to [1]

$$E_{kin} = \frac{1}{2}\sum_{i}^{N} m_i \cdot \vec{v}_i^2 = \frac{3}{2}NkT \qquad (1)$$

Such simulations show the time evolution of the given system under the thermodynamic conditions specified and allow us to judge the stability of the given structural alignment, constitution, or conformation relative to some reference state. For

this reason, simulation of molecular dynamics is a quite popular way of performing conformation-searches, especially for large molecular systems. By extending the analysis to the various aspects of entropy, affinity can also be estimated, at least on a molecular level.

A Model for Dynamical Affinity of Molecular Systems

In practice, molecular dynamics simulations are performed by discretized integration of the respective equations of motion. [2] Since the characteristic frequencies of all relevant degrees of freedom must be resolved by the integration step-size, simulations of molecular dynamics are usually very lengthy and produce huge data sets (trajectories) if applied to large systems.

There is, however, a way to avoid very long simulations. The idea is based on the concept of collective modes of oscillation, which exist in stable molecular alignments. Indeed, the existence of such collective modes can be taken as a criterion for stability, because they make the difference between an unstable scattering state and a stable bound state of a molecular aggregate. According to quantum mechanics, their respective Eigenfrequency can identify such modes. Using a so-called Drude model, [3] which was originally

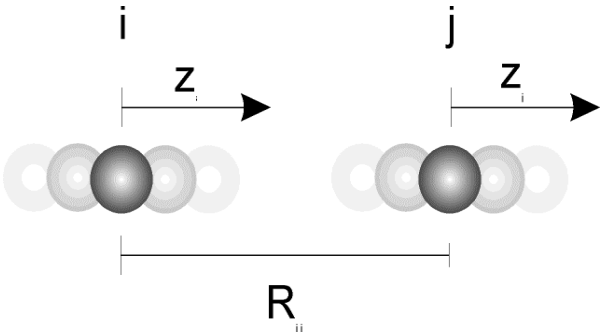


Figure 2: Coupled oscillators

developed for the electronic dispersion interaction of atoms and molecules by London, [4] this can be shown quite easily. Interacting molecules are represented by pairs of coupled harmonic oscillators (Figure 2). For simplicity, we take a pair of one-dimensional, coupled, identical harmonic oscillators positioned on the z-axis. The corresponding Hamiltonian is given by

$$H = T + V = \tfrac{1}{2} m\left[\dot{z}_i^2 + \dot{z}_j^2\right] + \tfrac{1}{4} K\left[z_i^2 + z_j^2 + 2a \cdot z_i \cdot z_j\right] \tag{2}$$

where m is the mass of each oscillator, K the force constant and a the coupling constant, which is a function of the distance between the equilibrium positions of the oscillators. After separation of variables one has

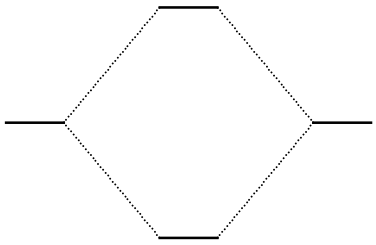
$$\begin{aligned} H &= \tfrac{1}{4}\left(\dot{z}_i + \dot{z}_j\right)^2 + \tfrac{1}{4} K(1+a)\left(z_i + z_j\right)^2 + \tfrac{1}{4} m\left(\dot{z}_i + \dot{z}_j\right)^2 \\ &+ \tfrac{1}{4} K(1-a)\left(z_i . z_j\right)^2 \\ &= H\left(z_i + z_j\right) + H\left(z_i - z_j\right) \end{aligned} \tag{3}$$

The first term represents the coherent motion of the center of gravity of the pair of oscillators and the second term the relative 'breathing' motion. Since both oscillators have a ground state frequency ω_0, coupling results in a symmetric split of energy levels as shown in the following scheme (for $a > 0$)

$$\omega_+ = \omega_0 \cdot \sqrt{1+a}$$

$$\omega_0 = \sqrt{K/m}$$

$$\omega_- = \omega_0 \cdot \sqrt{1-a}$$



(for $a < 0$ in reversed order)

The energy of an oscillator is $\varepsilon = \frac{1}{2}\hbar \cdot \omega$, so that the splitting is given by

$$\varepsilon - \varepsilon_0 = \tfrac{1}{2}\hbar(\omega_+ + \omega_- - 2\omega_0) \qquad (4)$$
$$= \tfrac{1}{2}\hbar\omega_0\left[\sqrt{1-a} + \sqrt{1-a}\right] - 2$$

For small $|a|$ one has

$$\varepsilon - \varepsilon_0 \approx -\hbar\omega_0 \cdot a^2 \qquad (5)$$

which is a typical second order, resonance-like effect.

Coming back to classical mechanics, one can calculate the sum over states ($|a| << \omega_0$) of the system

$$Q = \frac{kT}{\hbar\omega_+} \cdot \frac{kT}{\hbar\omega_-} = \frac{\left(\frac{kT}{\hbar\omega_0}\right)^2}{\sqrt{1-a^2}} \qquad (6)$$

The Helmholtz free energy is

$$A = -kT \cdot \ln Q = A_0 + \frac{kT}{2}\ln\left(1-a^2\right) \qquad (7)$$

and since

$$S = -\left(\frac{dA}{dT}\right)_v = k \cdot \ln Q + kT \cdot \left(\frac{d\ln Q}{dT}\right)_v,$$

and

$$S = S_0 - \tfrac{1}{2}k \cdot \ln\left(1-a^2\right),$$

it is clear that

$$A - A_0 = -T \cdot (S - S_0) = T \cdot \tfrac{k}{2}\ln\left(1-a^2\right) \approx \frac{kT}{2}a^2 \qquad (8)$$

Therefore, in terms of thermodynamics, coupling of oscillators adds a contribution to the energy of the overall system, which is mainly an entropy effect. It should be noted that the difference of the energy levels is independent of the sign of the coupling constant, since it is proportional to a^2. The energetic order of ω_+ and ω_-, however, is a function of the sign of the coupling constant. By analogy to the analysis of the (electronic) dispersion interaction by London, [4] this contribution to the entropy of molecular complexes can be called mechanical dispersion or, because of its stabilizing effect, dynamical affinity.

Tracing Dynamical Affinity in Molecular Dynamics Simulations

In a molecular dynamics simulation dynamical affinity can be traced, mapping coherent and breathing motion by correlation functions.

$$G_{ij} = \frac{1}{T}\int_{t_0}^{T} g_i(t+\delta)\cdot g_j(t)\ ;$$

$$g = \begin{cases} \text{: position correlation function} \\ \text{: velocity correlation function} \end{cases}$$

i, j : centers of correlation

δ *: correlation time*
$T\text{-}t_0$ *: time of measurement (simulation time)*
$i = j$ *: autocorrelation*
$i \neq j$ *: cross-correlation*
$G_{ij}(0) = 1$ *: normalization*

For harmonic oscillators, one can define the autocorrelation functions for coherent and breathing motion

$$G_{ij}^{+}(\delta) = \frac{1}{T}\int_{t_0}^{T} \tfrac{1}{2}\left[g_i(t+\delta) + g_j(t+\delta)\right]\cdot \tfrac{1}{2}\left[g_i(t) + g_j(t)\right]\cdot dt \qquad (9)$$

$$G_{ij}^{-}(\delta) = \frac{1}{T}\int_{t_0}^{T} \tfrac{1}{2}\left[g_i(t+\delta) - g_j(t+\delta)\right]\cdot \tfrac{1}{2}\left[g_i(t) - g_j(t)\right]\cdot dt \qquad (10)$$

Now, it can be shown that the second derivative of these correlation functions is $-\omega^2$ for zero correlation time ($\delta = 0$). This means that the whole simulation can be condensed to just two independent numbers, ω_+, ω_-, and perhaps $\Delta\omega = \omega_+ - \omega_-$.

G^+ and G^- are determined by selecting two centers (atoms or groups of atoms) i and j. The only condition to meet, is that i and j should be

Molecule	ΔG kcal mol^{-1}	ΔH kcal mol^{-1}	TΔS (297 K) kcal mol^{-1}	$K_{binding}$ (M^{-1})	ΔG_{Biotin} / ΔG_{HABA}	$T\Delta S_{Biotin}$ / $T\Delta S_{HABA}$
Biotin	-18,3	-32,0	-13,7	$2{,}5 \times 10^{13}$	3,5	-2,0
HABA	-5,27	1,70	6,97	10^4	1	1

Table 1: Thermochemical data of Streptavidin complexes with Biotin and HABA.

influenced in their dynamics by both, the coherent and the breathing mode.

Since ω_+ and ω_- are determined from the trajectories of the full simulation ensemble, they are frequencies from the phonon spectrum of the whole system and not just frequencies of local molecular vibrations. In fact, splitting into ω_+ and ω_- is a sensitive indication of the existence of a common, non-local mode of vibration for both oscillators. This of course shows that the interaction between the molecules has lead to a stable bound state and not an unstable scattering state.

Streptavidin and Biotin

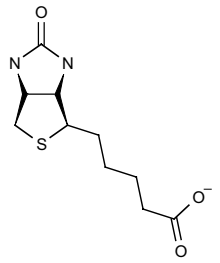
The example given below is a complex of two biomolecules, the protein Streptavidin and the vitamin Biotin. They form a specific complex with the largest binding constant known between biomolecules in nature. Therefore, this system is frequently used for immobilization of biomolecules. Surprisingly, experimental studies with molecules slightly different from Biotin show significant loss of stability and document the high specificity of the Biotin/Streptavidin complex.

For example, 2-(4'-hydroxyphenylazo)-benzoic acid (HABA) also binds to streptavidin, but with a binding constant which is 9 (!) orders of magnitude lower.

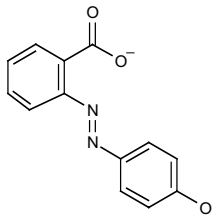
The thermochemical data measured for these complexes are given in Table 1. [5]

Apart from the remarkable values of the binding constants, it should be noted that the sign of the entropy contribution to the free binding energy changes going from Biotin to HABA. This is an indication of a change in the role of entropy.

Biotin



2-(4'-hydroxyphenylazo)-benzoic acid



$$K_{binding}^{Biotin} = \frac{[Streptavidin{:}Biotin]}{[Streptavidin][Biotin]} \cong 10^{13}\, M^{-1}$$

$$K_{binding}^{HABA} = \frac{[Streptavidin{:}HABA]}{[Streptavidin][HABA]} \cong 10^{4}\, M^{-1}$$

From the theoretical point of view, it is of course a challenge to model such a system. Fortunately, crystal structures of both complexes have been published. [6] Molecular dynamics simulations starting from these crystal structures have been run using the AMBER 3.0 [7] force field in NVT ensembles with water and counterions at 300 K. The ensembles have been thermalized during 30 psec simulations. Subsequently, another 15 psec were used to sample the trajectories from which oscillator correlation functions have been estimated. Table 2 summarizes the results of the simulations. For Biotin, four different orientations of the ligand in the binding pocket of Streptavidin have been simulated, for two HABA (Table 3). Columns 2 and 3 show the frequencies of the coupled oscillator motions derived from the autocorrelation functions G^+ and G^-. For the crystal structure orientation of Biotin (1) the coherent motion has the lower frequency and the breathing motion is significantly faster. In the first row of Table 2 the Biotin-results of a simulation without water and counterions are

System / binding mode	ω_+ (GHz)	ω_- (GHz)	$\Delta\omega$ (GHz)	Type of coupling	Splitting kcal.mol^{-1}	TΔS (297 K) [6] kcal.mol^{-1}
1STP [8]/ 1[9]	5.4	14.6	-9.1	a<0	-0.87	
1STP / 1	2.9	12.4	-9.6	a<0	-0.91	-13.70
1STP / 2	6.1	0.76	5.4	a>0	0.51	
1STP / 3	13.5	6.9	6.6	a>0	0.63	
1STP / 4	10.7	2.2	8.4	a>0	0.80	
1HBA [8] / 1	8.7	4.7	4.0	a>0	0.38	6.97
1HBA / 2	8.6	13.8	-5.3	a<0	-0.50	

Table 2: Results of molecular dynamics simulations of Streptavidin complexes with Biotin and HABA. Starting from the crystal structures published, different orientations of the ligands have been studied. See text for further details.

given. The values do not differ very much from the results for the solvated system, which indicates the robustness of the method.

The interesting result is that the entropy contribution from oscillator coupling found in the molecular dynamics simulations shows the same relationship between Biotin and HABA as does the experimentally determined quantity $T \cdot \Delta S$. The agreement is 13% with respect to the experimental value, which is adequate for the force field chosen, the size of the simulation ensembles, and the simulation time.

$$\eta = \frac{T \cdot \Delta S_{Biotin}}{T \cdot \Delta S_{HABA}}$$

	Experiment	MD Simulation
	297 K	300 K
η	-1.97	-2.27

This underlines the role of oscillator coupling as indicator for stability of a given molecular alignment. At the same time it demonstrates the potential of data reduction that is given by this approach.

In terms of model-based data compression, we have the following situation. The original data production is the molecular dynamics algorithm in combination with the force field model of the system. The trajectories are the original data. Now,

Ligand orientation	Description	System
1	crystal structure	Biotin, HABA
2	upside down	Biotin, HABA
3	reversed	Biotin
4	upside down and reversed	Biotin

Table 3: Orientations of the ligands Biotin and HABA bound to Streptavidin

the model production is given by the coupled oscillators, the corresponding data scheme by the oscillator correlation functions, and the data representation by the oscillator frequencies. The representation of the information, *i.e.* the descriptor of stability, is given by the level splitting, calculated from the frequencies.

AN EXAMPLE FROM BIOMETRY

Quite a different approach to model based data compression is possible in the area of kinetic studies of bacterial pathogenicity. Such studies are very important in infectious disease research. In a very general view, the key issue is the interaction between pathogens and the hosts they infect.

Besides the medicinal aspects of infection, pathogen-host interactions are the primary focus of target and lead compound search in the pharmaceutical industry. It is a complex phenomenon with several degrees of freedom.

Dynamics of Infectious Disease Progression

Progression of an infectious disease is, in a generalized sense, always the result of several types of growth processes, which are characteristic for different phases of disease progression. [10] If one wants to identify targets for anti-infective drugs, the early phases of disease progression are of special interest.

The first phase is the invasion of the pathogen. This is some kind of transport phenomenon, which often is coupled to specific surface interactions and recognition steps.

What follows is a phase of establishment that usually results in a growth of the pathogen population. In this phase chemical communication between pathogen and host may occur, which can facilitate the pathogen's establishment significantly. The chemical 'messages' pathogens send to the host are called virulence factors. Typically, they serve to subvert normal functions of the host cells. Sometimes they have an immuno-suppressive effect. [11]

Next is the formation or enrichment of so-called pathogenicity factors. Very often they are toxins secreted by the pathogen. But also bacterial enzymes, which, for example cause necrotic degradation of host tissue belong to this class of factors.

Last but not least, the development of disease symptoms is related to the amount of pathogenicity factors formed. In all theses phases, however, there is some kind of host response to defend against the pathogen. For more complex host organisms it is an immune response.

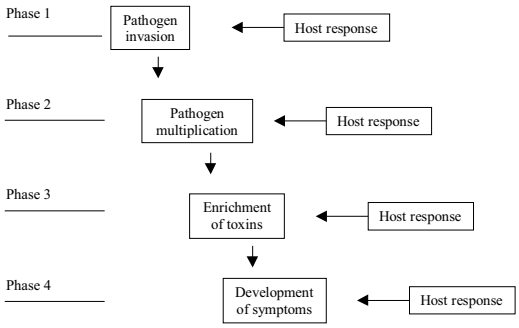


Figure 3: Phases of infectious disease progression. See text for details.

The scenario described above can be summarized in terms of the categories pathogenicity, virulence, and susceptibilty. Even though in literature pathogenicity and virulence are often used synonymously, a distinction based on genomical and disease progression considerations is possible. Pathogenicity is first of all a property of a pathogen that manifests in the formation of pathogenic factors like, for example toxins. [12] This, of course, depends on genotypic, as well as phenotypic conditioning of the pathogen. To be specific, what matters is type and amount of pathogenicity factors produced by the pathogen inside, or in contact with the host. The amount of factors formed, however, also depends on the size

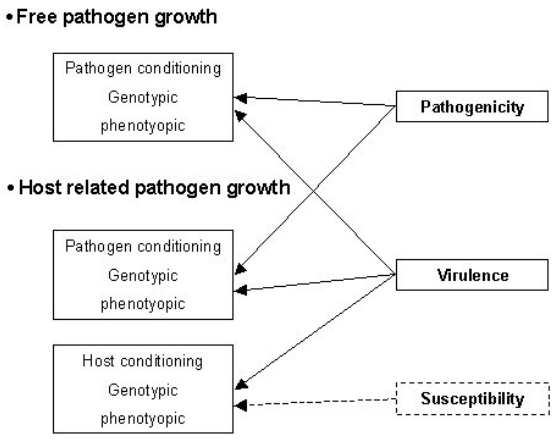


Figure 4: Pathogenicity, virulence, susceptibility, genotypic, and phenotypic conditioning.

of the pathogen population inside the host, which, in turn, depends on genotypic and phenotypic conditioning of the pathogen.

Due to host response, however, pathogen multiplication also depends on genotypic and phenotypic conditioning of the host. In principle, there are two degrees of freedom for the pathogen. These are, on the one hand its ability to produce pathogenicity factors, and on the other hand the size of population of pathogenicity factor producing pathogens inside the host.

Since virulence factors are often host specific, many authors refer the notion virulence to the combined effect of pathogenicity factor formation and population growth.

The third degree of freedom (see Figure 5) is the host's susceptibility to infection by the pathogen. Here, genotypic and phenotypic conditioning of the host are the important features.

Any research in the field of infectious diseases aimed at understanding the large variety of strategies pathogens have developed during evolution must analyze the kinetics related to the different phases. First of all, descriptors have to be identified that allow us to follow the individual processes by experimental measurements (see Table 4).

Phase	Type of process	Descriptors
Invasion	- transport phenomenon, first/zeroth order kinetics; - target recognition, signal transduction	invasive pathogen count, [13] optical densities of culture media specific interactions
Pathogen multiplication	- free pathogen population growth - invaded pathogen population growth (dependent on host response)	pathogen count, [13] optical densities pathogen count, [13] disease marker concentration, antibody titer
Toxin enrichment	- secretion of toxins and other pathogenicity factors - pathogen population growth	toxin concentration, antibody titer, disease marker concentration pathogen count [13]
Development of symptoms	- host population decrease	disease marker concentration, antibody titer

Table 4: Processes in disease progression

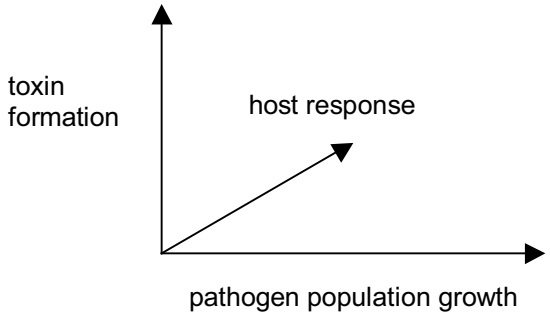


Figure 5: Genetic degrees of freedom in infectious diseases

A key problem in handling living organisms is reproducibility. Usually, this is taken care of by running replicate experiments and forming averages. In addition, time-resolved measurements are necessary to analyze the associated kinetics. To do so, the following model assumptions are useful.

A Model for Infectious Disease Dynamics

The normal way to measure pathogenicity starts from a set of N_0 host organisms, which are infected. In the course of the experiment, decrease of the host population is measured. Typically, one obtains a sigmoid curve (Figure 6), which can be represented by the solutions of the following differential equation (DE). It is called the logistic, autocatalytic, or autokatakinetic differential equation [14]

$$\frac{dN(t)}{dt} = \left[k - g \cdot N(t)\right] \cdot N(t) \qquad (11)$$

describing growth processes with feedback. It is the equation of an exponential growth, which is modified by the second term in the square brackets. This second term depends on the population N at time t and constitutes the feedback. It can be agonistic ($g<0$), as well as antagonistic ($g>0$). The

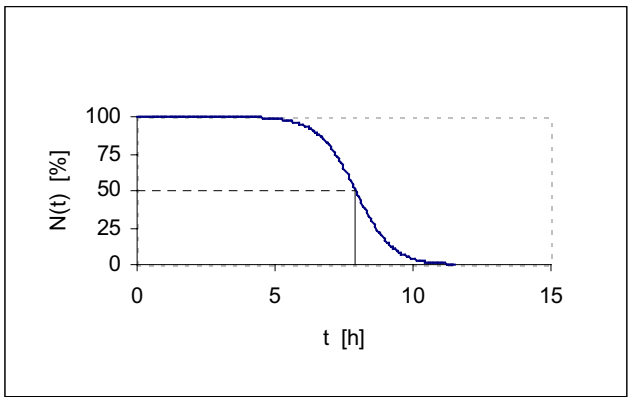


Figure 6: Decrease of a host population after infection. The time of the population's half-life is indicated.

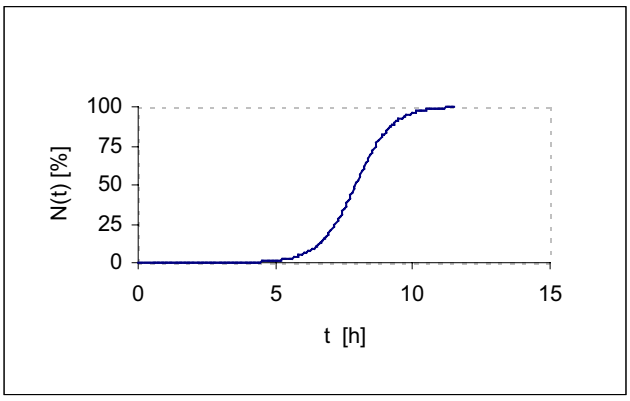


Figure 7: Increase of a pathogen population after infection.

general form of the solution is

$$N(t) = \frac{k \cdot N_0 \cdot e^{k \cdot t}}{k + g \cdot N_0 \cdot \left(e^{k \cdot t} - 1\right)} \qquad (12)$$

With the integration constant N_0, the initial size of the population, plus the rate constant k, and the feedback constant g there are three independent parameters. The combination of $N_0>0$ and a negative value of k describes the decrease of a population (Figure 6).

In contrast, vanishing N_0 together with a positive value of k describes a population that grows into a saturation state. With such parameters a growing pathogen population can be described (Figure 7).

Applying this equation to infection experiments, one has, first of all, equation (11) for the decrease of the host population. According to the considerations outlined above, it is easy to imagine that the kinetic constant k in fact depends on the growing pathogen population $P(t)$ and is thus a pseudo-constant. Therefore,

$$k = k\left[P(t)\right] \qquad (13)$$

The growth of the pathogen population may either be unrestricted (free exponential growth)

$$\frac{dP(t)}{dT} = \kappa \cdot P(t), \text{where } P(t) = P_0 \cdot e^{\kappa \cdot t} \qquad (14)$$

or restricted,

$$\frac{dP(t)}{dT} = \left[\kappa - \lambda \cdot P(t)\right] \cdot P(t), \quad \text{where } P(t) = \frac{\kappa \cdot P_0 \cdot e^{\kappa \cdot t}}{\kappa + \lambda \cdot P_0 (e^{\kappa \cdot t} - 1)} \qquad (15)$$

reaching a saturation level due to host response. As

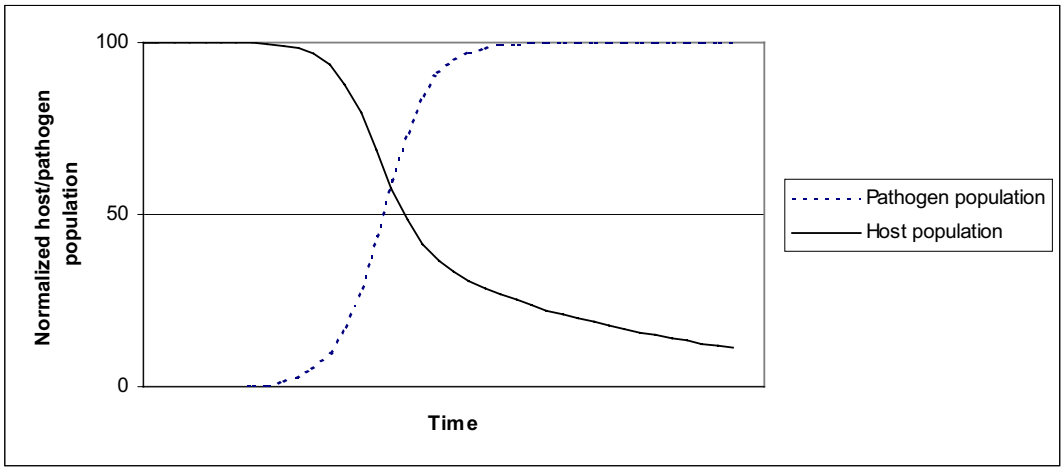


Figure 8: Decrease of host population coupled to growing pathogen population.

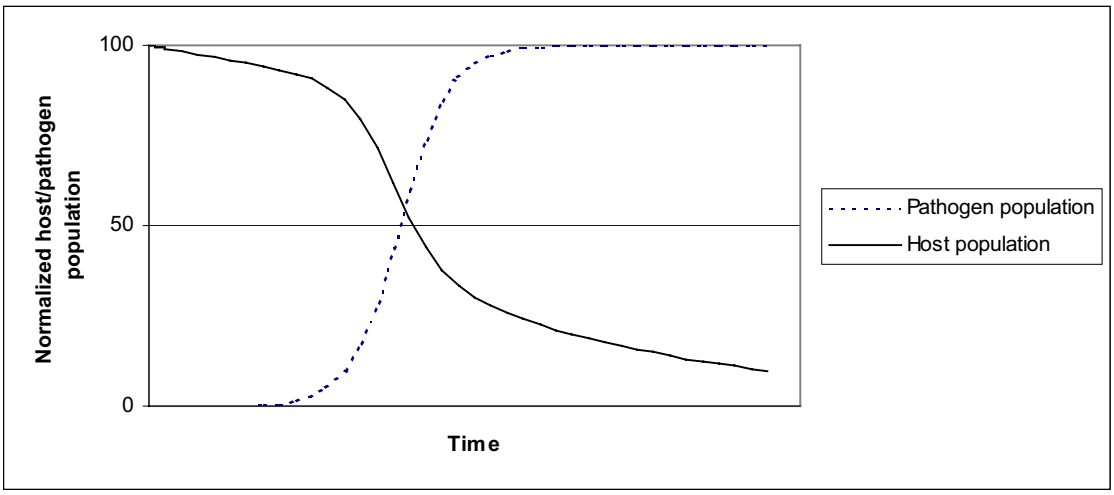


Figure 9: Decrease of host population coupled to growing pathogen population. A clear modulation of the host curve by the feed-back term can be seen

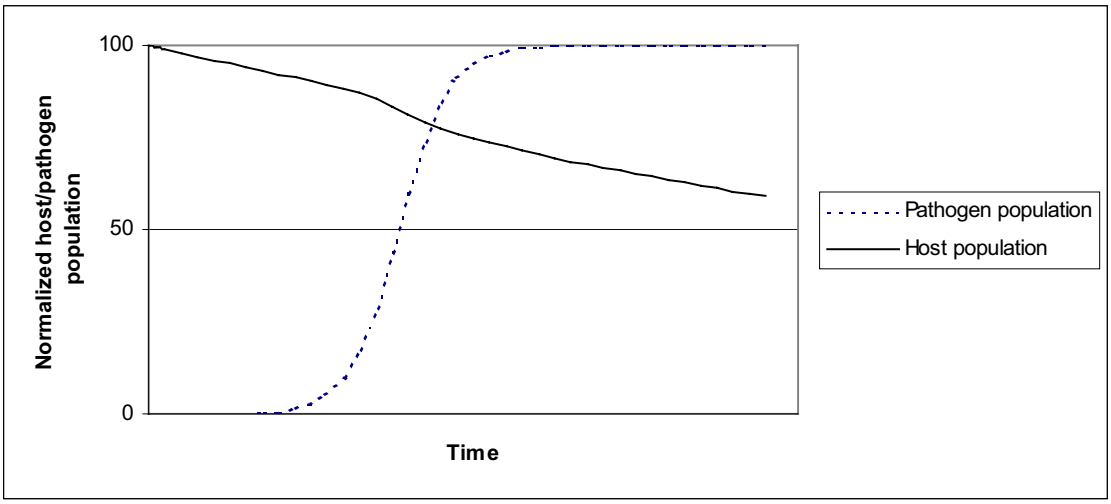


Figure 10: Decrease of host population coupled to growing pathogen population. Feed-back overrides the effect of the pathogen population

mentioned above, P_0 is small, and κ is positive.

The simplest form of combining the two processes is to set

$$k = -\eta P(t), \qquad (16)$$

which, for example results in the situation shown in Figure 8.

It is obvious that the effect of the growing (not constant) pathogen population can be seen as a deformation of the host population curve. The degree of deformation increases with η. There is, however, a further type of deformation of the host population curve. It comes from the feedback term and can be seen in Figures 9 and 10. This certainly reflects host conditioning.

Practical Applications

In experiments with time-resolved measurements, one usually has data reflecting the decrease of a host population that consists of test organisms such as, for example insects, mice, rats, or nematodes [15] (Figure 11).

Traditionally, the simplest way to measure pathogenicity is to count the host population after some predefined time $t_{scoring}$, which gives an *ad hoc* score as a percentage.

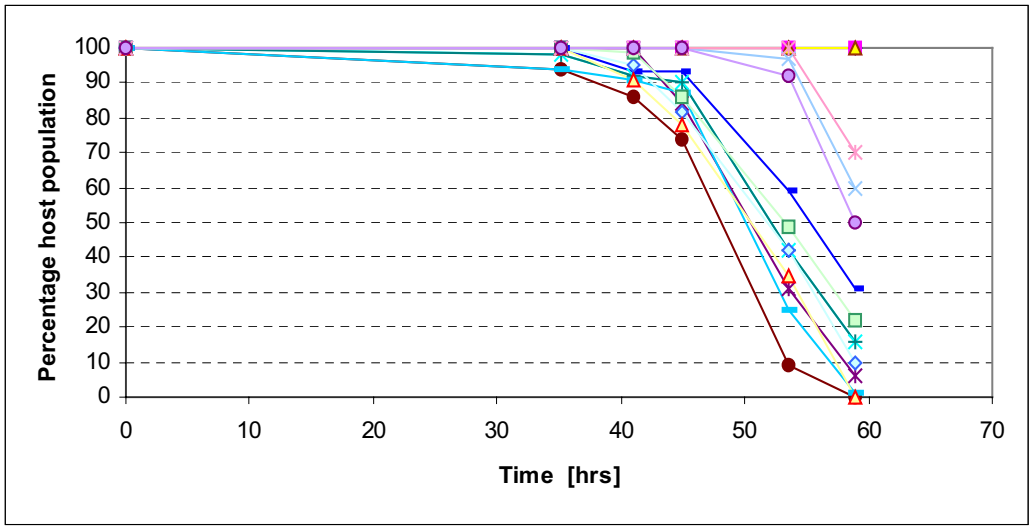


Figure 11: Time-resolved measurements of a C. elegans population infected by P. aeruginosa.

$$S_{adhoc} = \frac{N_0 - N_{t_{scoring}}}{N_0} \cdot 100\% \qquad (17)$$

Unfortunately, this way of measuring pathogenicity depends very much on the choice of $t_{scoring}$. Standardization is accomplished by normalization with the wild type of the pathogen:

$$S_{normalized} = \frac{\frac{N_0 - N_{t_{scoring}}}{N_0}}{\frac{N_0^{wildtype} - N_{t_{scoring}}^{wildtype}}{N_0^{wildtype}}} \cdot 100\% \qquad (18)$$

Very often, time series are run until the host population has reached half its original size and

$$t_{1/2} = t\left(N_0/2\right) \qquad (19)$$

is taken as the measure of pathogenicity. This condenses the whole series of measurements to one single value. For a given host organism those pathogens with a low $t_{½}$ are more pathogenic than those with a higher value.

Further Developments

In general, however, this is not sufficient to distinguish all possible effects that may modulate the interactions between a host and a pathogen. Coming back to the solutions of the logistic equation, steepness of the population curve at $t_{½}$ can tell a lot about pathogens, as well as hosts. [15] To improve the analysis, one has to fit solutions of the logistic equation (12) to the experimental data. This can be done, for example using the method by Marquardt and Levenberg. [16] The set of parameters k, g, and eventually κ, λ, or even η allow the identification of those bacterial mutants that show extraordinary behavior. This allows scanning the genome for so-called pathogenicity and virulence genes. Furthermore, different mechanisms of infection can be distinguished.

Together with the huge amount of genomic bacterial information available in the near future, such methods can be used to look for entirely new ways of fighting infectious diseases. One can, for example try to target genes or gene products involved in the very first step of an infection. This would not kill a pathogen, but disable its establishment and later on multiplication in the host. Such a ‘gentle’ way of infectious disease prevention is very likely not to trigger the development of resistances. Since the pathogens can survive ‘outside’ the host, there is only little

selective pressure.

Based on the situation described above, strategies for fighting infectious diseases must be defined. This also defines the type of targets to be searched and later on has an impact on the assays used for identification of active substances (lead compounds).

The normal strategy in target finding is to deactivate (knock out) genes systematically and to check by suitable assays with model organisms, to what extent pathogenicity, virulence, and, perhaps susceptibility are affected. Both steps are rather critical and need careful evaluation of the data generated and a very critical assessment of the results obtained.

SUMMARY

Whenever it is possible, model-based data compression serves two purposes. First of all it can be a great help to condense even huge data sets to very few numbers. Furthermore, the definition of the model necessary for compression is a very challenging step that often helps to gain deeper insights into the matter. It can reveal inconsistencies and facilitate the recognition of unknown phenomena. Together with the condensed data it offers possibilities to represent the information behind the data in a very efficient way.

However, any kind of modeling is an abstraction and idealization. A model always has to skip part of the reality. This of course limits the applicability of model-based data compression and defines the due diligence that must be applied using it.

REFERENCES AND NOTES

[1] McQuarrie, D. A.;*Statistical Mechanics*, Harper & Row, New York, **1976**.

[2] van Gunsteren, W. F.; Weiner, P. K. (Eds.), *Computer Simulation of Biomolecular Systems*, ESCOM, Leiden, **1989**.

[3] Drude, P. K. L.; *The Theory of Optics*. Longman, London, **1933**.

[4] London, F. *Z. Physik* **1930**, 63, 245; *Z. phys. Chem. B* **1930**, 11, 222; *Trans. Faraday Soc.* **1937**, 33, 537.

[5] Weber, P. C.; Ohlendorf, D. H.; Wendoloski, J. J.; Salemme, F. R. *Science* **1992**, 243, 85.

[6] Weber, P. C.; Wendoloski, J. J.; Pantoliano, M. W.; Salemme, F. R. *J. Am. Chem. Soc*. **1992**, 114, 3197.

[7] Weiner, S. J.; Kollman, P.A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Amer. Chem. Soc*. **1984**, 106, 765.

[8] Brookhaven PDB entry code

[9] 'in vacuo' (without water and counterions) simulation of the crystal structure of the Streptavidin/Biotin complex

[10] Finlay,B. B.; Falkow, S. *Microbiol. Rev*. **1989**, 53, 210.

[11] Mescas, J.; Strauss, E. J. *Emerg. Infect. Diseases* **1996**, 2, 271.

[12] Toxin is considered here to be not only a substance produced endogenically by the pathogen, but also toxic substances that are formed by host response to the pathogen.

[13] Measured as CFUs (colony forming units). See, *e.g.* Brock, T. D. *Biology of Microorganisms*, Prentice Hall, London **2000**.

[14] Dost, F. H. *Grundlagen der Pharmakokinetik*, Georg Thieme Verlag, Stuttgart **1968**; Nisbet, R. M.; Gurney, W. S. C. *Modelling fluctuating populations*, Wiley, New York, **1982**; Skehan, P. *Growth* **1986**, 50, 496; Renshaw, E. *Modelling biological populations in space and time*, Cambridge University Press, Cambridge, **1991**; M. Marusic, M.; Bajzer, Z.; Vuk-Pavlovic, S.; Freyer, J. P. *Bull. Mathem. Biology* **1994**, 56, 617.

[15] Mahajan-Miklos, S.; Tan, M. W.; Rahme, L. G.; Ausubel, F. M. *Cell* **1999**, 96(1), 47; Mahajan-Miklos, S.; Rahme, L. G.; Ausubel, F. M. *Mol. Microbiol*. **2000**, 37(5), 981.

[16] Dennis, J. E.; Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Non-linear Equations*, Prentice-Hall, **1983**. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C*, Cambridge University Press, Cambridge **1992**.

STRUCTURAL BROWSING INDICES AS HIGH-THROUGHPUT SAR ANALYSIS TOOLS

MARK JOHNSON[a] AND YONG-JIN XU[b]

[a] Computer-Assisted Drug Discovery, Pharmacia Inc., 301 Henrietta Street, Kalamazoo, MI 490006, USA.

[b] Discovery Medicinal Chemistry, Pharmacia Inc., 800 North Lindbergh Blvd., Creve Coeur, MO 63167, USA.
E-mail: mark.a1.johnson@am.pnu.com

Received: 22nd June 2000 / Published 11th May 2001

ABSTRACT

Structural browsing indices (SBIs) have been proposed as tools for organizing and exploring large sets of chemical structures in a manner complementary to that addressed by substructure and similarity-based methodologies. Molecular equivalence indices (MEQIs) comprise a special subclass of SBIs that play a central role in constructing a suite of SBIs appropriate to a variety of browsing, chemical-diversity, and SAR tasks. After presenting a general definition of a molecular equivalence index, three different ways of constructing SBIs based on MEQIs will be illustrated. The first index uniquely identifies the chemical graph of a compound and will be used to identify the sets of geometric and stereoisomers in a compound collection as well as to visually assess the overlap of two compound collections. The second index identifies a largest set of nonoverlapping functional groups of a compound and will be used to visually identify a functional-group-based receptor-relevant subspace associated with ACE inhibitors. The third index provides a hierarchical ordering of compounds whose use will be illustrated in the context of browsing structures and SAR relationships.

INTRODUCTION

The problem of organizing collections of molecular structures has been with us in one form or another since the dawn of modern chemistry. The development of substructure-searching algorithms was one of the initial pursuits in the creation of databases specifically structured for chemists and reflects the natural partial ordering of compounds with respect to the substructure-relationship. The last 15 years has seen the development of sophisticated algorithms for similarity searching, another way of exploring the compounds in a large collection based on the computation of a distance relationship between them. However, neither of these two methods provides a systematic way of assuring that all of the compounds in a collection have been examined.

Clustering and projection methods have long been available as statistical tools for organizing objects embedded in a high-dimensional space that does facilitate systematic browsing. Projection methods organize the objects into a low-dimensional space, usually the plane, so that distances in the points reflect distances between the points in the high-dimensional space. Clustering methods traditionally organize the objects along a line so that related clusters tend to occur together.

Long predating chemistry, humankind faced the problem of organizing large collections of hand-made objects as market places evolved. Modern department stores now display millions of items. Yet in one day you browse a large department store

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/johnson/johnson.pdf

for a sense of what it sells, and if you wish to buy a shirt, you are likely to find the shirts that interest you in reasonably close proximity of each other. This is an organizational feat that merits study by investigators in cheminformatics.

How do the department store managers do it? Basically they form a hierarchy of equivalence classes. Appliances, clothes, cosmetics…. Within appliances: stoves, refrigerators, washing machines…. Stoves might then be organized by size or manufacturer. Lastly the items are ordered along aisles in a manner consistent with this organizational hierarchy.

The organizational hierarchy can be distinguished from the clustering and projection methods we have just mentioned in that the equivalence classes are in some sense inherent in the nature of the object. We needn't see a stove in a cluster of other stoves, refrigerators, and washing machines to recognize that stove as an appliance and not a cosmetic.

The molecular equivalence indices presented here were developed with this department store analogy in mind, only, in this case, the equivalence classes are entities such as the chemical graph, the cyclic system, and chemical formula of a molecule, its side chains, ring systems and functional groups. Rouvray [1] reviews a number of the notions of structural equivalence that have played an important role in the development of chemistry. The formal perception of various integral components of a molecule has its origin in the dawning of cheminformatics, [2] as does the perception of an exhaustive set of a particular genre of components. [3] The idea of looking at a formalized notion of molecular equivalence and studying the resulting equivalence classes is more recent [4] as is the notion of hierarchically organizing structures by means of numbers. [5] The notion of systematically incorporating various notions of molecular equivalence into browsing indices whose values essentially serve as names for the resulting equivalence classes [6] forms the subject of this study.

After describing the set of structures that will serve to illustrate the concepts, a general definition of a molecular equivalence index (MEQI) will be given. A simple, yet fundamental, MEQI that assigns each chemical graph a unique code [7] will be presented and used to find the sets of geometric and stereoisomers in collection of compounds and to illustrate a simple mechanism for determining which structures occur in each of two collections. A more general MEQI identifies a largest set of nonoverlapping functional groups of a compound and will be used to visually identify a functional-group-based receptor-relevant subspace associated with ACE inhibitors. Finally, a MEQI specifically designed to hierarchically order compounds with respect to their cyclic systems and arrangement of their side chains will be illustrated in the context of browsing structures and SAR relationships.

AN ACE-INHIBITOR DATASET

In a recent paper, Pearlman and Smith [8] develop the concept of a receptor-relevant subspace using 78 angiotensin-converting enzyme (ACE) inhibitors. In Figure 3 of that study, these 78 compounds are positioned in a localized area of a three-dimensional BCUT space when viewed against a backdrop of a "5% diverse subset of the total MDDR [9] population." Bob Pearlman graciously sent us the structures of those 78 ACE inhibitors and Veer Shanmugasundaram kindly provided us with a similar diverse subset of 3932 compounds based on a comparable three-dimensional BCUT space from the MDDR collection at Pharmacia. Choosing a "comparable" subset of the MDDR compounds to serve as a backdrop was thought to increase our chances of finding a receptor-relevant subspace using MEQIs,

a concept that will be discussed in the section on the alpha-augmented functional group ensemble MEQI. No attempt has been made to verify the suitability of this expectation.

Defining a Molecular Equivalence Index

If a chemical descriptor is viewed broadly enough to include any function that maps the space of compounds to a linearly ordered set, a MEQI can be viewed as a special case of a chemical descriptor. However, in the case of a MEQI, this mapping can always be viewed as a composite mapping in that it first maps the space of compounds to a space of visually interpretable representations and then maps this intermediary space to a linearly ordered set.

This decomposition of a MEQI is illustrated in Figure 1. A few comments are needed to explain the figure. For computationally purposes, one must replace the compounds by some approximate mathematical representation. In Figure 1, we use a slight generalization of the chemical graph in which both the vertices and the edges are labeled. Mathematicians call this a colored or labeled graph. By allowing for loops and multiple edges, one obtains a labeled pseudograph. Thus, in our case, the equivalencing function always maps the space of labeled pseudographs onto itself. The particular equivalencing function in Figure 1 deletes all single-degree vertices labeled 'H' for hydrogen. In particular, it converts all chemical graphs to their hydrogen-reduced counterparts, but note that our definition of this equivalencing function is operationally defined for any labeled pseudograph.

The second mapping assigns each labeled pseudograph a unique code. This code depends only on the labeled pseudograph on which it is computed, and not on the compound mapped to that pseudograph by the equivalencing function. This code could be a number base 10, a number base 36, such as is used in car license plates, or a character string. However, the resulting values must be linearly ordered. In some cases, these assigned values depend on the sequence in which graphs are presented to the naming algorithm with the first graph labeled number 1, the second 2, *et cetera*. [6] *A priori* naming procedures [5,7] depend only on the labeled pseudograph and will consequently be independent of the time and place in which the naming is carried out. Obviously, the utility of a MEQI diminishes rapidly if this naming function is

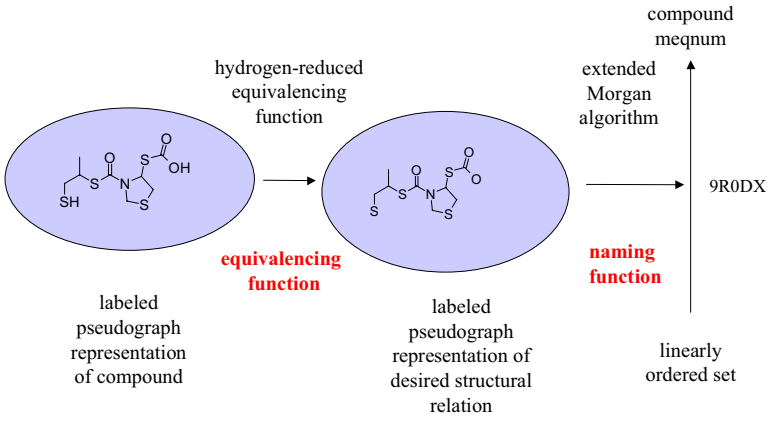


Figure 1. Two basic components of a molecular equivalence index mapping a compound to its compound meqnum.

not unique for all practical purposes, i.e. nonisomorphic labeled pseudographs are assigned distinct values. (It remains an open question if there exists a one-to-one naming function that lies outside the NP-completeness class. [10])

In this study we will be using an extension [11] of the Morgan algorithm [12] to compute an *a priori* naming function. We have yet to encounter a case of nonuniqueness. This algorithm assigns a number base 34. (I's and O's are not used because of their possible confusion with 0's and 1's.) We refer to this number as a molecular equivalence number or meqnum for short.

For every distinct equivalencing function, we obtain a distinct MEQI. When the equivalencing function maps a compound to its hydrogen-reduced graph as in Figure 1, we call the resulting assigned numbers "compound meqnums." In an analogous way, we obtain compound-skeleton meqnums, cyclic-system meqnums, cyclic-system skeleton meqnums, *et cetera*.

THE COMPOUND MEQNUM

Finding Geometric and Stereoisomers

The compound meqnum identifies a compound up to geometric and stereoisomerism. Even this simple meqnum has interesting uses. For example, the pharmacological activity of a compound is often stereospecific, whereas most chemical descriptors are not. This would seriously diminish the utility of most chemical descriptors in lead-optimization contexts if it were not for the fact that lead optimization efforts in drug discovery quickly focus on those compounds with the desired handedness at the critical stereocenters. However, there are often cases in which both stereoisomers are present and one must remove the compound with the undesired handedness before proceeding further. This is easily done by computing the compound meqnums for all of the compounds and then constructing the histogram given in Figure 2. We will assume that the compound with the desired handedness will be synthesized whenever the compound with the undesired handedness is synthesized. Consequently, the compound meqnum of any compound with the undesired handedness will occur twice since the corresponding stereoisomer will also be present and have the identical chemical graph.

Emerging graphical capabilities are enabling us to visualize relationships involving high-content

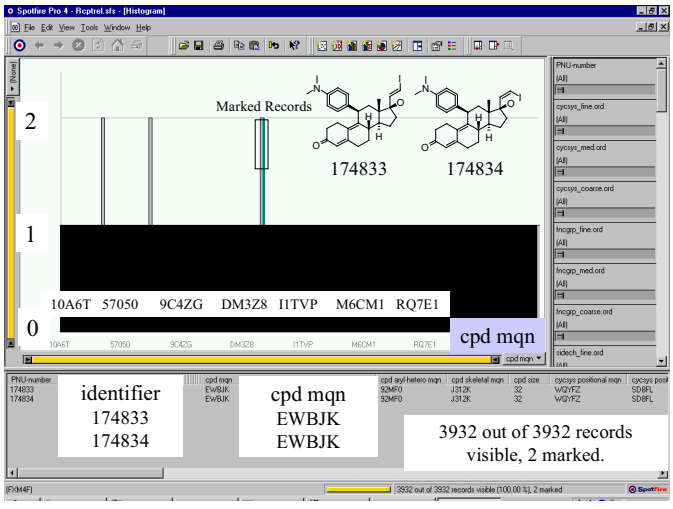


Figure 2. Histogram for finding geometric and stereo isomers with with the compound meqnums along the x-axis. The two geometric isomers associated with the marked bar of height two are displayed.

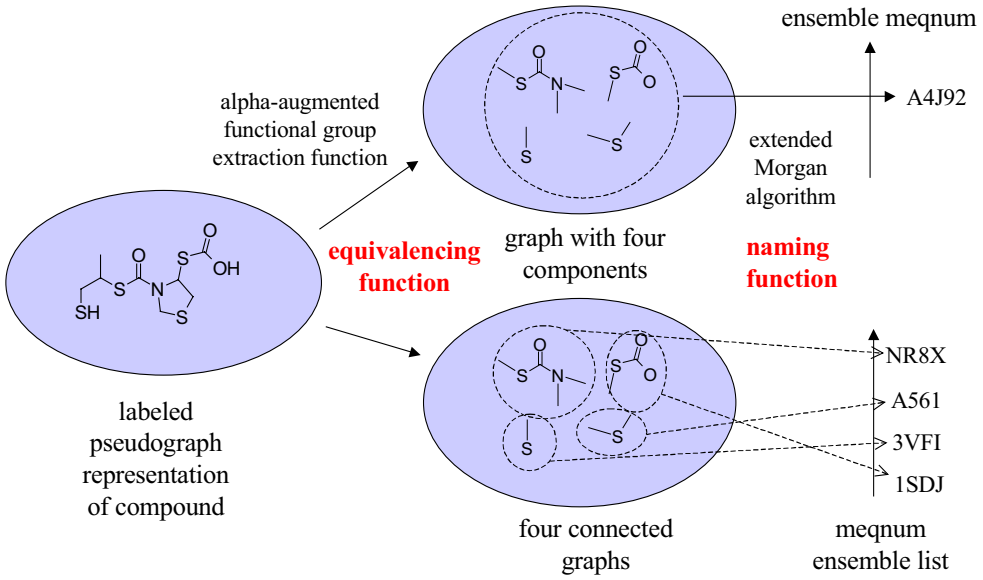


Figure 3. Construction of two alpha-augmented functional group MEQIs using a naming function that generates a single meqnum and a list of meqnums for multicomponent graphs, respectively.

variables such as MEQIs. Spotfire [13] allows the use of string-valued variables for the axes of a plot and provides many of the navigational aids required for efficient browsing. By simply selecting the compound meqnum variable for the x-axis in the histogram view in Spotfire, Figure 2 pops into view.

Out of a data set of roughly 4000 compounds, one quickly and visually isolates all the pairs of geometric and stereoisomers. These pairs correspond to the three thin bars of height 2 representing 6 compounds. The structures can be seen by moving the mouse diagonally across its top to form an enclosing rectangle which "marks" the compounds. One of the bars of height 2 in Figure 2 is marked. The details window gives the identifiers of the two tallied compounds as **174833** and **174834** and gives EWBJK for the common compound meqnum. The remaining 3926 compounds are represented by corresponding bars of height 1 compressed so tightly as to give the visual impression of a solid black horizontal bar of that height.

Comparing Two Compound Collections

A similar logic allows one to quickly find the intersection in two compound collections. Again, compounds that occur in both collections would be represented by bars of height 2 or greater. These can be marked appropriately and the other compounds deleted. The remaining bars can then be proportionally colored by source. Multicolored bars would reflect chemical graphs found in both collections. Monocolored bars would represent isomers and other compounds with the same chemical graph found in only one collection.

An Alpha-Augmented Functional Group MEQNUM Ensemble

The concept of a receptor-relevant subspace as developed by Pearlman and Smith [8] can be viewed generally as any formal specification of a class of compounds in which compounds with the desired receptor affinity are highly concentrated. In this section, we would like to illustrate another group of MEQIs by developing one that provides a

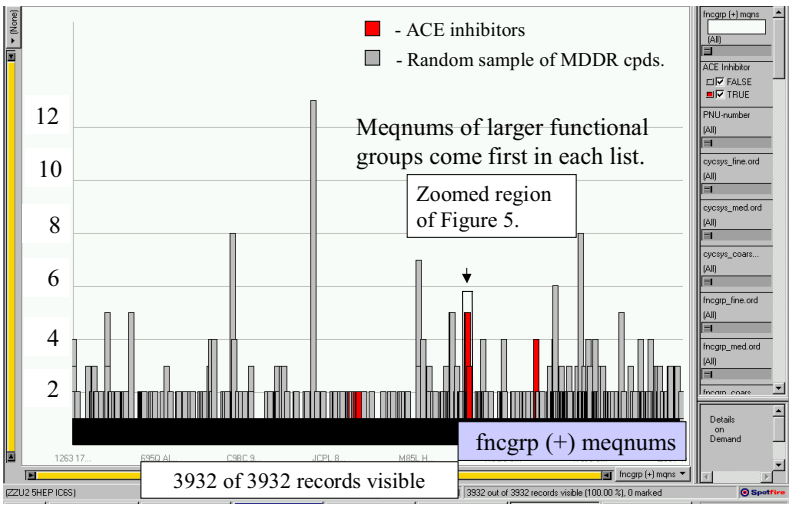


Figure 4. Histogram of alpha-augmented functional group list.

simple means of specifying a receptor-relevant subspace for the 78 ACE inhibitors in our data set.

Figure 3 shows two distinct MEQIs involving the same equivalencing function, but two different, yet related naming functions. To define the equivalencing function, divide the atoms of a chemical graph into separating atoms and non-separating atoms. Call a largest-connected subgraph consisting only of non-separating atoms a maximal group. By letting the separating vertices be any carbon atom that does not share a double bond with any oxygen, nitrogen, or sulfur or share a triple bond with nitrogen, we obtained the maximal functional groups. By augmenting these maximal functional groups with their adjacent alpha carbon atoms, we obtain the alpha-augmented functional groups (AFGs) that form the disconnected graph of four components depicted Figure 3.

We now have a choice of naming functions. We can use the one in Figure 1 which always assigns a single number to a graph whether connected or not. This gives the ensemble meqnum A4J92 in the

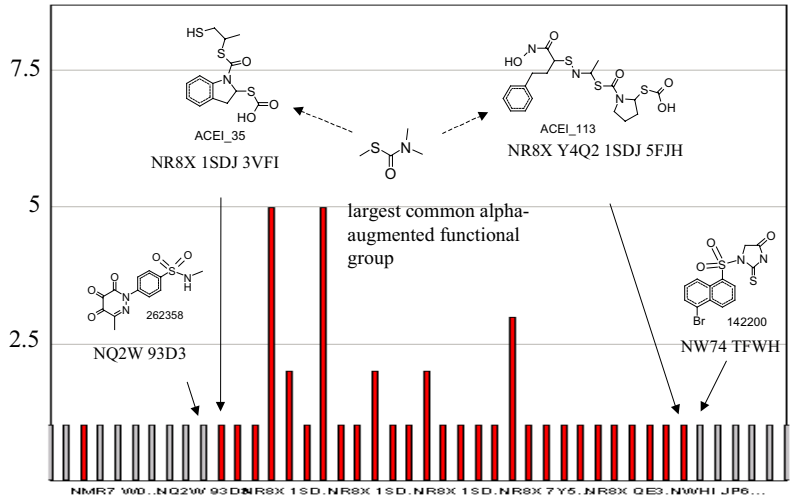


Figure 5. Zoomed region of histogram in Figure 4 of the alpha-augmented functional-group meqnum-ensemble list showing a grouping of ACE inhibitors with respect to their largest perceived functional group.

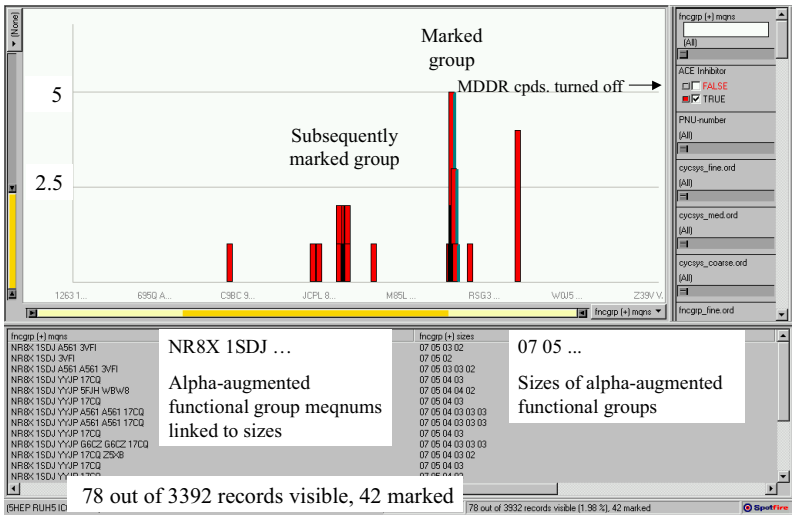


Figure 6. Marked region of ACE inhibitors suggesting alpha-augmented functional groups associated with ACE activity.

upper portion of Figure 3. Alternatively, we could apply the naming algorithm to each of the components and, in that way, obtain a list of numbers. This is illustrated in the lower portion of Figure 3, in which the outcome of the naming function is a meqnum ensemble list. There are k! ways of ordering a list of k numbers. To order the AFG lists canonically, we order the names first by the number of atoms in the corresponding component. When two or more components have the same number of atoms, the numbers are ordered lexicographically. The ensemble meqnum is nice when a short number is required. The meqnum ensemble list gives us substring access to its components and will be used here.

Figure 4 is obtained by simply selecting the AFG meqnum ensemble list variable for the x-axis of the histogram and coloring the bars to indicate the proportion of ACE inhibitors amongst the compounds with a particular set of alpha-

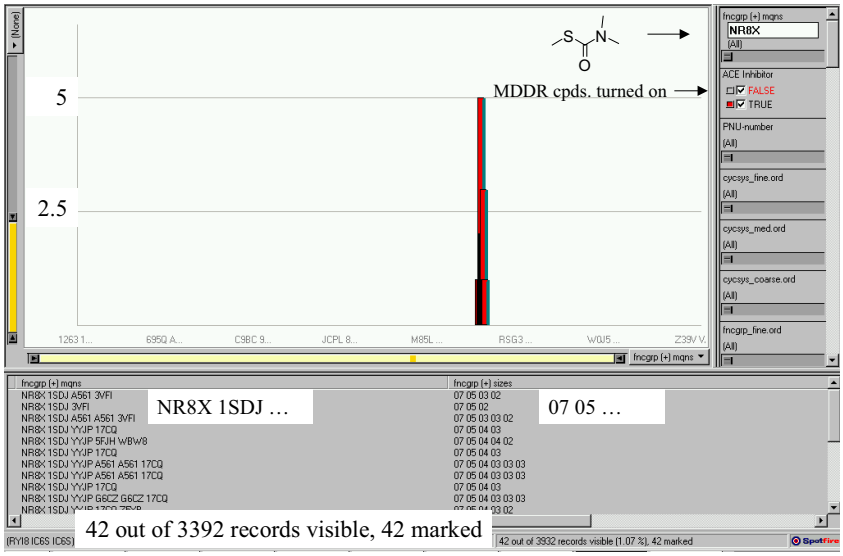


Figure 7. Substring search demonstrating the specificity of a suggested alpha-augmented functional group with meqnum NR8X.

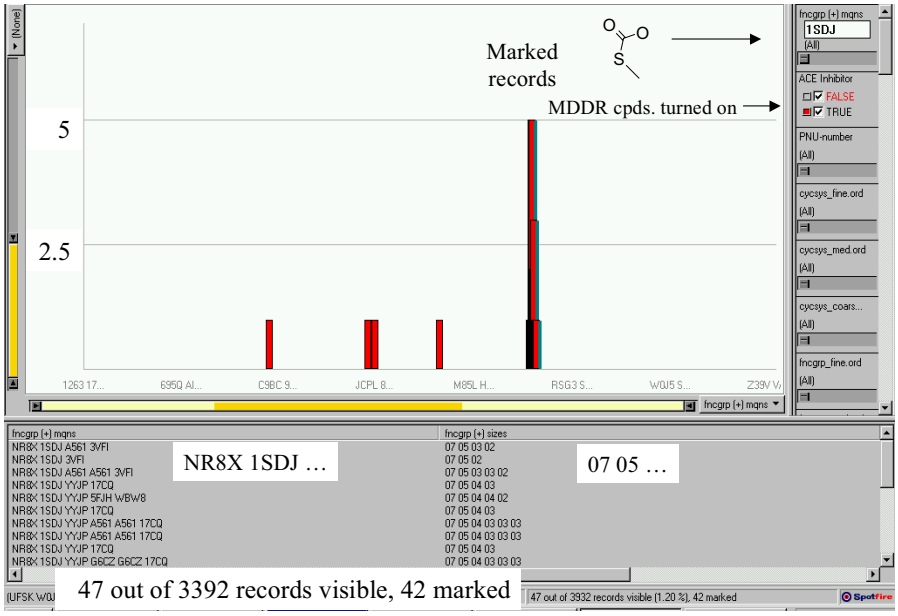


Figure 8. Substring search demonstrating the specificity of a co-occurring alpha-augmented functional group with meqnum 1SDJ

augmented functional groups.

We immediately see that one combination of AFGs is shared by 13 non-ACE inhibitors, and another combination of AFGs is common to 5 ACE inhibitors. However, most of the compounds have a unique combination of AFGs, and consequently, we obtain the black horizontal bar of height 1 along the bottom. The importance of using an meqnum ensemble list rather than a ensemble meqnum is revealed when we use the x-axis slider to zoom in on the narrow region on either side of the red bar corresponding to the 5 ACE inhibitors. This gives rise to Figure 5. Since the AFG meqnums in each ensemble list are ordered first by size, and since the carbamothioate AFG with meqnum NR8X is the largest AFG in quite a few ACE inhibitors, but is not the largest AFG in any non-ACE inhibitors, we obtain a very interesting interval of uninterrupted ACE inhibitors.

Zooming back out and turning off the non-ACE inhibitors, we obtain Figure 6. One can now easily mark the interval of ACE inhibitors displayed in Figure 5. This reveals the AFG lists for each of the marked compounds. Again we note that each begins with NR8X.

To check if the associated AFG occurs on any other compounds, which would necessarily contain another AFG of 7 or more atoms, one enters NR8X in the substring search window for the AFG slider as indicated in the upper-right portion of Figure 7. When finished, all compounds without that AFG are removed from view. In Figure 7, we see that the non-ACE inhibitors have been turned back on! Consequently, we see that all compounds containing the NR8X functional group are ACE inhibitors.

But Figure 6 also reveals that the thiocarbonate AFG 1SDJ is present whenever NR8X is present. Searching for those compounds that contain 1SDJ, we obtain Figure 8. There are 47 such compounds, all ACE inhibitors. The data are inadequate to determine if only one or both of these functional groups is critical to activity in this subseries of the ACE inhibitors.

It is informative to repeat this logic by marking the compound in the “subsequently marked region” in

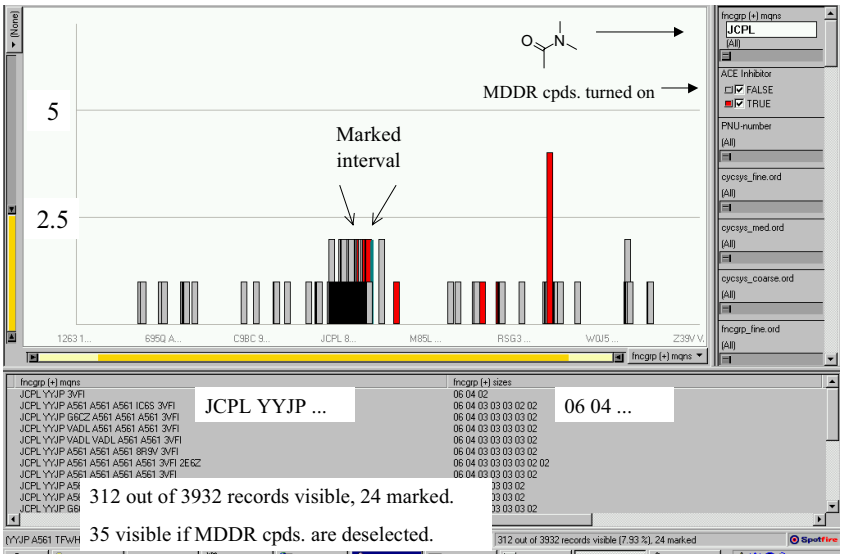


Figure 9. Substring search demonstrating the nonspecificity of a suggested alpha-augmented functional group with meqnum JCPL.

Figure 6. The results are summarized in Figure 9. We see that there were 24 marked compounds whose largest AFG is the amide JCPL. The corresponding substring search reveals a total of 312 compounds with that AFG, 35 of which are ACE inhibitors. Consequently, we conclude that this amide AFG is not ACE-receptor specific, even though it may still contribute to activity when other more receptor-specific structural features are present in a particular arrangement.

A Designed Cyclic System-Ordering Browsing Structures

Efficient systematic browsing requires that structures be linearly ordered. If we are to look at every structure m in a collection of n structures without looking at any one more than once, we would necessarily encounter them in some sequence. One of the most common sequences is defined by the registry number of compounds. Figure 10 shows the first 12 structures one would encounter when lexicographically ordering the 3854 MDDR structures in our data set by their registry number. Although very useful for finding particular compounds when the registry number is known, this ordering does not facilitate our finding a particular cyclic system or getting a good sense of its representatives.

Now suppose the structures were ordered by a MEQI that maps each structure to its cyclic system. Then, for each cyclic system, there would be a single largest interval of compounds comprised of all the compounds with that particular cyclic system. Long/short intervals would represent cyclic systems represented by many/few compounds, respectively. However, adjacent intervals would generally represent compounds coming from entirely unrelated cyclic systems. For example, an interval of steroids might be adjacent to an interval of indoles.

This raises the question as to how one gets closely related cyclic systems to be associated with closely positioned intervals. The natural solution is to develop a hierarchical ordering so that, for example, the compound intervals associated with cyclic systems sharing the same cyclic skeleton are grouped together. Such groupings are easily obtained as follows:

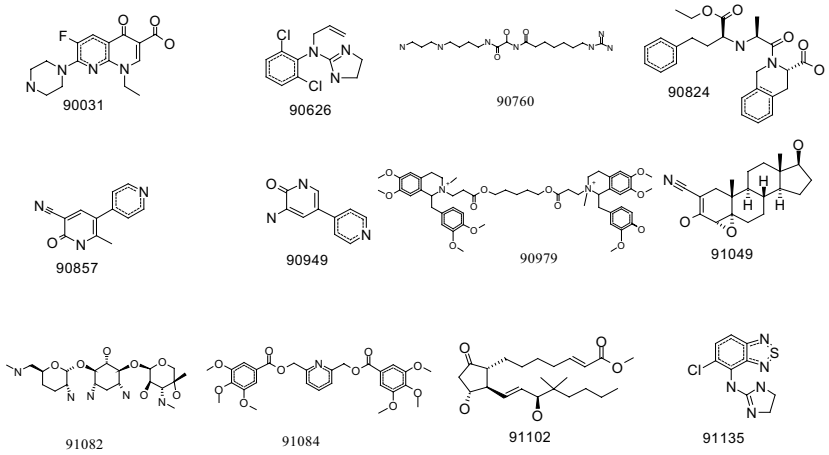


Figure 10. First 12 of 3854 random MDDR structures as traditionally ordered by registration number.

Let SBI_j, j = 1,…,J, be any finite sequence of SBIs. These will usually be a combination of MEQIs associated with the cyclic system skeleton, the set of component ring systems, *et cetera* and suitably chosen counts of the number of atoms, number of component ring systems, *et cetera*. If m denotes an arbitrary structure, then 'SBI_1(m) SBI_2(m) … SBI_J(m)' is a keyword list. A variable taking such keyword lists as values hierarchically orders structures when its values are lexicographically ordered. For example, if J were 2 and SBI_1 and SBI_2 were MEQIs representing a cyclic-skeleton meqnum and cyclic-system meqnum, respectively, we would immediately accomplish our purpose of assuring that compound intervals associated with cyclic systems sharing the same cyclic skeleton were grouped together.

The proof of the relevance of a particular sequence of SBIs in constructing a hierarchical ordering lies in the relevance of the compound orderings that emerge. Such relevance is best demonstrated though numerous examples in a variety of contexts. Space restrictions allow only a rather superficial demonstration of a rather involved cyclic system ordering we are exploring.

The first SBI in the construction of this ordering is the number of ring systems. Since this number is 0 for acyclic structures, all acyclic structures precede all non-acyclic structures in our ordering. Consequently, to extract a short section of the 3854 MDDR structures in our data set that shows that our cyclic system ordering groups related cyclic systems, we list structures 1001 – 1012 in our ordering. The list, given in Figure 11, consists of 12 aromatic, single-ring-system structures beginning with 6 quinoxalinediones, followed by a 1,2,3,4-tetrahydropyrido[4,3-d]pyrimidine-2,4-dione, and then 5 1,2,4-benzotriazin-3-ones. Our perception program currently treats a ketone as an acyclic group. Consequently, the first quinoxalinedione has 3 acyclic groups, the next three have 4, and the last two have 5. Because of this ordering of the number of acyclic groups within a cyclic system, we know there are exactly 3 and 2 single-ring-system quinoxalindiones with 4 and 5 acyclic groups, respectively. Similarly, the interval of 1,2,4-benzotriazin-3-ones begins with two compounds with 2 acyclic groups. The last three compounds have 3 such groups. Consequently, we know there are exactly 2 single-ring-system 1,2,4-benzotriazin-3-ones with 2 acyclic groups in this subcollection of the MDDR.

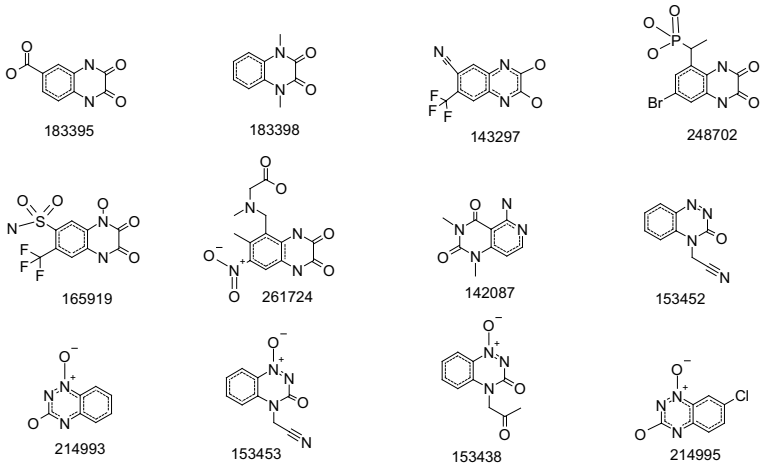


Figure 11. Compounds 1001-1012 in the fine-grained cyclic-system rrdering of 3854 random MDDR structures

Browsing a Structure-Activity Relationship

A visual analysis of a structure-activity relationship (SAR) provides an intuitive feel for the structures on which it is based and roughly determines which structural features are critical to activity. There are many aspects to a comprehensive visual analyses of an SAR. One aspect that is repeatedly encountered is to find a group of compounds with a common cyclic system and similarly positioned side-chains. This is easily facilitated with the joint use of a medium and fine-grained cyclic system ordering. The medium-grained ordering only distinguishes between compounds with different cyclic systems. The fine-grained ordering further distinguishes the compounds by the number of side-chains, how they are positioned, and the particular set of side chains. Figure 12 illustrates how the two levels of resolution work together. The figure is restricted to the 78 ACE inhibitors. The upper histogram has the medium-grained cyclic-system ordering along the x-axis. The lower scatter plot has the fine-grained

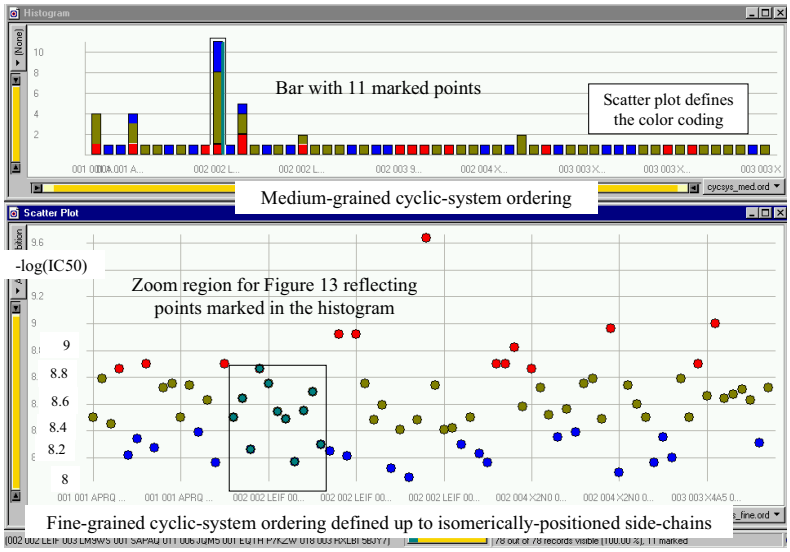


Figure 12. Linked histogram and scatter plot of 78 ACE inhibitors with medium and fine-grained cyclic-system orderings for the x-axes.

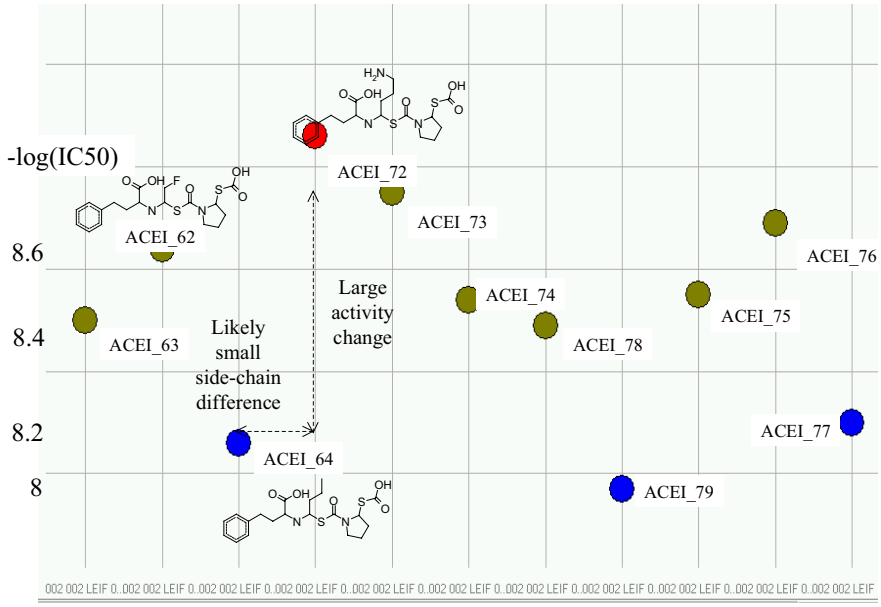


Figure 13. An interval in a fine-grained cyclic-system ordering which uncovered a structure-activity cliff based on small side-chain difference.

cyclic-system ordering along the x-axis and minus the log of the IC50 concentration for the y-axis. The tallest bar in the histogram indicates the presence of a cyclic system represented by 11 compounds. When one "marks" this tallest bar, the corresponding points in the scatterplot are marked as well. These 11 marked points form an interval of contiguous marked points because the fine-grained ordering is simply a further elaboration of the medium-grained ordering.

Because the cyclic-system orderings are based purely on structure, one has no guarantee or even expectation that a particular activity will relate to that ordering. However, one can expect to see closely related structures positioned close to one another. Should these similarly positioned structures differ markedly in activity, we will have found a "structure-activity cliff" where a small structural change is accompanied by a large change in activity. Such an occurrence identifies a critical position in the SA analysis. Figure 13, a blow-up of the marked region in the lower scatter plot of Figure 12, illustrates such an occurrence. Notice that ACE inhibitors **62**, **64**, and **72** have side-chains at the same position and that the number of atoms in the side-chains increases as we move along this particular part of the ordering. As we go from the propyl group to the aminopropyl group, a marked increase in activity is observed, revealing a structure-activity cliff.

POSITIONING MOLECULAR EQUIVALENCE INDICES IN CHEMINFORMATICS

MEQIs are another tool in a long line of tools for organizing and browsing structures. Figure 14 is an attempt to put these tools into a comparative perspective, not with respect to the pros and cons of the possible uses to which such tools have been put, but wit h respect to their mathematical and inferential structure. The major categories along the

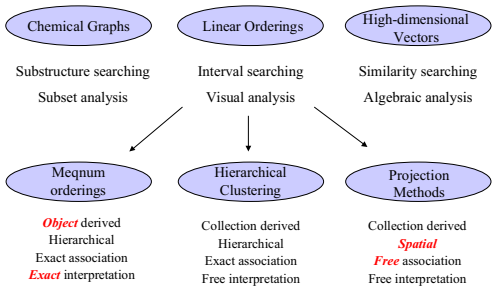


Figure 14. Positioning molecular equivalence indices as natural tools for visually organizing large compound collections.

first row of the figure groups these tools by the underlying mathematical space.

Although complex and difficult to navigate, the space of chemical graphs, partially ordered by the substructure relation, is arguably the most fundamental of the three representations. Because a chemical graph is such a rich storage vehicle, substructure searching gives the user exquisite control in retrieving specific subsets of structures. On the other hand, manually specifying such subsets is too time consuming and the resulting inferential structure too restrictive for most purposes of SAR analysis.

High-dimensional chemical-descriptor spaces have become increasingly important with the advent of similarity searching and the development of data-mining software, especially recursive partitioning programs. The component chemical descriptors usually have very limited structural content by themselves, but taken all together, they can encode a very significant amount of the structural information in a molecule. These spaces are arguably the most simple in that one can define an algebra over them. Consequently, one can "automate" analyses. On the other hand, these high-dimensional spaces are visually unintuitive and often what actually takes place in this automation can differ significantly from what one believes is taking place. (See the paper of this Beilstein workshop by Stanley Young for recent developments along these lines.)

Structural browsing indices are variables whose values are linearly ordered, but there is no restriction that they behave as numbers admitting algebraic operations. The only requirement is that intervals along this linear ordering represent some type of structural commonality. The more such intervals there are, the arguably more rich is the information content of the corresponding index. (One could think of a single fragment chemical descriptor that might be a component of a high-dimensional descriptor space as a browsing index, but it would be a relatively uninformative one. One of its intervals would represent the compounds with the structural fragment and the other would represent the remaining compounds.)

Structural browsing indices have been around for a long time, and have always played an important role in visualizing chemical space. The idea of capturing in a few variables much of the distance information in a high-dimensional point cloud has a long history in statistics and in cheminformatics. Often two principal components suffice. Although some information is sacrificed, much is gained by being able to visualize the captured information in a two-dimensional point cloud. Hierarchically clustering objects and then correspondingly ordering the objects along a line also has a long history, but is receiving renewed interest from the scientific visualization community. (See the papers of this Beilstein workshop by Jeff Saffer for recent developments in visualization methods based on projection and hierarchical clustering.) Meqnum orderings provide a third alternative.

The three types of orderings can be operationally distinguished four ways. First, a MEQI is distinguished from the other two indices in that it can be computed on a single object. The other two types of clustering and projection indices only make sense with respect to a collection of compounds. Their values change with changes in that collection.

Second, the visual grouping of structures is hierarchically organized for MEQI and clustering-based methods whereas these groupings are spatially distinguished in projection methods. This distinction leads naturally into the third distinguishing criteria. Since spatial distinctions rely upon the eye to say whether or not a particular point is or is not in a cluster, the user has

considerable freedom in deciding which groupings of points are clusters and which are not. Operationally, the structural groupings are exactly set fourth when using MEQI and clustering-based methods.

MEQIs are again distinguished from the other two visualization categories when it comes to interpreting the clusters. The interpretation of a meqnum is set forth by the equivalencing function. Moreover, the labeled pseudograph to which a compound is mapped by that function serves as a visual specification of its class with respect to that equivalencing function. This contrasts markedly with the groupings set up via the other two visualization methods. Sometimes these methods generate clusters which admit obvious specifications that distinguish the clusters, but it would seem to be a rare instance where this would be the case if all possible structures were represented in the collection of compounds that was clustered.

SUMMARY AND CONCLUSION

In this study we have attempted a rather broad overview of the types of MEQIs that can be generated and the variety of uses to which they can be put. Our overview is far from exhaustive, and the examples invite further development. Hopefully, this brief sketch of some of the directions we are pursuing in delineating roles MEQIs might play in cheminformatics and structure-activity analysis will suggest areas of interest to others.

REFERENCES AND NOTES

[1] Rouvray, D. H. The Evolution of the Concept of Molecular Similarity. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A.; Maggiora, G. M., Eds.; Wiley Inter-Science: New York, NY, 1990; pp 15-42.

[2] Adamson, G. W.; Creasey, S. E.; Eakins, J. P.; Lynch, M. F. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part. V. More Detailed Cyclic Fragments. *J. Chem. Soc. Perkin I*, **1973**, 2071-2076.

[3] Randić, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for all Self-Avoiding Paths for Molecular Graphs. *Comput. & Chem.* **1979**, 3, 5-13.

[4] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, 39, 2887-2893.

[5] Lawson, A. J. The Lawson Similarity Number (LN): Offline Generation and Online Use. In *The Beilstein Online Database*; Heller, S., Ed.; ACS Symp. Ser **1990**, 436, 143-155.

[6] Johnson, M. A. Browsable Structure-Activity Datasets. In *Advances in Molecular Similarity*; Carbó-Dorca, R.; Mezey, P., Eds.; JAI Press Inc., 1998, 2, 153-170.

[7] Randić, M. Molecular ID Numbers: By Design. *J. Chem. Inf. Comput. Sci.* **1986**, 26, 134-136.

[8] Pearlman, R. S.; Smith, K. S. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 28-35.

[9] Modern Drug Data Report database is distributed by MDL Information Systems, San Leandro, CA.

[10] Read, R. C.; Corneil, D. G. The Graphs Isomorphism Disease, *J. Graph Theory* **1977**, 1, 339-363.

[11] Xu, Y.-j.; Johnson, M. Extending the Morgan Sequence Algorithm for the Needs of Structural Browsing Indices, to appear in *J. Chem. Inf. Comput. Sci*., **2001**.

[12] Morgan, H. L. The Generation of a Unique Machine Description for Chemical Graphs - A Technique Developed at Chemical Abstracts Service, *J. Chem. Doc.* **1965**, 5, 107-113.

[13] Spotfire is a product of Spotfire, Inc. Cambridge, MA (www.spotfire.com).

COMPUTATION AND ANALYSIS OF LARGE CHEMISTRY DATA SETS

S. STANLEY YOUNG AND CHRIS E. KEEFER

Glaxo Wellcome Inc., Research Triangle Park, NC 27709, USA.

E-mail: ssy0487@glaxowellcome.com ; cek43215@glaxowellcome.com

Received: 25th May 2000 / Published 11th May 2001

ABSTRACT

Very large screening data sets are becoming available; hundreds of thousands of compounds are screened against panels of biological assays. There is a need to make sense out of the data; screeners need to know which compounds to screen next and medicinal chemists need to know which series of compounds are active and what features are associated with activity. We use the statistical technique recursive partitioning and simple molecular descriptors, atom pairs and topological torsions, to analyze these data sets based upon the 2D representation of the compounds. We use more general features and a special 3D representation of the compounds for pharmacophore identification. The benefit of this work is that we can rapidly evaluate screening data and make sound recommendations for additional screening work or how to proceed with lead optimization.

INTRODUCTION

Enormous numbers of compounds are now available for screening. Large companies will have over five hundred thousand compounds in inventory; over one million compounds are available commercially; library synthesis offers many millions of possible compounds. It is not feasible to screen all available compounds in all screens. Indeed, with the ongoing genetics efforts there will be an explosion of drug targets over the next several years, increasing the number of available screens.

There is a need to be able to examine screening data and make recommendations on how to proceed.

Which compounds should be screened next? Which compounds acquired for screening? When to stop screening and move to lead optimization? For lead compounds, what are the important features? Statistical analysis of large screening sets can help with all of these questions. In this paper we describe the use of recursive partitioning for the analysis of large chemistry data sets.

METHODS

We use rather simple compound descriptors. See Figure 1 for examples of atom pairs, [1] atom triples, and topological torsions. [2] For pharmacophore identification we use standard pharmacophore features. [3]

Recursively splitting a data set into homogeneous subsets was first proposed by Morgan, and Sonquest. [4] Statistical methods for univariate recursive partitioning are described by Hawkins and Kass, [5] Hawkins *et al*. [6] and Rusinko *et al*. [7] Basically, all potential variables are examined and the single variable that will best split the entire data set into two daughter data sets is selected and the split made; those compounds with the feature go to the right daughter node and those without the feature go to the left. See Figure 2.

Atom pair
N(3,0) - 7 - S(2,0)

Atom triple
N(3,0) - 7 - S(2,0)
S(2,0) - 6 - N(2,0)
N(3,0) - 12 - N(2,0)

Topological torsion
N(3,0)C(2,0)-1-
C(3,1)O(2,0)

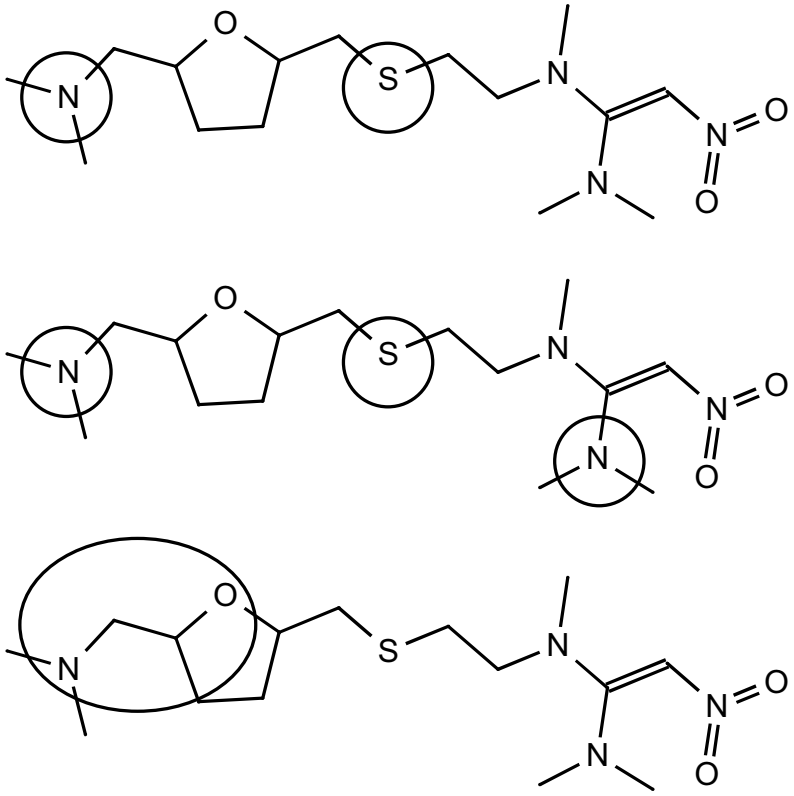


Figure 1. Atom pair, atom triple and topological torsion molecular descriptors.

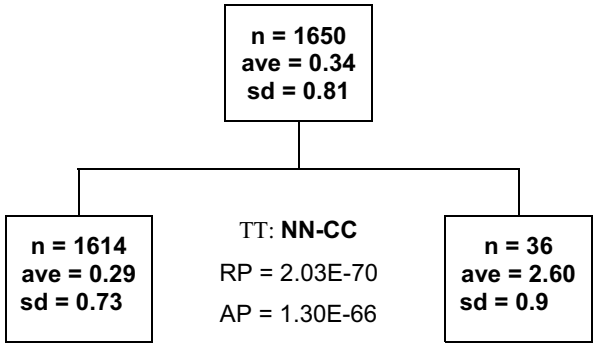


Figure 2: The data set is split using a t-test.

Each daughter node is split in turn. Splitting stops when there are no statistically significant splits remaining. For multivariate recursive partitioning we replace the Student t-test with the Hotelling T^2. [8]

RESULTS

Recursive partitioning is capable of identifying multiple chemical classes of compounds from a data set, and is thus a method for deconvoluting mixtures. [7] Figure 3 gives a skeleton of the recursive partitioning tree.

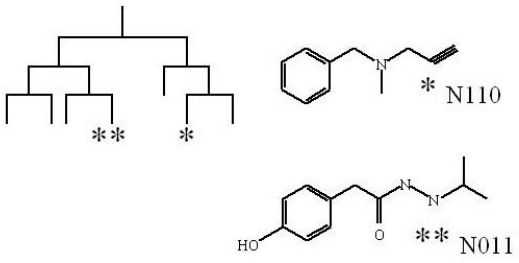


Figure 3: Tree and active compound classes identified.

Also given are representative compounds from two of the terminal nodes. These compounds act through different mechanisms to block the MAO enzyme, see references in Rusinko *et al*. [7]

A data set of 20989 compounds with 4 tumor responses was obtained from the NCI website.

Multivariate recursive partitioning was run. Figure 4 gives a skeleton tree with blowups of two of the terminal nodes. Terminal node N0101 has a

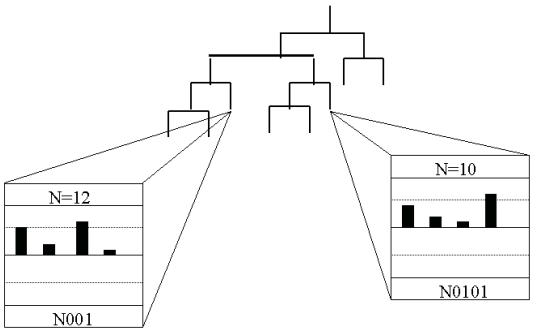


Figure 4: Multivariate recursive partitioning tree, NCI data.

relatively high incidence of the first and last tumor types, Lung and Melanoma, and a relatively low incidence of the second and third tumor types, Colon and Breast. Terminal node N001 has a high incidence of the first and third tumor types. The bits in the node names note the absence, 0, or presence, 1, of chemical features characteristic of compounds in the terminal nodes.

An internal data set of 1444 compounds with IC50 values for the kinase CDK2 was analyzed using typical pharmacophoric features, H-bond donor, H-bond acceptor, *etc*. [3] Multiple conformations were computed and distance between features were binned. After each split, constrained conformations were computed. A total of about 1.4M conformations were computed and the analysis took about 14 hr. CPU time. The resulting recursive partitioning tree is given in Figure 5. The resulting 3D pharmacophore was comparable to crystal structure results, Figure 6.

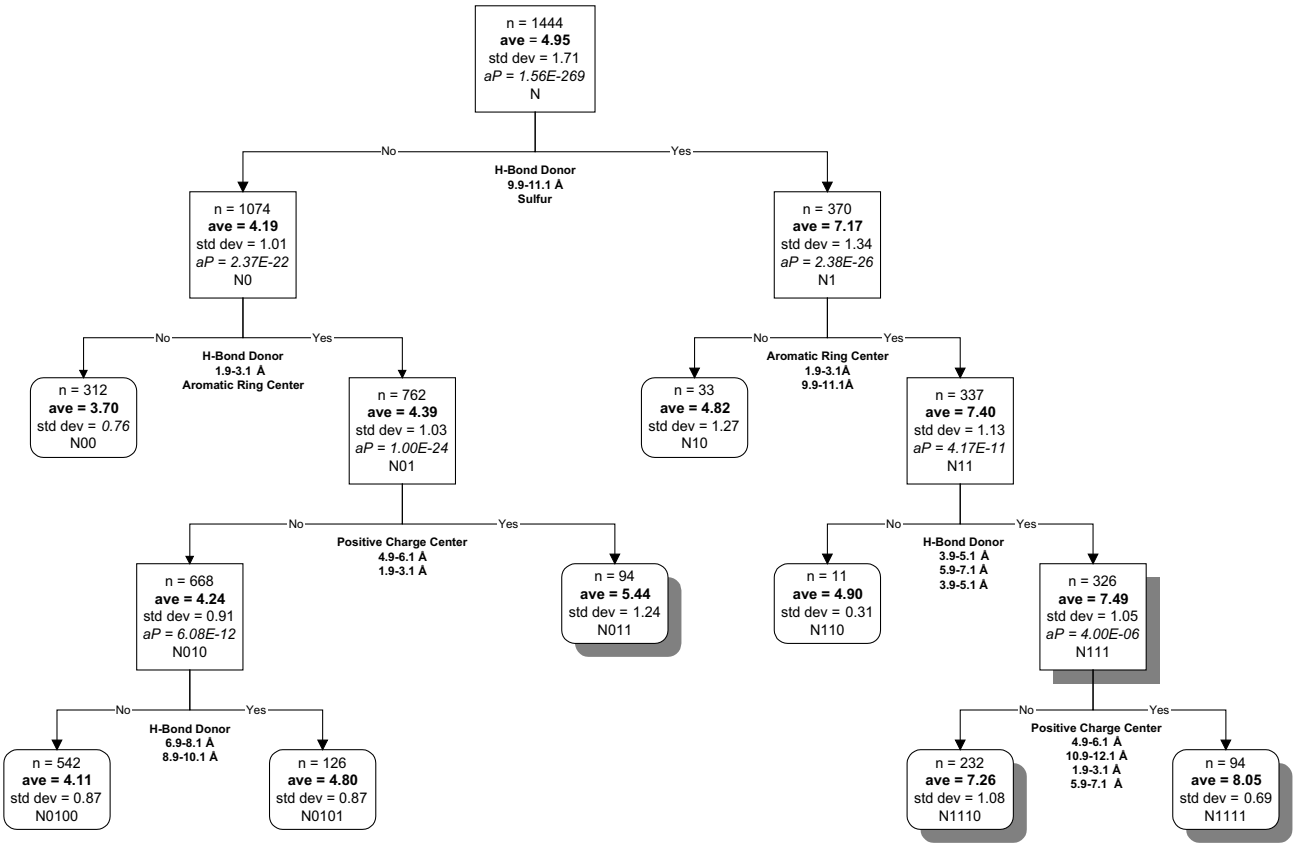
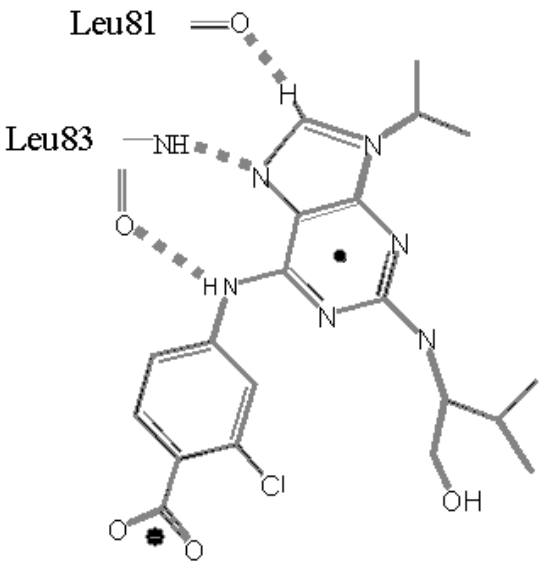
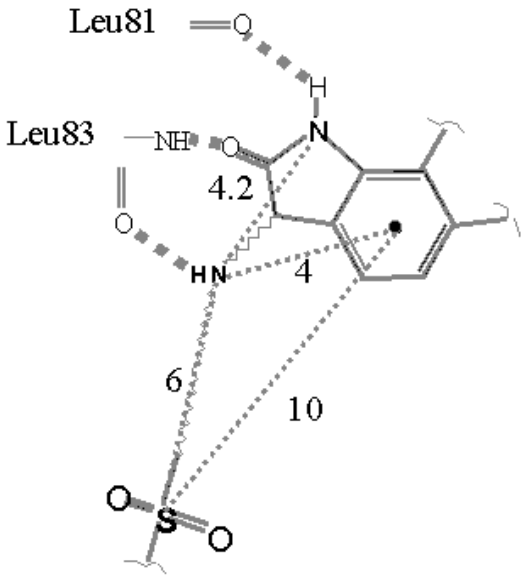


Figure 5: 3D recursive partitioning tree for CDK2 data set.



CDK2-purvalanol B interactions determined by X-ray (*Science* **1998**, 281, 533)



Key interaction features of CDK2 inhibitors determined by 3DSCAM.

Figure 6: Node N111 in CDK2 Tree

DISCUSSION

The key problem to be overcome in the analysis of HTS data sets is that there are likely to be multiple, biological mechanisms. Some molecules may act through one mechanism and others by another. Some might bind in one orientation, others in a different orientation or even at a different location. In the case of the Abbott MAO data set, two mechanisms are known and compounds following each mechanism are found by recursive partitioning. For a large HTS data set there are likely to be multiple mechanisms and even for a single binding pocket, different compounds might bind in different orientations. Most statistical methods assume that there is one underlying model of a single process. If there are two processes, e.g. regular binding site and alosteric binding site, then the features important for one process are very unlikely to be important for the other. Most statistical methods, *e.g.* linear regression, will average the effect for each feature over the two processes. Results are likely to be bad and could be entirely misleading. Recursive partitioning is a simple statistical method that can deal with multiple mechanisms. A feature is identified and the data split based upon this feature. If the feature is important for a specific mechanism, then compounds with that feature (and binding by that mechanism) are separated out from the main body of the data. Following this set of compounds, the analysis is limited to just these compounds; other compounds in the data set have no affect on the subsequent analysis. In this manner, multiple mechanisms can be identified.

A second important problem with HTS data is that assay results for individual compounds are often only crudely determined. Speed and cost are important aspects of HTS. The main goal is to rapidly eliminate the vast majority of compounds

from further consideration. Recursive partitioning does not depend upon a single assay value. Recursive partitioning is driven by averages of compounds with a specific feature and averages are much more stable than single assay values. The node average is the average of all the compounds that have the features that lead compounds into that node. Because the recursive partitioning process is driven by averages, the derived structure-activity rules can have great statistical validity; p-values less than 10^{-100} are common even if the measured effects, increases in binding of less that five percent, are small.

A great deal of effort has been expended implementing these algorithms to make these codes fast. Univariate recursive partitioning runs in seconds for modest data sets, twenty five thousand compounds and ten thousand descriptors. Multivariate recursive partitioning is also fast. This speed has proven to be very useful. Obviously, time is money so completing an analysis quickly can help speed a drug to market. Just as important is that the speed can be used to explore alternative analyses. Medicinal chemists and biologists can interact with the data in real time increasing the likelihood that alternatives are considered and good decisions are made. The statistical methods are rigorous, *e.g*. p-values are adjusted for multiple testing, [9] and help keep the exploratory analysis soundly based.

Atom pairs and topological torsions could be criticized as too simple to be of use for structure activity determination. It is clear that binding into a protein is a three dimensional process; optical isomers often have very different effects. Knowledge of the binding conformation would seem to be essential for good SAR determination. It is clear both theoretically and empirically that these descriptors do capture some structural information. Our empirical results demonstrate that these simple descriptors, coupled with recursive partitioning, are effective in building simple, but useful, structure-activity models.

Building three dimensional pharmacophore models from large data sets is a challenge. We report here on modestly sized data sets, less than 2,000 compounds, where IC50 data is available. Computational speed for 3D recursive partitioning is good relative to commercial codes, but it would be helpful to increase speed. We are working on methods to increase speed with the goal of real-time analysis. In theory, 3D pharmacophore models should be better than 2D methods, but the superiority of 3D over 2D is largely undemonstrated. We plan benchmarking studies to address this question.

REFERENCES AND NOTES

[1] Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 64.

[2] Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1987**, 27, 82.

[3] Chen, X.; Rusinko, A. III; Tropsha, A.; Young, S. S. *J. Chem. Inf. Comp Sci.* **1999**, 39, 887.

[4] Morgan, J. A.; Sonquest, J. N. *J. Amer. Stat. Assoc.* **1963**, 58, 415.

[5] Hawkins, D.M.; Kass, G.V. In *Topics in Applied Multivariate Analysis*; Hawkins, D. H., Ed.; Cambridge University Press, **1982,** 269.

[6] Hawkins, D.M.; Young, S.S.; Rusinko, A. III. *Quant. Struct.-Act. Relat.* **1997**, 16, 1.

[7] Rusinko, A, III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. *J. Chem. Inf. Comp Sci.* **1999**, 39, 1017.

[8] Hotelling, H. *Ann. Math. Stat*. **1931** 2, 360.

[9] Miller, R. G. *Simultaneous Statistical Inference*. Springer-Verlag, **1981**

PREDICTION OF PHARMACEUTICALLY IMPORTANT PROPERTIES FROM MONTE CARLO SIMULATIONS

WILLIAM L. JORGENSEN[a] AND ERIN M. DUFFY[b]

[a] Department of Chemistry, Yale University, New Haven, CT 06520-8107, USA

[b] Central Research Division, Pfizer Inc., Groton, CT 06340, USA
E-mail: william.jorgensen@yale.edu; eduffy@achillion.com

Received: 12th June 2000 / Published 11th May 2001

ABSTRACT

Monte Carlo statistical mechanics simulations have been carried out for more than 250 organic solutes in water. Physically significant descriptors such as the solvent-accessible surface area, numbers of hydrogen bonds, and indices for cohesive interactions in solids are correlated with pharmacologically important properties including the octanol/water partition coefficient (log P), aqueous solubility (log S), and brain/blood concentration ratio (log BB). The regression equations for log P and log S only require 4 - 5 descriptors to provide correlation coefficients, r^2, of 0.9 and rms errors of 0.7. The descriptors can form a basis for structural modifications to guide an analog's properties into desired ranges. For more rapid application, a program that estimates the significant descriptors, QikProp, has been created. It can be used to predict the properties for ca. 1 compound/sec. with no loss of accuracy.

INTRODUCTION

The aqueous solubility (log S), octanol/water partition coefficient (log P), and brain/blood concentration ratio (log BB) of a drug are important factors in determining its bioavailability. Log S reflects the concentration S of the drug in mol/l for a saturated aqueous solution in equilibrium with the crystalline material, while log P and log BB give the log of the concentration ratio of the drug at equilibrium partitioning between octanol and water phases or brain and blood. These quantities affect the ability of a drug to reach significant concentrations in the blood stream and to distribute into tissue. In view of their importance, numerous procedures have been developed for their estimation.[1-8] Most methods start with a structure drawing and have numerical increments associated with large numbers of molecular fragments. For example, the CLOGP procedure of Hansch and Leo uses more than 200 fragment and correction terms to predict log P values. [1]

We recently reported an alternative approach in which a Monte Carlo (MC) simulation is run for the solute in water. [9a] Configurationally averaged results are obtained for physically significant quantities including the solute-water Coulomb and Lennard-Jones interaction energies, solvent-accessible surface area (SASA) and numbers of donor and acceptor hydrogen bonds. Correlations were obtained between these descriptors and gas to liquid free energies of solvation in hexadecane, octanol, and water, and log P. Linear regressions with only 3-4 descriptors yielded fits with correlation coefficients, r^2, of 0.9 in all cases. The regression equation for log P was developed using over 200 diverse compounds and only requires four

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/jorgensen/jorgensen.pdf

descriptors to provide an rms error of 0.55, which is competitive with the best fragment-based methods. Extension of the method to log S was reported using a database of 150 compounds including more than 80 drugs and related heterocycles. [9b] A more rapid procedure, QikProp, has been developed, which uses algorithms to estimate the significant descriptors including the hydrogen-bond counts. No degradation in the quality of the results is found in comparison to the full simulation results. Its application is illustrated here including for predictions of log BB.

COMPUTATIONAL METHODS

The computational details have been described in the earlier work. [9] Briefly, the MC calculations are performed for a single solute in a periodic cube with 500 TIP4P water [10a] molecules at 25 °C and 1 atm. Each simulation consists of sampling 3.2 million configurations for equilibration and 10 million configurations during the averaging phase. The potential energy is represented by harmonic bond-stretching and angle-bending terms, a Fourier series for each dihedral angle, and Coulomb and Lennard-Jones non-bonded interactions. The parameters come from the OPLS-AA force field; [10b] however, since OPLS-AA partial charges are not available for some functional groups, all partial charges are obtained from PM3 calculations using the CM1P procedure. [11] These charges, which are appropriate for the gas phase, are scaled by a factor of 1.3 for neutral molecules in the simulations to reflect the enhanced polarization in the liquid state. The TIP4P water molecules undergo only rigid-body translations and rotations, while the sampling for the solutes also covers all internal degrees of freedom. The MC calculations are run with the BOSS program [12] in an automated manner; only the atomic numbers and a set of starting coordinates are required for the solute.

Twelve descriptors are averaged including the solute-water Coulomb (ESXC) and Lennard-Jones (ESXL) interaction energies, the number of freely rotatable bonds other than for CXYZ (X,Y,Z = H, halogen) groups, SASA and its hydrophobic, hydrophilic and aromatic components, and the numbers of solute as donor (HBDN) and acceptor (HBAC) hydrogen bonds. [9] Hydrogen bonds are defined using a geometric cutoff of 2.5 Å for solute H/water O and solute N, O, or S/water H distances. Results were obtained for more than 250 compounds for log P, [9a] 150 compounds for log S, [9b] and 61 compounds for log BB, [13] that have available experimental data. [1-8,13] Emphasis was placed on representation by diverse structures, functionality, and drugs. The database was maintained and analyzed with the JMP program. [14] F ratios (regression model mean/error mean square) were used to establish the significance of the descriptors; the descriptors reported in the regression equations satisfy the condition that the probability of a greater F value occurring by chance (Prob>F) is less than 0.0001. Cross-validated r^2 values, q^2, were obtained by a leave-one-batch-out procedure using 15 batches of 10 randomly chosen compounds. The database was not split into training and test sets since this is only statistically meaningful for significantly larger data sets.

MONTE CARLO RESULTS

From the Monte Carlo simulations, it was found that log P is well predicted by eq. (1), where the dominant terms are the total surface area and the number of hydrogen bonds accepted by the solute. Corrections are included for the number of non-conjugated amine groups, #amine, and the total number of nitro and carboxylic acid groups, #(nitro+acid).

log P = 0.01448•SASA – 0.7311•HBAC – 1.064•#amine + 1.1718•#(nitro+acid) –1.772 **(1)**

The need for the corrections was traced to deficiencies in the CM1P charge distributions for these functional groups. Increasing size favors solvation in octanol or other organic solvents, while hydrogen-bond acceptor sites favor solvation in water. [3,9] The similar hydrogen-bond accepting ability of octanol and water eliminates the significance of a term for the number of donated hydrogen bonds (HBDN). This simple equation yielded an r^2 of 0.90, q^2 of 0.89, a rms error of 0.55, and a mean unsigned error of 0.44 log unit for the database of 250 compounds.

For solubility, Yalkowsky has noted that log S correlates well with log P with an additional term involving the melting point (MP) for crystalline solutes, eq. (2). [4] MP can be regarded as a gauge of cohesive interactions in the crystal such that a higher MP leads to lower solubility.

log S = 0.8 – log P –0.01(MP – 25) **(2)**

Thus, we initially set out to supplement eq 1 with measures of the cohesive interactions, which could be extracted from the computed descriptors in water. None of the measures of the electrostatic interactions such as the Coulomb energy, ESXC, or the total number of hydrogen bonds, HBAC + HBDN, proved useful. However, ESXC/SASA is a statistically significant descriptor. It can be deemed the Coulomb tension and is large in magnitude for small, highly polar molecules, which have high melting points. Augmentation of eq. (1) with this term led to an equation that yields an r^2 of 0.82 and a rms error of 0.88. However, analysis of the compounds with significant errors pointed especially to heteroaromatic molecules such as pyridines, pteridines, and cytosine, which have an excess of hydrogen-bond acceptor over donor sites. If the sites are not in balance and oriented properly, substantial hydrogen-bonding does not occur in the crystal. To reflect the needed balance, HBDN x HBAC was tried in place of ESXC/SASA, but it did not improve the correlation. However, adjusting this for size with HBDN x HBAC/SASA yields an r^2 of 0.86 and rms error of 0.78. Significant outliers are then prostaglandin E2, chloramphenicol, and mannitol, which have unusually high numbers of hydrogen-bond donor and acceptor sites, and are predicted to have log S values that are too low by 2 - 3 units. With that many hydrogen-bonding sites, it is unlikely that they can all be satisfied simultaneously in the crystal. So, a saturating effect is expected. This can be introduced by applying a fractional power in the descriptor. We arrived at HBAC x HBDN^{½}/SASA as a reasonably simple and effective cohesive index, and the best five-descriptor equation that could be found is eq. (3). The correction for carboxylic acids is no longer significant and has been dropped.

log S = 0.3158•ESXL + 0.6498•HBAC +2.192•#amine – 1.759•#nitro – 161.6•HBAC• HBDN^{½}/SASA + 1.181 **(3)**

Eq. (3) gives an r^2 of 0.88, q^2 of 0.87, a rms error of 0.72, and a mean unsigned error of 0.56 for the 150 compounds. Uncertainty in the experimental data makes it unlikely that predictive schemes for such diverse collections of compounds can yield rms errors below 0.5. [8]

QIKPROP RESULTS

With QikProp, the same descriptors are found to be the most significant as from the Monte Carlo simulations. However, the solute-water Coulomb and Lennard-Jones energies are no longer available, and it is often found that a somewhat larger number of descriptors, ca. 8, are found to be fully significant from the F ratios. For log P, the

regression equation for 270 compounds yields an r^2 of 0.92 and rms error of 0.55. For log S, the corresponding figures for 190 compounds are 0.88 and 0.69. And, for log BB, eq. (4) has an r^2 of 0.84 and rms error of 0.31 with the dataset of 61 compounds.

log BB = 0.001300•FOSA – 0.004332•FISA +0.6337•#amine – 0.0751•μ –0.1369•#rotor + 0.04192 **(4)**

There are only five significant descriptors; hydrophobic surface area and non-conjugated amines increase the brain concentration, while increased polarity, as reflected in the hydrophilic surface area and dipole moment, and flexibility increase the concentration of the compound in blood. The results are illustrated in Figure 1.

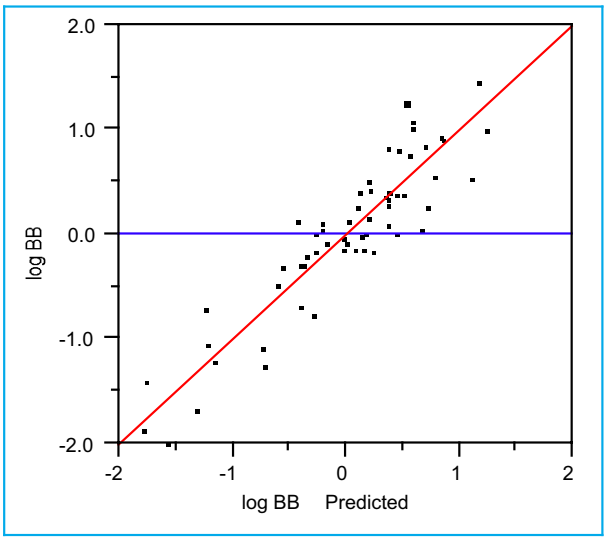


Figure 1: QikProp vs experimental results for log BB.

CONCLUSION

In summary, log P, log S, and log BB can be predicted well using regression equations with only 4-8 descriptors. The descriptors correspond to easily interpreted quantities. They suggest changes that can be made in a structure to guide an analog's properties into a desired range. The current methods are applicable to any neutral molecule with atoms having PM3 parameters, i.e., H, C, N, O, F, Al, Si, P, S, Cl, Br, and I. Improvements are possible through the addition of new descriptors, performance of simulations in different media, and use of alternative partial charges. The descriptors can also be applied to develop correlations for other properties or for refined analyses of narrower classes of compounds.

ACKNOWLDEGMENTS

Gratititude is expressed to the National Science Foundation for support of this research.

REFERENCES AND NOTES

[1] Hansch, C.; Leo, A. "Exploring QSAR – Fundamentals and Applications in Chemistry and Biology", American Chemical Society: Washington, 1995.

[2] Sangster, J. "Octanol-Water Partition Coefficients: Fundamentals and Physical Chemistry",Wiley: Chichester, 1997.

[3] Buchwald, P.; Bodor, N. *Curr. Med. Chem.* **1998**, *5*, 353-380.

[4] Yalkowsky, S. H. "Solubility and Solubilization in Aqueous Media", Oxford University Press: Oxford, 1999.

[5] Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1-18.

[6] Huuskonen, J.; Salo, M.; Taskinen, J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450-456.

[7] Mitchell, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489-496.

[8] Abraham, M. H.; Le, J. *J. Pharm. Sci.* **1999**, *89*, 868-880.

[9] (a) Duffy, E. M.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 2878-2888. (b) Jorgensen, W. L.; Duffy, E. M. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1175-1178.

[10] (a) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926-935. (b) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

[11] Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Aided Mol. Design* **1995**, *9*, 87-110.

[12] Jorgensen, W. L. *BOSS Version 4.2*; Yale University: New Haven, CT, 2000.

[13] (a) Luco, J. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 396-404. (b) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemerr, J.-P. *Pharm.*

Res. **1999**, *16*, 1514-1519.

[14] *JMP Version 3*; SAS Institute Inc., Cary, NC, 1995.

QUANTUM CHEMINFORMATICS: AN OXYMORON?

TIMOTHY CLARK

Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Nägelsbachstraße 25, D-91052 Erlangen, Germany. Tel. +49-(0)9131-8522948; Fax +49-(0)9131-8526565.
E-mail: clark@chemie.uni-erlangen.de

Received: 13th June 2000/ Published 11th May 2001

ABSTRACT

The use of semiempirical MO-theory for complete databases is demonstrated using the example of the Maybridge Chemical Company Database (53,000 compounds). 3D-Descriptors derived from the quantum mechanical wavefunction are used to set up QSPR-models using neural nets as the interpolation technique. Techniques for cross-validation of such models and for calculating individual error estimates for each compound are discussed. The examples are illustrated for properties such as logP, the vapor pressure, aqueous solubility and boiling points. The multi-net method of estimating individual error bars appears to give a good approximation of error limits of ± one standard deviation for several datasets.

INTRODUCTION

Until recently, quantum mechanics calculations were thought of as CPU-intensive and only applicable to perhaps tens of moderately sized (typically under 100 atoms) molecules within a reasonable cost in computer resources. The often described phenomenal increase in the performance of computer hardware has, however, been accompanied by a similar increase in the efficiency of quantum mechanics software, so that, for instance the geometry optimization of ascorbic acid with MNDO, [1] which took about 40 minutes CPU-time on a Convex C1 superminicomputer at the end of 1983, now takes only 5 seconds on an average PC under Windows NT. This, and the fact that most cheminformatics applications are inherently massively parallel through the trivial parallelization of calculating one molecule per processor, make quantum mechanical techniques applicable to tens of thousands of compounds within a single day, as we were able to demonstrate a few years ago. [2] This article is intended to describe the use and applications of semiempirical molecular orbital techniques (exclusively AM1 [3] and PM3 [4]) to complete databases and for the prediction of physical properties. Such techniques are equally well suited to the estimation of biological activity, but this will be the subject of a second article. [5] This article will concentrate on the advantages of using quantum mechanical, rather than classical mechanical, methods and on the derivation of robust, reliable and accurate quantitative structure-property relationships (QSPRs) with individual error estimation for each.

WHY QUANTUM MECHANICS?

Classical mechanical (force field) techniques employ a simple mechanical model of the molecular system. It is therefore not surprising that they do not do as good a job of describing properties that can be derived from the electron density of the molecule such as the molecular

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/clark/clark.pdf

electrostatics, polarizability, ionization potential etc. as quantum mechanical techniques that treat the electrons explicitly. This is illustrated be the molecular electrostatic potentials shown for guanine in Figure 1. Figures 1(a) and 1(b) show the solvent-excluded surface [6] of guanine color coded according to the electrostatic potential at the surface. The color scale is the same for the two figures. Figure 1(a), however, shows the quantum mechanically calculated molecular electrostatic potential (MEP), whereas Figure 1(b) shows the MEP obtained from an atomic multipole model in which the partial atomic charges were fitted to the quantum mechanical MEP using the VESPA technique. [7] Thus, Figure 1(b) represents almost the best approximation to the quantum mechanical results obtainable from an atomic monopole model (not quite the best as VESPA fits to charges outside the molecular surface).

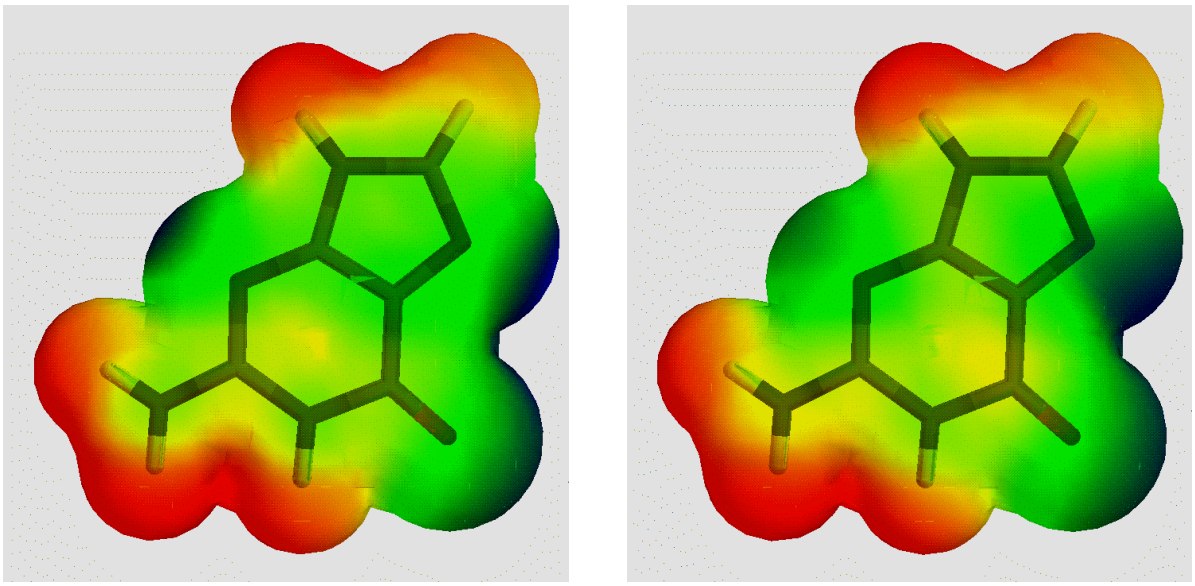


Figure 1: Color coded MEP-surface of guanine (red is positive, blue negative) calculated (left) using the NAO-PC technique [7] from the AM1 wavefunction and (right) using VESPA-derived [6] atomic monopoles.

Figure 2 shows the areas of the surface in which the difference between the two different MEPs is 10 kcal mol^{-1} or more. The surface is now color coded according to the difference in MEPs at the surface. Only the areas in which the absolute difference exceeds 10 kcal mol^{-1} are shown. Red indicates a positive difference and blue negative. The red circles indicate the nitrogen H-bond acceptor regions and the blue ellipse the H-bond acceptor region above the ring system.

The importance of the data illustrated by Figure 2 lies not in the magnitudes of the deviations, although these are significant, but in their positions, The largest concentrations of deviations between the two types of MEP lie at the two hydrogen-bond acceptor site on the ring nitrogens (marked by red

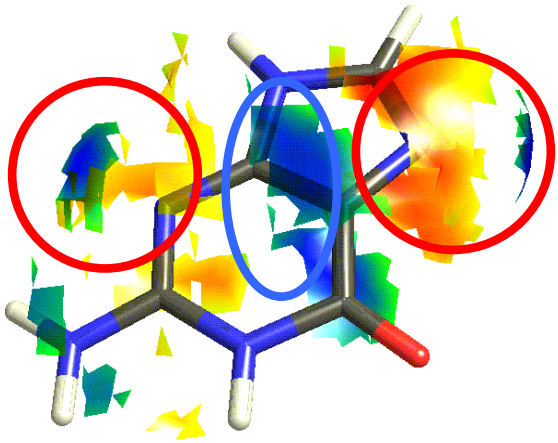


Figure 2: Difference {QM-monopole} of the two MEPs shown in Figure 1, again shown as a color-coded

circles) and at the H-bond acceptor site on the face of the ring system (marked by the blue ellipse). Thus, by projecting the quantum mechanical charge distribution onto an atomic monopole model we lose significant information exactly where it is important for intermolecular interactions.

Thus, we can expect that quantum mechanical methods should describe strong (electrostatic) intermolecular interactions better than atomic monopole based force field techniques. This is, however, not the only advantage of quantum mechanical techniques. Properties such as polarizability, ionization potentials, electron affinities, multipole moments etc. are readily available. Descriptors based on these properties can be expected to play a significant role in QSPRs designed to predict common physical properties.

THE MOLECULAR POLARIZABILITY

Apart from the often dominant and longrange electrostatic interactions, weak intermolecular forces (dispersion) play a major role in determining intermolecular interactions. [8] In order to treat these forces, which dominate for intermolecular interactions between nonpolar molecules, correctly, we need to be able to calculate the molecular electronic polarizability accurately. There are several types of calculational technique available for calculating the polarizability from the molecular wavefunction, but most are too unwieldy to be used routinely for applications on complete databases. Among these are the finite field perturbation method, [9] which, however, is compute-intensive and requires a large, flexible basis set in order to give good results, and the perturbational sum-over-states (SOS) technique. [10] The latter, however, requires a configuration interaction calculation in order to obtain the excited states and is therefore also very compute-intensive. The SOS-method does, however, have the advantage that it can give frequency-dependent polarizabilities.

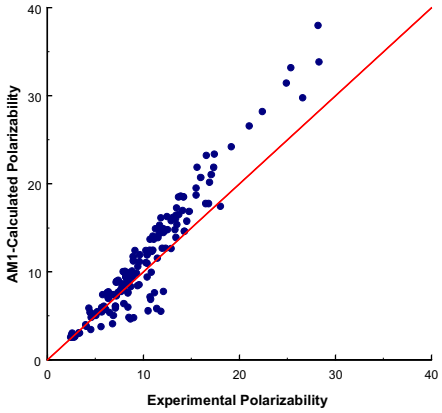


Figure 3: Calculated [11] and experimental molecular electronic polarizabilities (Å^3) using the original variational technique [10] with AM1.

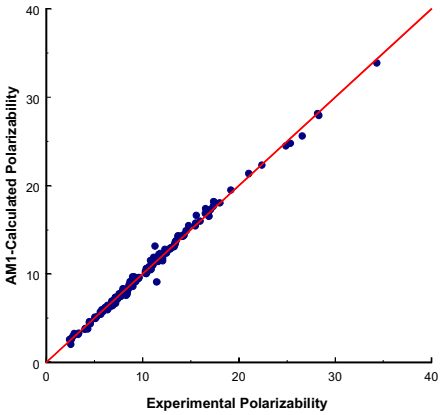


Figure 4: Comparison of calculated and experimental molecular electronic polarizabilities (Å^3) using the parametrized variational technique [11] with AM1.

A more computationally tractable technique that we have used for some years is the variational method developed by Rivail and his coworkers. [11] This technique requires only some multipole integrals and the density matrix and can therefore be appended to a normal SCF-calculation much like a population analysis and without increasing the time of the calculation significantly. Figure 3 shows the results of such calculations with AM1 for a test set of organic molecules.

The results show a systematic deviation for the

larger molecules and a significant scatter for the smaller ones. The standard deviations between calculation and experiment for MNDO, AM1 and PM3 are 1.94, 2.99 and 4.44 Å^3, respectively.

Thus, although the original variational method fulfils the computational requirements for use in a cheminformatics application, it is not accurate enough. In order to remedy this situation, we developed a parameterized variational technique. [12]

If the multipole integrals, which are normally a function of the Slater exponents and ordinal numbers, are treated as variable parameters and the optimized for a set of 156 organic molecules, the results shown in Figure 4 are obtained for the independent test set of 83 organic molecules also shown in Figure 3.

The standard deviations between calculation and experiment for MNDO, AM1 and PM3 are now 0.78, 0.70 and 0.74 Å^3, respectively. Thus, the parameterized variational method offers a computationally economical and accurate method for determining molecular electronic polarizabilities. It also offers the advantage that, with certain restrictions, it can be partitioned into atomic polarizability tensors, which, although not physically measurable, are particularly useful for additive, atom-atom dispersion models.

AM1 Optimizations for a Complete Database

The computational software must fulfill two conditions for a semiempirical technique such as AM1 or PM3 to be applied to a database of perhaps hundreds of thousands of compounds. It must be fast and it must be extremely reliable. Perhaps surprisingly in the light of the introduction, speed is not really a problem. Database applications can use the full power of massively parallel architectures, or even of large compute clusters with relatively slow communication. This is of course because the computational effort per molecule is relatively large and data transfers relatively small and seldom. We reported [1] a benchmark application of AM1 to the Maybridge database [13] a few years ago. The computational protocol necessary to process a 2D-database like Maybridge is shown in Table 1.

Table 1: Processes, software and failure rates for processing the Maybridge database. [1]

Process	Software	# of failures
Data cleanup	SDFClean [14]	211
2D → 3D Conversion	CORINA [15]	41
AM1 optimization	VAMP [16]	68
Generate descriptors	PROPGEN [17]	0
Apply models	PROPHET [18]	0

The data cleanup process is necessary because, even if each structure were entered perfectly, the structures needed for quantum mechanical calculations are not necessarily those entered in databases. Ion pairs, for instance, may be entered as covalently bound structures, free base plus counterion, or in other less standard ways. Because generally the counterion is not considered in quantum mechanical calculations, it must be eliminated and the correct protonation site determined if the free base is entered. Finally, it is also necessary to check that the structures entered in the database make chemical sense. This process resulted in 211 compounds from Maybridge being marked for manual processing, mostly because the exact site of protonation was not absolutely clear. We note here that for many applications it may be preferable to calculate the free base, or even both the base and its conjugate acid.

The 2D to 3D conversion process has been discussed in detail before [19] and will therefore

not be treated here. We used CORINA [15] for the Maybridge run, which resulted in only 41 failures.

The optimization of the molecular geometries with AM1 or PM3 is the most time-consuming step in the entire process. This was performed in parallel (one molecule per processor) on a 128-processor Silicon Graphics Origin 2000. At the time of the run, two processors were defective, giving a total number of processors used of 126. The details of this run have been published, but the essence is that the molecules in the database were optimized within 14 hours elapsed time with only 68 failures. [2] We have since repeated this run several times on distributed moderately parallel machines and on heterogeneous UNIX/Windows NT® clusters with excellent results. Using a Compaq-Alpha two-processor server, a Hewlett-Packard four-processor server and two Intel-based two-processor Windows-NT® machines, for instance, Maybridge can be processed in a weekend. [20]

The descriptors necessary to calculate physical properties can be calculated from the complete electrostatic information stored in the database in a relatively fast step (the most time-consuming task is to generate the potential-derived charges using the VESPA-technique [20]). Finally, the descriptors generated, which are added to the molecular description in the database, are used to calculate properties such as logP [21], the vapor pressure at 25° [22] or the aqueous solubility. [23]

What Factors are Important in QSPR-Models?

Figure 5 shows an overview of typical QSPR-techniques.

The yellow boxes indicate the descriptors used to characterize the molecule. These may be atoms or groups, in which case the interpolation technique used (colored light blue) consists of a set of increments. Such atom- or group-additive methods assume that such increments are transferable and are best suited for properties where this is most likely to be true, such as heats of formation [24] or ^{13}C-chemical shifts. [25] There are a large variety of 2D-descriptors such as, for instance, the range of Kier and Hall indices, [26] although there are very many others. These indices are remarkably successful in treating a large number of properties. They have the advantage that they treat the molecular conformation, if at all, implicitly, so that there is no requirement to locate the most stable conformation or even perform a Boltzmann averaging over a number of conformations. 3D-descriptors, which will be used in the work described here, are derived from the molecule at a given geometry. They are often calculated from the electron density given by quantum mechanical calculations, but this must not be the case. Many descriptors, such as those introduced by Politzer and Murray, [27] describe a property such as the electrostatic potential at the molecular surface. 3D-descriptors are, however, conformationally dependent. This is in principle an advantage, but in practice practically always a disadvantage. This is because the search for the global conformational minimum or a representative set of stable conformations is an extremely compute-intensive task for molecules with a large number of rotatable bonds. Thus, many QSPR-models based on 3D-descriptors actually only use one conformation. This point will be discussed below. Table 2 shows

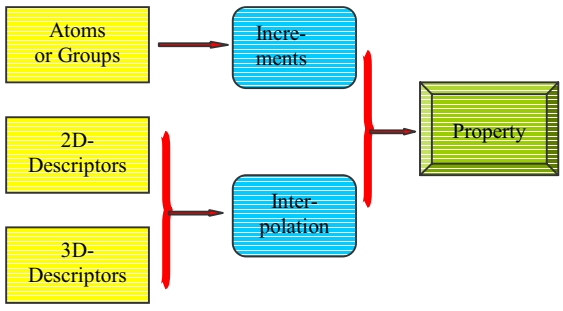


Figure 5: The typical features of QSPR models

the main characteristics of the different types of descriptors:

Table 2: The principal characteristics of different types of molecular descriptors.

Increments	2D-descriptors	3D-descriptors
Fast	fast	can be slow
not universally applicable	general	General
best for additive properties (heats of formation, chemical shifts)	good for many properties	good for properties involving intermolecular interactions
no conformational information	treats conformation implicitly (?)	conformationally dependent

The most traditional interpolation technique is a regression analysis in some form. Alternatives include nearest neighbor techniques, in which the property in question is estimated from those of the most similar known molecules, and artificial neural nets. When used carefully, the latter are extremely powerful but, like all interpolation techniques, they are open to misuse and can simulate a far better performance than they can actually deliver. This leads to a set of requirements for the interpolation used in a QSPR model:

The model should be well validated. This is typically done by some sort of cross-validation procedure in which the predictive ability of the technique, rather than its ability to reproduce known results, is assessed.

The second requirement is that the technique should be as robust as possible. This requirement is often translated as meaning that the model should give a small standard deviation from experimental values for a wide variety of compounds. I suggest, however, that the largest observed error is the most indicative variable for a the robustness of a QSPR-model. The largest likely error is a quantity that defines the reliability of the model for many experimentalists.

Leading from the requirement for robustness is the further desirable feature that the QSPR-model should be able to assess the likely reliability of its prediction *for each individual compound*. Clearly, the properties of s compound that is similar to many in the training set will be predicted more reliability than for one that lies outside its range. The ideal model should not only give its predicted value, but also its estimated error limits.

Quantum Mechanical/Neural Net QSPR-Models

We have in recent years developed a series of QSPR-models based on 3D-descriptors derived from semiempirical MO-calculations and using simple feedforward neural nets with one hidden layer as the extrapolation technique. The general scheme of such techniques is shown schematically

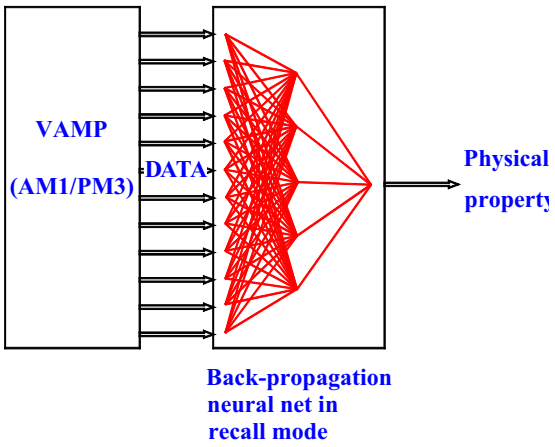


Figure 6: Schematic view of a typical QM/NN-QSPR-model.

in Figure 6.

However, such simple models do not usually satisfy the general conditions for a good QSPR-model given above. We must therefore address the questions of cross-validation and individual error estimates.

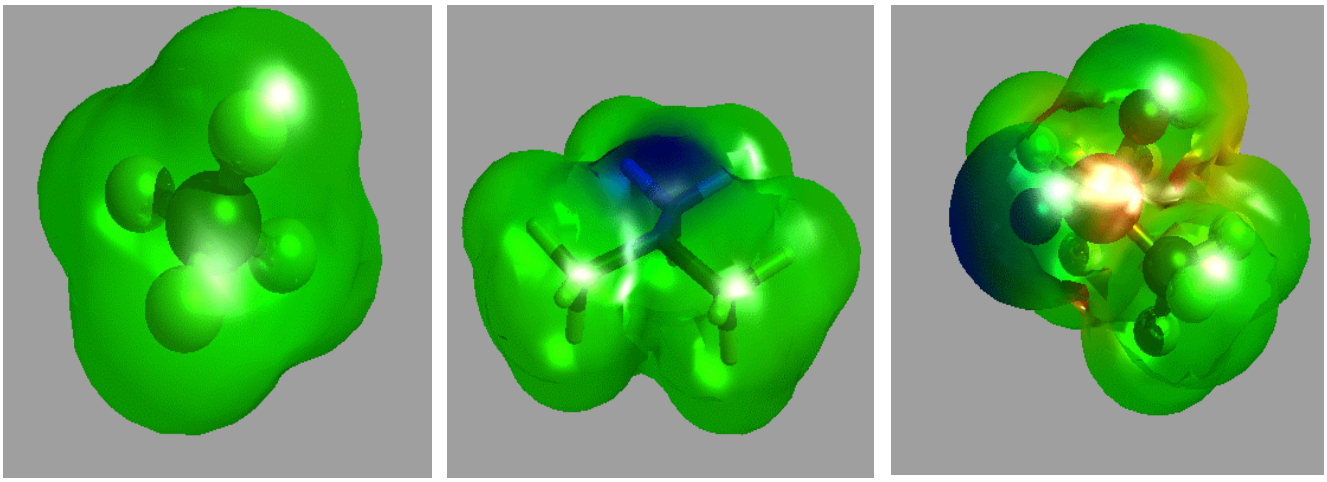


Figure 7: Molecular electrostatic potential surfaces for (from left to right) methane (total variance =5.4, balance parameter = 0.144), trimethylamine (total variance = 446.6, balance parameter = 0.009) and *bis*-(trifluoromethyl)phosphinic acid (total variance = 651.0, balance parameter = 0.246)

We [22] have approached cross-validation by dividing the dataset into about 10 equal, random portions and training 10 separate nets, each using one of the random portions as a test set. This results in 10 different nets, all of which use the same descriptors but which all have different test and training sets. The mean of the results of the 10 nets is used as the predicted value for the model as a whole and the results of the nets for which the compound in question was in the test set are used for cross-validation. In this way, cross-validated results are obtained for each compound in the dataset for a neural net in which it was a part of the test set.

The descriptors used for the QM/NN-models are often those introduced by Politzer and Murray for density functional calculations using the isodensity molecular surface. [27] We use semiempirical MO-theory with the NAO-PC model [28] for the molecular electrostatic potential at the solvent-excluded surface [6] of the molecule. Briefly, Politzer and Murray descriptors describe the statistics of the electrostatic potential distribution at the surface of the molecule. Figure 7 shows some illustrative examples. Methane is essentially nonpolar with very little variation of the electrostatic potential. This leads to a very low variance (5.4). Trimethylamine exhibits an area of negative potential due to the lone pair. This results in a higher variance (446.6) but, because there is no equivalent positive area, a very low balance parameter (0.009). The far more polar *bis*-(trifluoromethyl)phosphinic acid, with both positive and negative areas on the electrostatic potential surface, has an even higher total variance (651.0) and also a high balance parameter (0.246). Such descriptors were designed to describe the intermolecular electrostatic interactions. They have been used in all our QSPR models that estimate physical properties that depend on intermolecular forces. Table 3 shows the parameters used for our published logP model. [21]

These descriptors, of which the sums of the ESP-derived charges probably function as extended atom-counts, can all be linked to logP conceptually. It is noteworthy that the molecular polarizability and the molecular volume, parameters that are generally very strongly correlated, are both necessary in order to generate a reliable model. Figure 8 shows the results obtained using the cross-

validation technique described above.

Table 3: Descriptors used for logP.

[21]

Descriptor	Definition
α	Molecular polarizability
μ	Dipole moment
A	Molecular surface area (SES)
V	Molecular volume
N_{sum}	Sum of ESP-derived charges on N-atoms
O_{sum}	Sum of ESP-derived charges on O-atoms
P_{sum}	Sum of ESP-derived charges on P-atoms
S_{sum}	Sum of ESP-derived charges on S-atoms
X_{sum}	Sum of ESP-derived charges on halogens
V_{max}	Maximum MEP at the SES
V_{min}	Minimum MEP at the SES
M_+	Mean positive MEP at the SES
M_-	Mean negative MEP at the SES
σ^2_{tot}	Total variance of the MEP
ν	Politzer/Murray balance parameter
G	Globularity [29]

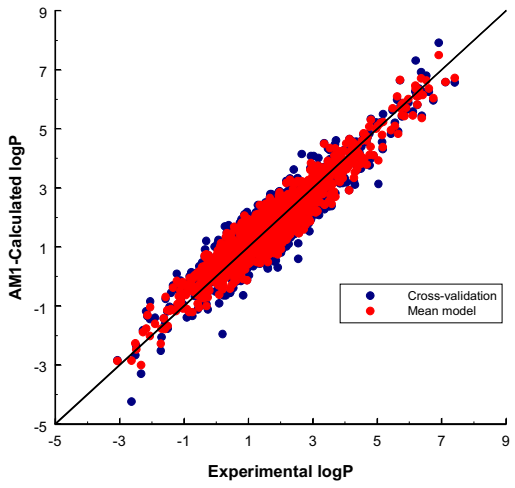


Figure 8: Mean and cross-validated results for the logP model. [21]

Table 4 gives the performance of the mean model and the cross-validation.

The above model appears to be robust as the cross-validation results are comparable to those of the mean of the ten nets. It does not yet, however, give error estimates for individual compounds.

In order to be able to assess individual errors, we [22] calculated the standard deviations of the 10 net predictions for each compound. In principle, the larger the disagreement among the 10 nets, the less reliable should be the predicted value. If now the absolute difference between the calculated (mean model) and experimental value for each compound is divided by the standard deviation of the 10 net predictions for that compound, we obtain the histogram shown in Figure 9.

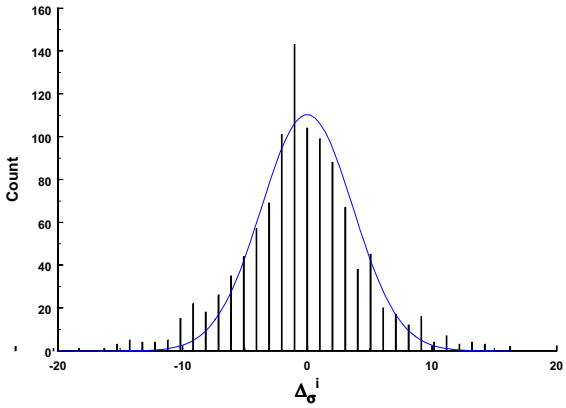


Figure 9: Histogram of the experimental errors in units of the standard deviations of the predictions of the 10 nets for the logP model. [21, 22]

Table 4: Analysis of the mean model and the cross-validation results for the logP model.

Parameter	Mean model	Cross-validation
Std. dev	**0.47**	**0.56**
Max. error	**1.21**	**2.15**
r^2	**0.91**	**0.87**
slope	**1.01**	**0.97**
intersect	**0.01**	**0.06**

The mean absolute value of the deviation in units of the individual standard deviation for each compound is 3.58. We therefore suggest that an intuitively reasonable error estimate for each compound is simply the product of the standard deviation of the net predictions times this mean deviation for the training dataset. [22] If we calculate the error bars in this way for the logP model, we obtain the data shown in Figure 10.

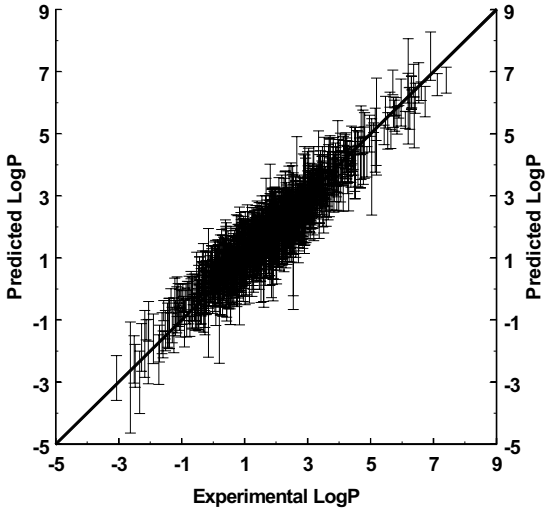


Figure 10: Performance of the logP model with error bars. [21, 22]

This results in 408 compounds (37%) with errors outside the error bars, which corresponds fairly closely to an error estimate of ± one standard deviation. Two questions remain. Is this behavior general for all models and how appropriate are the error bars for completely unseen data?

In order to answer the latter question, we investigated the dataset of nucleotides published bay ACD-labs. [30] These data are not only outside our dataset, but also apply to a class of models explicitly excluded from our data because of the ambiguity of the exact form of the compounds in different media. The results obtained are shown in Figure 11.

Table 5: Performance of three QM/NN-QSPR models.

	Aqueous solubilty	**Vapor pressure**	**Boiling point**
Reference	[23]	[22]	[31]
Units	Log (solubility)	Log (vapor pressure)	°C
Number of compounds	559	551	6,000
Std. dev.	0.51	0.29	16.5
mean unsigned error	0.40	0.22	11.8
maximum error	1.67	1.00	-119
r^2	0.90	0.94	0.96
slope	1.03	1.01	1.01
intersect	0.08	-0.01	-4.6
mean Δ	2.11	2.98	2.15
compounds outside the error bar	201 (35%)	199 (36%)	2244 (37%)

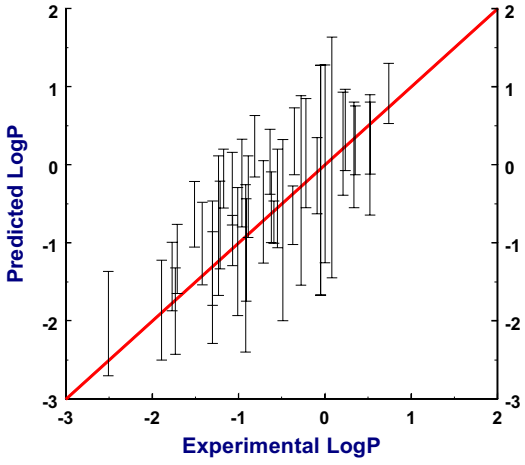


Figure 11: LogP results obtained for the nucleotide dataset. [22, 30]

In this case only 8 compounds (20%) are outside the error bars. This, however, is an anomalous result probably caused by the very low diversity of the dataset, as will be seen in the following

examples. Table 5 shows the statistics of the results obtained for three further models, aqueous solubility [23], vapor pressure at 25° [22] and boiling points at atmospheric pressure. [31] In all cases, the error estimates given by the multi-net technique described above are close to those expected from error bars of ± one standard deviation, confirming the hypothesis that the multi-net technique as described gives reliable error estimates.

Thus, the purely empirical technique of error estimation appears to give reliable results for a variety of QSPR-models and can help to point to compounds for which the neural nets are attempting to extrapolate outside the range of their training sets.

The Effect of Conformational Changes

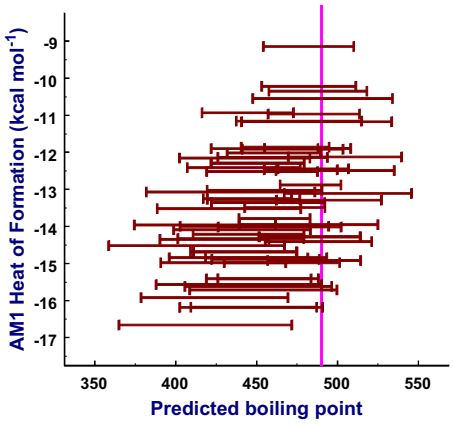


Figure 12: Calculated boiling points for different conformations of bis-(2-aminoethyl)amine plotted against the heat of formation of the individual conformers.

The above models used only one molecular conformation per molecule – that obtained from the Corina-calculated structure after AM1-optimization with VAMP. In principle, models based on 3D-descriptors such as these should be able to describe conformational effects on the property. However, the available data, most of which is for flexible compounds, does not provide us with the necessary experimental resolution to be able to produce a conformationally dependent model. We thus rely on the standard computational protocol to provide us with reasonable conformations. How does this affect the results, however? In order to investigate this effect, we [31] calculated all the minimum energy conformations of *bis*-(2-aminoethyl)amine using the systematic torsional search facility in VAMP. The boiling point model was then applied to each of these conformations, some of which, for instance, contain internal hydrogen-bonds. The results are shown in Figure 12.

In general, the fluctuations in the calculated boiling point are of the same order as the error estimate. The Boltzmann-averaged calculated boiling point is 444±36°, compared with an experimental value of 480°. We therefore feel justified in using the present single conformation approach.

Summary and Conclusions

The techniques described here have demonstrated the applicability of quantum mechanical techniques to cheminformatics. Surprisingly for some, the CPU-requirements are not the major disadvantage of such techniques, but rather the lack of reliable and consistent experimental data and, to some extent, the limitations of current semiempirical methods. For some properties such as aqueous solubility, the published experimental data is too sparse and too noisy to produce a first class QSPR-model. In any case, the available data do not usually allow us to produce a conformationally-dependent model, although normal boiling points may be an exception to this rule. Modern techniques allow us to store essentially the entire electrostatic and polarizability information about a molecule as well as a host of other quantum mechanically derived parameters, so that an

amazingly complete description of the molecules is now available form databases of this type.

Just as the work reported here was impossible at the time of the first Beilstein Workshop (1988), so will the techniques described here be superseded in ten years time? A prime requirement is a semiempirical MO-method that does not suffer the weaknesses of the current techniques for heavy atoms, hydrogen bonds, branching errors and weak interactions. We are currently developing such a technique, which should then provide an even better description of the molecules. However, the "magic limit" of about ±0.5 log units mean error for QSPR-models of physical properties is only likely to be lifted when large (10^3-10^4) numbers of consistent and accurate datapoints become available.

ACKNOWLEDGEMENTS

I thank my postdocs and students, Dr. Bernd Beck, Dr. Andrew Chalk, Dr. Andreas Breindl, Dr. Peter Gedeck, Gudrun Schürer, Maik Gottschalk, Bodo Martin, Matthias Hennemann and Michael Will for their constant support and constructive contributions to this work, which was supported by the Deutsche Forschungsgemeinschaft and Oxford Molecular.

LITERATURE AND NOTES

[1] Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc*., 1977, **99**, 4899; 4907; Thiel, W., *MNDO*, in *Encyclopedia of Computational Chemistry*, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F. III; Schreiner, P. R. (Eds), Wiley, Chichester, 1998, **3**, 1599.

[2] Beck, B.; Horn, A.; Carpenter, J. E.; Clark, T. *J. Chem. Inf. Comput. Sci*. 1998, **38**, 1214.

[3] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P.; *J. Am. Chem. Soc*., 1985, **107**, 3902; Holder, A. J.; *AM1* in *Encyclopedia of Computational Chemistry*, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F. III; Schreiner, P. R. (Eds), Wiley, Chichester, 1998, **1**, 8.

[4] Stewart, J. J. P.; *J. Comput. Chem*., 1989, **10**, 209; 221; Stewart, J. J. P. *PM3*, in *Encyclopedia of Computational Chemistry*, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F. III; Schreiner, P. R. (Eds), Wiley, Chichester, 1998, **3**, 2080.

[5] Clark, T. *QSAR 2000; proceedings of the 13th European Symposium on QSAR*, to be published.

[6] Pascal-Uhuir, J. L.; Silla, E.; Tuñon, I., *J. Comput. Chem.* 1994, **15**, 1127.

[7] Beck, B.; Clark, T.; Glen, R. C.; *J. Comput. Chem.*, 1997, **18**, 744.

[8] Stone, A., *The Theory of Intermolecular Forces*; Vol. 32 of *International Series of Monographs in Chemistry*; Oxford University Press, Oxford, 1996.

[9] Kurtz, H. A.; Stewart, J. J. P.; Dieter, K. M., *J. Comput. Chem*. 1990, **11**, 82; Cardelino, B. H.; Moore, C. E.; Stickel, R. E., *J. Phys. Chem.* 1991, **95**, 8645.

[10] Docherty, V. J.; Pugh, D.; Morley, J. O. *J. Chem. Soc. Faraday Trans. 2*, 1985, **81**, 1179; Zamini-Khamiri, O.; Hameka, H. F. *J. Chem. Phys.*, 1979, **71**, 1607.

[11] D. Rinaldi and J.-L. Rivail, *Theoretica Chmica Acta* 1974, **32**, 243; J.-L. Rivail and A. Carter, *Mol. Phys.* 1978, **36**, 1085.

[12] G. Schürer, P. Gedeck, M. Gottschalk and T. Clark, *Int. J. Quant. Chem*. 1999, **75**, 17.

[13] Maybridge Chemicals Company Ltd., Trevillet, Tintagel, Cornwall PL34 OHW, England.

[14] Beck, B. Oxford Molecular, 1999.

[15] Sadowski, J.; Gasteiger, J. *Corina v. 1.8*, Oxford Molecular, Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, UK.

[16] Clark, T.; Alex, A.; Beck, B.; Chandrasekhar, J.; Gedeck, P.; Horn, A.; Hutter, M.; Martin, B.; Rauhut, G.; Sauer, W.; Schindler, T.; Steinke, T. *Vamp 7.0*, Oxford Molecular, The Medawar Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, United Kingdom, 1999/2000.

[17] Beck, B.; Burkhardt, F.; Clark, T., *Propgen 1.0*, Oxford Molecular, The Medawar Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, United Kingdom, 1999.

[18] Beck, B.; Burkhardt, F.; Clark, T., *Prophet 1.0*, Oxford Molecular, The Medawar Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, United Kingdom, 2000.

[19] Sadowski, J.; Gasteiger, J. *Chem. Rev*. 1993, **93**, 2567.

[20] Beck, B. unpublished results

[21] Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. *J. Mol. Model* 1997, **3**, 142.

[22] Beck, B.; Chalk, A.; Clark, T. *J. Chem. Inf. Comp. Sci*., in the press.

[23] Beck, B.; Clark, T. unpublished.

[24] Benson, S. W. *Thermochemical Kinetics* 2nd ed.; Wiley: New York, 1976; Clark, T.; McKervey, M. A. Saturated Hydrocarbons in *Comprehensive Organic Chemistry*, Barton; D. H. R. and Ollis, W. D. Eds.; Pergamon Press: Oxford, 1979, Volume 1, Chapter 2, 37-120.

[25] Kalinowski, H.-O.; Berger, S.; Braun, S. *Carbon-13 NMR Spectroscopy*, Wiley, Chichester, 1988.

[26] Hall, L. H.; Kier, L. B. *Reviews in Computational Chemistry*, Lipkowitz, K. B; Boyd, D. B. (Eds), VCH, Weinheim, 1990, p. 367; Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976; *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Wiley, Letchworth, England, 1986.

[27] Murray, J. S.; Lane, P.; Brinck, T.; Grice, M. E.; Politzer, P. *J. Phys. Chem.* 1993, **97**, 9369

[28] Rauhut, G.; Clark, T. *J. Comput. Chem*., 1993, **14**, 503; Beck, B.; Rauhut, G.; Clark, T. *J. Comput. Chem.*, 1994, **15**, 1064.

[29] Meyer, A. Y. *Chem. Soc. Rev*. 1986, **15**, 449.

[30] http://www.acdlabs.com

[31] Chalk, A.; Beck, B.; Clark, T. *J. Chem. Inf. Comput. Sci*., 2000, **40**, 1046.

Beilstein-Institut Chemical Data Analysis in the Large, May 22nd – 26th 2000, Bozen, Italy

MOBILE ELECTRONS IN MOLECULES: THE ANISOTROPY OF THE CURRENT-INDUCED DENSITY (ACID) [1]

RAINER HERGES* AND ANDREA PAPAFILIPPOPOULOS

Institut für Organische Chemie, Technische Universität Braunschweig, Hagenring 30, 38106 Braunschweig, Germany.
E-mail: r.herges@tu-bs.de

Received: 19th January 2001 / Published 11th May 2001

ABSTRACT

We have shown that the anisotropy of the induced current density (ACID) can be interpreted as the density of the delocalized electrons in molecules. The ACID scalar field, which can be plotted as an isosurface, is a powerful and generally applicable method for investigating and visualizing delocalization and conjugative effects, *e.g.* stereoelectronic effects in reactions, the anomeric effect, aromaticity, homoaromaticity *etc*.

INTRODUCTION

The problem of localized versus delocalized bonding is almost as old as chemical structure theory itself. The first localized structures were probably drawn by A. S. Couper in 1859 in *Ann. Chim*. [1] and by Kekulé in 1860 in his famous *"Lehrbuch der Organischen Chemie"*. [3] The latter formulae are known as "Wurstformel" (sausage formula).

Only a few years later Kekulé realized that ascribing fixed bonds to carbon does not explain the properties of benzene [4] and he suggested that the six carbon atoms are somehow combined in a common nucleus. In today's terminology we would say that he realized that the localized bonding concept fails in the case of benzene. His rather fuzzy description was criticized by contemporary colleagues, who tried to preserve the fixed bonding concept by proposing localized structures (Claus, [5] Städeler, [6] Kolbe, [7] Ladenburg, [8] Wichelhaus [9] and Meyer [10]). Driven either by his genius or simply by the need to save his six-ring structure, Kekulé proposed a mechanical collision or vibration of the six carbon atoms exchanging double and single bonds. Even though this view might seem quite close to our understanding today, Kekulé did not have a real chance to provide an answer on a sound physical basis.

Delocalization is a phenomenon that can only be explained by quantum theory. Thus the community had to wait for quantum mechanics to enter the field of chemistry. Erich Hückel published the decisive papers on delocalization in 1931 [11] and 1932. [12] He not only explained aromaticity, but also other forms of π-conjugation.

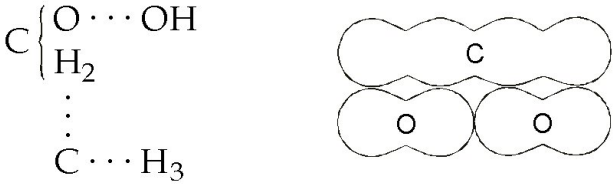


Figure 1: Historical localized bonding concepts of Couper (CH_3CH_2OH, left) and Kekulé (CO_2, right).

Today we use two different concepts to explain delocalization: VB- and MO-theory. In its simplest and most approximate application, valence bond theory describes delocalization by drawing mesomeric structures (mixing VB configurations). MO theory inherently considers delocalization by a linear combination of atomic orbitals to a set of molecular orbitals that extend over the whole molecule. Both methods, however, exhibit the drawback that they are "*unanschaulich*" (not easily interpretable). In larger, and particularly in non-planar systems, the situation becomes complicated and conjugative effects are difficult to "extract" from a number of other phenomena.

Our main goal, therefore, was to develop a method to visualize delocalized (mobile) electrons in molecules. Moreover, the method should also provide a simple means to quantify conjugation. Since delocalization is a quantum theoretical property, (even though it is not an observable) we searched for a suitable interpretation of a quantum chemical observable that avoids empirical parameters.

MAGNETIC PROPERTIES OF MOLECULES, THE ACID METHOD

There are a number of criteria derived from the observables energy and geometry to describe delocalization and conjugation. Conjugation usually leads to changes in energy and geometry with respect to a reference system without conjugation. The choice of the reference system is ambiguous and so are the numbers representing the strength of conjugation. Moreover, the numbers calculated by energy and geometry considerations are not suitable for visualization.

Magnetic properties of molecules have been used to describe aromaticity, which is a special type of cyclic delocalization. The magnetic susceptibility, the anisotropy of the magnetic susceptibility and the NICS method (based on the magnetic shielding) provide numbers that must be compared with reference systems to quantify aromaticity.

Even though these methods provide valuable information, they are restricted to aromaticity and are difficult to visualize as a molecular property with spatial resolution. Closest to a visualization concept are the so-called current density plots. The current density is a vector field obtained by calculating the current induced by an external magnetic field at each point in space. Remember from high school physics that a magnetic field induces a current that follows the "left hand rule" (if the thumb points parallel to the magnetic field **B** the remaining fingers indicate the direction of the induced current **J** e.g. in a solenoid).

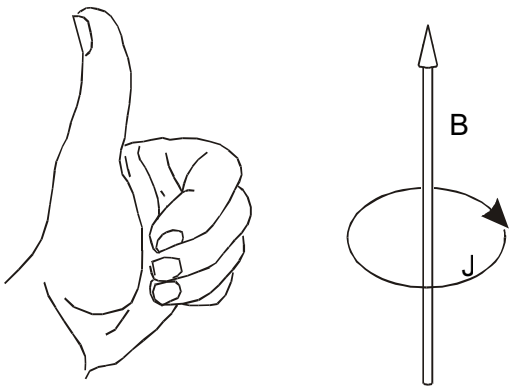


Figure 2: The "left hand rule" for determining the direction of an irduced current.

In quantum mechanics, the situation is more complicated. The quantum theoretical equation for the calculation of the induced current density $\vec{\mathbf{J}}^{(1)}$ is obtained by a first order perturbation treatment and can be expressed in vectorial form as follows: [13,14]

$$\vec{\mathbf{J}}^{(1)} = -i\left(\frac{eh}{2m_e}\right)\sum_{n=1}^{N}\left(a_n - a_n^*\right)\left(\Psi_n\nabla\Psi_0 - \Psi_0\nabla\Psi_n\right) - \frac{e^2}{m_e}\vec{\mathbf{A}}\rho \qquad (1)$$

The summation includes all solutions of the

Schrödinger equation for the unperturbed system. Ψ_n are the corresponding wavefunctions and $\rho = \Psi_0^2$ is the electron density in the unperturbed system. **A** is the vector field. The coefficients a_n are obtained by applying perturbation theory using the magnetic field as the perturbation. Since a vector field is difficult to visualize (a vector is assigned to each point in space), a reference plane in which the current vectors are projected is usually selected (see Figure 3):

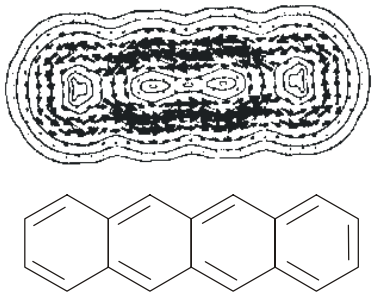


Figure 3: π-current density of tetracene, calculated in a plane parallel to and at a distance of one a_0 from the plane of the molecule (Steiner, E.; Fowler, P. W. *Int. J. Quant. Chem.* **1996**, 60, 609)

The arrows in Figure 3. represent the "interatomic currents" (a notation introduced by London), [15] which are interpreted as mobile or delocalized electrons. Currents that follow the left hand rule are called diatropic and are characteristic of aromatic systems. Those flowing in the reverse direction are paratropic and are observed in antiaromatic systems. The analysis of induced currents is a powerful tool for investigating aromaticity and NMR shielding effects.

However, there are three major drawbacks:

1. Since a graphical 3-D representation of a vector field is impossible (a vector is assigned to each point in space) the method is restricted to planar systems or arbitrary chosen sectional planes.
2. The current density is a function of the overall electron density (see last term in Eq. (1)). Hence, the largest currents are induced close to the nuclei, where the electron density is highest. Since these local currents are much larger than the interatomic currents, they often obscure delocalization effects.
3. Current density maps in terms of delocalization are only interpretable in case of cyclic conjugation (aromaticity and antiaromaticity).

To avoid these problems we must satisfy the following conditions:

1. The parameter representing delocalization should be a scalar field to allow plotting as an isosurface.
2. The scalar field should be independent of the relative orientation of the molecule and the magnetic field (the current density is not).
3. The scalar field should not be a function of the electron density (the isosurface should represent the density of delocalized electrons and not the density as a whole).
4. The method should be generally applicable, not only for aromatic systems but also for any kind of conjugation (through bond, through space, ...) in any kind of system (ground state, excited state, transition state, ...)

The anisotropy of the induced current density $\Delta T_S^{(1)}$ is such a parameter. It can be computed from the current density tensor according to the following equation: [1, 16]

$$\Delta \mathbf{T}_S^{(1)2} = \frac{1}{3}\left[\left(t_{xx}-t_{yy}\right)^2+\left(t_{yy}-t_{zz}\right)^2+\left(t_{zz}-t_{xx}\right)^2\right]+ \frac{1}{2}\left[\left(t_{xy}+t_{yx}\right)^2+\left(t_{xz}+t_{zx}\right)^2+\left(t_{yz}+t_{zy}\right)^2\right] \quad (2)$$

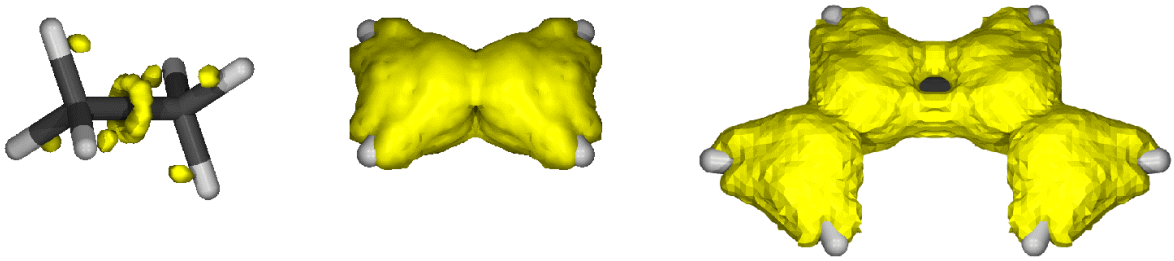


Figure 4: ACID surfaces of ethane, ethylene and *s-cis*-butadiene.

VISUALIZATION

We compute the current density tensor field using the continuous set of gauge transformation (CSGT) method, developed by Keith and Bader [17, 18] implemented in the Gaussian suite of programs. [19] Link 1002 was changed in such a way that the current density vector field was written to a file. The data was transformed to the scalar field of the anisotropy of the induced current density (ACID) according to the above equation and written in the in the cub file format. Isosurfaces were plotted using Povray. For 3D animations we used the Chime plugin, which is able to read cub files.

To provide additional information on the magnitude and direction of currents (e.g. diatropic or paratropic), current density vectors can be plotted onto the isosurface of ACID.

The only parameter that can be chosen in ACID is the isosurface value. This provides control over the sensitivity of the method and a way to quantify conjugative effects (small conjugative effects can be visualized using small isosurface values). We define the isosurface value at which the topology of the ACID boundary surface changes (e.g. breaks in two independent enveloping surfaces) as the critical isosurface value (CIV). The smaller the CIV between two atoms or groups the weaker is the conjugation.

EXAMPLES

We have tested our method extensively. In the first test stage we investigated small and well-known systems to prove consistency with current knowledge. Further emphasis was put on the fact that a broad range of conjugative effects should be covered to prove general applicability. The examples include different types of conjugation such as linear π–, cyclic π– (aromatic), through-bond- and through-space-conjugation. The systems investigated are ground states, excited states, and transition states.

In agreement with the general view of delocalization, alkanes such as methane, butane and cyclohexane do not exhibit delocalized bonds. This is represented by small ACID values around the nuclei and bonds. At isosuface values of 0.05 a.u. (the standard value used in most examples) only small areas of toroidal topology between two bonded nuclei (C-C and C-H) are visible, whereas double bonds exhibit ACID values at least two orders of magnitude larger. Interpreted in traditional terms, this means that the two electrons in a double bond are delocalized over both *p*-orbitals of the sp^2 carbons. In linearly π-conjugated molecules such as butadiene, delocalization is represented by a continuous boundary surface including all conjugated sp^2 carbons. However,

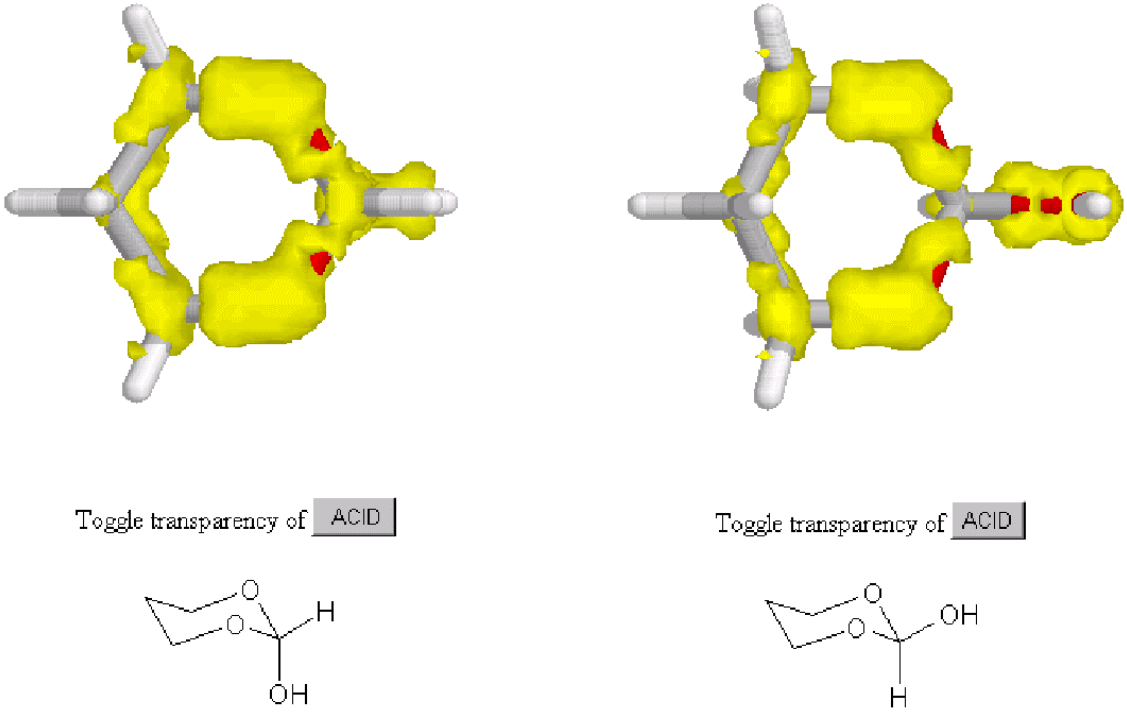


Figure 5: The anomeric effect in 2-hydroxy-1,3-dioxane.

again in agreement with the intuitive view, conjugation is less pronounced between the double bonds than within each double bond. The above defined critical isosurface value (CIV) is lower for the single bond between two double bonds than for the double bond itself. Thus, visualization of molecules using the ACID method is complementary to the information retained from the structural formulae, which only represent localized bonds. ACID plots for ethane, ethylene and *s-cis*-butadiene are shown in Figure 4.

More difficult to represent by traditional methods, and more interesting to investigate, are through-bond and through-space interactions. Figure 5 shows the anomeric effect in 2-hydroxy-1,3-dioxane as an example. For steric reasons (1,3-interactions) substituents in cyclohexane, tetrahydropyran, 1,3-dioxane and other six-membered rings with chair conformations usually prefer the equatorial over the axial position. Exceptions are heteroatom substituents in the α–position to a heteroatom in the ring. This is due to the conjugation of the σ* bond of the exocyclic C-heteroatom bond with the lone pair of the heteroatom in the ring. This conjugation is more favorable in the axial than in the equatorial position. What is difficult to explain within MO theory is instantly visible in the ACID plot.

There is a continuous isosurface from the lone pair of the endocyclic O to the exocyclic O-atom in the axial conformation and there is a discontinuity between the two O's in the equatorial conformation. Hence, the conjugation in the axial conformation is more pronounced, making it more stable. Note that there is also hyperconjugation between the ring O-atom and the neighboring CH_2-group. This is another example of a well-known effect that is difficult to explain in terms of MO-theory but instantly visible in our ACID plot.

IMPLEMENTATION IN A GRAPHIC ENVIRONMENT

Delocalization and conjugation are among the most important concepts in chemistry. These principles

are taught separately in organic, inorganic and physical chemistry from different points of view. The ACID method allows for the first time an integrated approach to teaching delocalization. To this end, we have implemented the ACID plots described above with additional 30 examples in a graphical environment for teaching purposes. We consider the following features to be important for didactical reasons:

1. Figures should replace text wherever possible.
2. Learning information should be divided into modules that fill one screen page (scrolling should be avoided).
3. One module (screen) should present only one main message.
4. The screen pages should present the information in such a way that the message becomes clear just by reading the titles and taking a close look at the pictures (self-explanatory as far as possible).
5. 3D-objects such as molecules and isosurface plots should be represented as 3D objects that can be translated, rotated and zoomed by the user. Additional information not necessary for understanding the main message should only be available in pull down menus.
6. Information containing dynamic data, such as conformational movements or reactions should be represented as dynamic objects (movie). Unlike videos that can be interrupted by pushing a (virtual) button, the movies should advance stepwise by interaction of the user (absolute control of the speed by the user, self-paced learning).
7. Important stages in a movie should be directly addressable by buttons.
8. If a movie contains 3D objects (e.g. molecules on a reaction coordinate) it should be possible to manipulate (translate, rotate, zoom) the 3D-objects in each frame of the movie by user interaction.
9. Interactive features should be used whenever possible (explorative learning), *e.g.* different isosurface values for representation of the ACID should be offered in a menu so that the user can determine the critical isosurface value by trial and error.

Our learning module so far includes 25 molecules as 3D objects, and 6 reactions as movies. The graphic interface (learning environment) will be further refined in an iterative process by testing the system with students.

REFERENCES AND NOTES

[1] The theoretical basis of the paper is outlined in a publication in *J. Chem. Phys.* (in the press)
[2] Couper, A. S. *Ann. Chem.* **1859**, 110, 46.
[3] Kekulé, A. *Lehrbuch der Organischen Chemie* Vol. I, Verlag von Ferdinand Enke, Erlangen **1860**.
[4] Kekulé, A. *Bull. Soc. Chim.* **1865**, 3, 98; *Annalen* **1866**, 137, 169
[5] Claus, A. *Theoretische Betrachtungen und deren Anwendung zur Systematik der organischen Chemie*, Freiburg, **1867**, 207.
[6] Städeler, G. *J. Prakt. Chem.* **1868**, 102, 105.
[7] Kolbe, H. *Über die chemische Constitution der org. Kohlenwasserstoffe*, Braunschweig, **1869**, 11.
[8] Ladenburg, A. *Berichte der deutschen chemischen Gesellschaft*, **1869**, 140.
[9] Wichelhaus, H. *Berichte der deutschen chemischen Gesellschaft*, **1869**, 197.
[10] Meyer, V. *Annalen* **1872**, 156, 295 and 159, 24.
[11] Hückel, E. *Z. Phys.* **1931**, 70, 204.
[12] Hückel, E. *Z. Phys.* **1932**, 76, 628.
[13] Keith, T. A.; Bader, R. F. W. *Chem. Phys. Lett.* **1993**, 210, 223.
[14] Stevens, R. M.; Pitzer, R. M.; Lipscomb, W. N. *J. Chem. Phys.* **1963**, 38, 550.
[15] London, F., *J. Chem. Phys. Radium*, **1937,** *VIII (series VII)*, 397.
[16] Wallenborn, E.-U.; Haldimann, R. F.; Klärner, F.-G.; Diederich, F. *Chem. Eur. J.* **1998**, 4, 2258.
[17] Wallenborn, E.-U.; Haldimann, R. F.; Klärner, F.-G.; Diederich, F. *Chem. Eur. J.* **1998**, 4, 2258.
[18] Keith, T. A.; Bader, R. F. W. *Chem. Phys.*

Lett. **1992**, 194, 1.

[19] *Gaussian 98, Revision A.6*, M. J. Frisch, M. J.;Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A. Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.;Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; . Pople, J. A. Gaussian, Inc., Pittsburgh PA, **1998**.

VISUALIZATION AND INTEGRATED DATA MINING OF DISPARATE INFORMATION

JEFFREY D. SAFFER,* CORY L. ALBRIGHT, AUGUSTIN J. CALAPRISTI, GUANG CHEN, VERNON L. CROW, SCOTT D. DECKER, KEVIN M. GROCH, SUSAN L. HAVRE, JOEL M. MALARD, TONYA J. MARTIN, NANCY E. MILLER, PHILIP J. MONROE, LUCY T. NOWELL, DEBORAH A. PAYNE, JORGE F. REYES SPINDOLA, RANDALL E. SCARBERRY, HEIDI J. SOFIA, LISA C. STILLWELL, GREGORY S. THOMAS, SARAH J. THURSTON, LEIGH K. WILLIAMS, AND SEAN J. ZABRISKIE

OmniViz, Inc., 3350 Q Avenue, Richland, WA 99352, USA.
E-mail: saffer@omniviz.com

Received: 7th July 2000 / Published 11th May 2001

ABSTRACT

The volumes and diversity of information in the discovery, development, and business processes within the chemical and life sciences industries require new approaches for analysis. Traditional list- or spreadsheet-based methods are easily overwhelmed by large amounts of data. Furthermore, generating strong hypotheses and, just as importantly, ruling out weak ones, requires integration across different experimental and informational sources. We have developed a framework for this integration, including common conceptual data models for multiple data types and linked visualizations that provide an overview of the entire data set, a measure of how each data record is related to every other record, and an assessment of the associations within the data set.

INTRODUCTION

Modern methods in the chemical and life sciences are providing data at an unprecedented pace. This is occurring in many areas with multiple types of information. For example, combinatorial chemistry and ultra-high-throughput screening methods are providing incredible numbers of, and information about, chemical compounds. Related screening methods, such as gene chip assays, and the associated expanding world of genome science is also providing information at a very high rate. And data annotations, scientific literature, patents, and a wide range of other documents have text information that is difficult to assimilate due to the sheer volume and complexity.

Given this flood of diverse information, effective and timely use of the results is no longer possible using traditional approaches. With large volumes of information, it is difficult to learn from long lists, tables, or even simple graphs, particularly with multidimensional data. Furthermore, it is clear that more valuable hypotheses can be derived by simultaneous consideration of multiple types of experimental data (e.g. chemical structural information in addition to activity data), a process that is problematic with large amounts of data.

As one solution for moving from large volumes of information to knowledge, we have developed an integrated data visualization and mining framework (OmniViz Pro¶). The primary premise upon which this framework was built is that discovery of the unexpected is a key goal of data mining. That is, in addition to searching for data records of well-defined behavior (testing specific hypotheses),

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/saffer/saffer.pdf

considerable value can often be obtained from assessing all the relationships within the full data set. To this end, there are several full data set overview visualizations that provide value to the analyst. The rationale behind these and the operational issues that have to be dealt with in their implementation are presented here.

CONCEPTUAL DATA MODELS

In working toward an integrated framework for data visualization and mining, we recognized that a common conceptual data model was essential. This conceptual model provides a familiar framework for the analyst and a common view that is independent of data type.

Functionally, this conceptual model can be considered similar to a spreadsheet where each record is a row in the data table and each column contains data describing a distinct attribute. This collection of attributes, or any subset, can be used directly in multivariate analyses. The goal is to use these attributes to define for each record a high-dimensional vector representation that can be used for cluster analysis as well as a common structure for visualization and interaction. Although the mental picture for this paradigm is two-dimensional, functionally the resulting vector space model can be multi-dimensional, providing a framework for integrating different analyses of the same data records.

Multiple data types can be used as attributes in this conceptual model, as with a spreadsheet, providing great flexibility. Numeric data (e.g., screening assay results), categorical data (e.g., functional classification or structure descriptors), genomic sequence (protein or nucleic acid), or even free text can be used. Some of this data can be used directly in high-dimensional vector representations. Other types of data may require the definition of specific descriptors or features, leading to the generation of a new collection of attributes. That is, a column of the data table is translated to a new set of one or more columns. As a result, each data record can ultimately be considered as a vector, whose dimensions are the attribute columns chosen for comparison. Some examples of how this might be accomplished are shown in Figure 1.

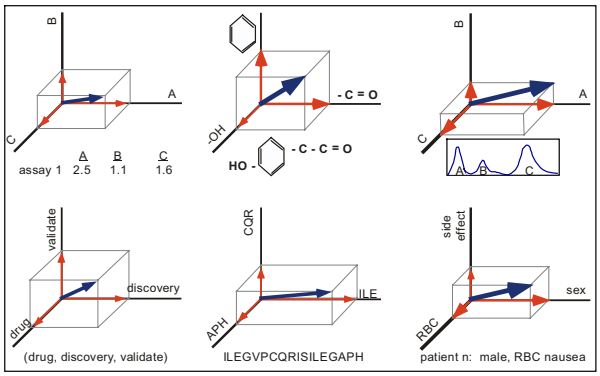


Figure 1: Examples of high-dimensional vector representations for several data types – numeric, chemical structure, chromatographic, text, genomic sequence, and mixed mode (numeric and categorical).

The methods for defining attributes or features for many data types are well known and will not be presented here. However, because of the relatively recent application of these approaches to genomic sequences, it is worth mentioning that a variety of sequence descriptors have been used that in many ways parallel the approaches used for chemical descriptors. For example, van Heel [1] has used a sequence-based method in which each protein sequence is represented the collection of amino acid dimers present in the sequence, somewhat analogous to using contiguous atom pairs for small molecule comparison. More diverse sequence properties have been employed by Hobohm and Sander; [2] in this case, protein sequences were translated to 144 attributes that included sequence components (amino acid composition and a subset of dimers) and several physical-chemical properties. More recently, as for chemical compounds, structural descriptors have been derived for comparing proteins. [3,4]

DATA VISUALIZATION – BASIC CONCEPTS

Exploratory data analysis requires a framework in which

1. the data can be organized along the lines of interest to the analyst and
2. a collection of tools is available for pursuing specific inquiries.

For both, the methods need to handle large volumes of data, with reasonable speed, and provide linkage among complementary views and to other tools.

Presenting data in an organized fashion requires appropriate data overviews, especially those that allow inference by comparison. For this, we have adopted visualization methods since they offer unequalled facility in presenting large volumes of data. In addition, the structure within a well-designed visualization can suggest relationships that might otherwise be overlooked. In that regard, it should be clear that data visualization methods assist, but cannot replace the analyst.

A key component of this approach is to use all the relevant attributes simultaneously for deriving the comparisons. With very large data sets, such as high-throughput screening, it is not possible for the analyst to examine the behavior of the data records a few columns at a time and be able to assess the overall behavior. The selection of attributes for comparison can be useful for testing specific hypotheses, but do not facilitate discovery of the unexpected. Hence, cluster-based methods that utilize all the appropriate data attributes simultaneously are preferred.

Even with mathematical methods that use all the data, no single visualization method can convey all of the information likely to be needed by the analyst and several complementary approaches are necessary. In that spirit, these should not be viewed as stand-alone entities, but linked together for continuity in data analysis. This becomes particularly important in an integrated analysis across different experimental data sets, for example, where distinct visualizations are used to organize the data from separate experimental regimens. The data overviews also need to be supported by complementary tools that support access to and, in many cases, visualization of the details of the data. The easy access to these tools is the foundation for progressing from visualization to data mining.

Given that the data exploration is necessary in the first place since the volume of data is too large to assimilate at once, the key features of the visualization methods are speed and progressive disclosure. Speed is essential since iterative analyses are necessary. Progressive disclosure is a specific type of iteration that is needed frequently. This goes beyond simply zooming in, but rather needs to allow a finer resolution based on comparison of a subset of data records. For example, the relationships uncovered from a subset may be driven by a very different set of attributes than in a full data set comparison.

Finally, recognizing that no exploratory data analysis package can do everything, the visualizations and tools need to provide easy access to external databases and analytical methods. For example, in the bioinformatics realm, the collection of public domain tools is enormous and rather than attempt to duplicate these, all that is necessary is easy export of data from a visualization into these tools and *vice versa*.

DATA OVERVIEW VISUALIZATIONS

As noted above, complementary data overviews are needed to address different aspects of a large data set. We classify these overviews into four types:

- overviews of the data itself,
- overviews of the relationship of each data record to every other record,
- overviews of the associations within the data set, and

- overviews specific to a particular data type.

To enable the discovery process, each of these visualizations must provide ready access to the underlying information and appropriate analytical tools. With these, it becomes possible to explore prior hypotheses as well as the unexpected relationships often suggested by the structure of visualizations of complex data sets.

CorScape

As one approach for viewing an entire data set, we have created the CorScape visualization (Figure 2A). Here, each data record is a row in the visualization and each attribute in the data table a column. Each cell in this visualization is color-coded to represent the actual data. The color-coding can be defined by continuous variables using a color gradient or specific colors for categorical data or missing values. Thus, this is like a spreadsheet with the individual cells color-coded and then shrunk to make it all visible in a single glance.

The rows in the CorScape are ordered for better recognition of the types of behaviors in the data set. Specifically, the records are first clustered (with cluster membership indicated by the alternating gray bars on the left), then the clusters are correlation ordered, and finally the records within each cluster are ordered using a Euclidean distance measure. The result of this layered ordering is the ability to see structure in the data. Furthermore, with large numbers of records (greater than the number of pixels available for the visualization), the ordering allows smoothing with minimal loss of ability to recognize types of behavior.

In addition to the record (row) ordering, the CorScape allows the columns to be ordered in a variety of ways as well. As for the rows, this provides useful structure in the visualization, but moreover provides an analytical tool. For example, consider a visualization of a number of compounds (rows) tested in several HTS assays (columns); arranging the assays by similarity, the analyst can immediately determine which assays may be providing redundant information and allow future screens to be done in a more cost effective manner.

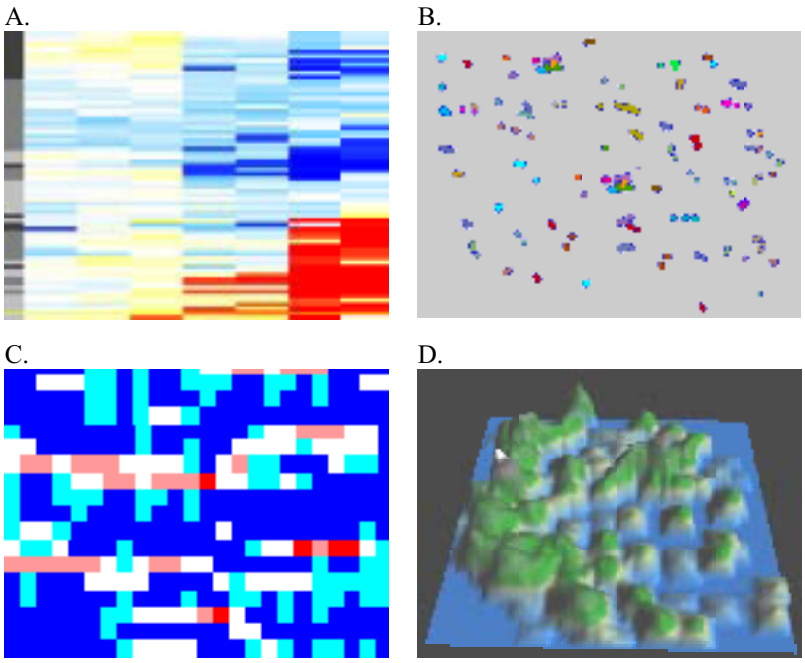


Figure 2: Visualization schemes. A. CorScape. B. Galaxy. C. CoMet. D. ThemeMap.

The CorScape simultaneously provides both a 'far view' which shows the entire data set in one frame and a 'near view' which provides a close-in view of a region of interest in a separate frame. Thus the far view provides the overall context for a data set and the near view allows detailed probing of the data. This two-tiered approach is particularly important for very large data sets.

The approach used in the CorScape visualization is similar in concept to methods employed by Eisen [5] and Weinstein, [6,7] but is done in a manner that is fully interactive. The end result, as implemented in the OmniViz Pro software, is a visualization that allows the analyst to understand the overall nature of the data, discern groupings of records and attributes, and explore the details quickly.

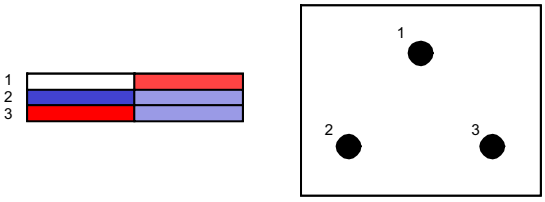
LINKING THE FAMILIAR WITH THE USEFUL

Besides providing a useful overview of all the data, the CorScape provides a link from the data table that is familiar to analysts and the higher-dimensional realm of multivariate data. It shows the information in what is essentially a data table, yet adds information about cluster membership. Thus, the CorScape along with tools, such as the NumericRecordViewer, which shows a portion of the data table with both the color code and numeric values, and familiar analytical tools, such as simple plots, provides a natural transition to higher order analyses.

GALAXY

Although the CorScape provides a ready overview of the overall data set, there is a limitation to the one-dimensional ordering in this type of view. Consider a group of three records in a CorScape, ordered 1-2-3 according to some measure of similarity (Figure 3). It may be that objects 1, 2, and 3 are in fact equally related, as in the diagram to the right. In this case, any order of the three records is correct, a complex relationship that can only be indicated in a higher-dimensional view.

Figure 3:



We have created such a visualization, the Galaxy view (Figure 2B), which is a projection of the data records from the high-dimensional space where the cluster analysis takes place to a two-dimensional view in which interactions can take place. In this view, a point represents each data record and a circle represents each cluster centroid. In particular, the Galaxy view shows how each data record is related to every other data record, with emphasis on the natural groups or clusters that occur within that information space. Thus, this visualization is a representation of the information space that allows the analyst to become oriented rapidly and assess global features of the information.

The Galaxy visualization has some features, such as the representation of an entire data set in single view, in common with other methods that have been applied in the chemical and biological sciences. For example, Sammon maps [8] have been used to compare protein sequences [9] and self-organizing maps [10] have been used for gene expression analysis . [11]

The Galaxy visualization also has several unique attributes to assist the analyst. To help with orientation in the Galaxy view, iconic representation of the behavior of the records in each group, for example, miniature plots for numeric data, provide immediate landmarks on the overall map and allow the user at a glance to see how many

records have what types of behavior. As implemented in OmniViz Pro, the Galaxy, as with the other visualizations, is fully interactive with ready access to data mining tools.

CoMet

To complement the insights about the data and the relationships of data records as gained from the above visualizations, a separate view of how the attributes associated with each data record are distributed is critical. This can be an assessment of how one or more attributes correlate with clusters of records (associating attributes with the group's behavior) or an assessment of how one set of attributes correlate with another set (independent of record to record relationships).

We have created the CoMet visualization (Figure 2C). This view is a data matrix with the rows and columns representing objects or attributes of interest. For example, if the association of a set of attributes with behavior (clusters) is of interest, the rows would represent each cluster (e.g., compounds grouped by biological activity) and the columns would represent the categorical values for each attribute (e.g., structural descriptor). Each cell in this matrix represents the records in that cluster that contain the attribute in that column and is color-coded according to raw occurrence frequency, percent occurrence, or, usually most valuable, the deviation from expected occurrence. In this way, it is easy to see which attributes contribute to the observed behavior. As with the CorScape, additional value in the visualization is derived by appropriate ordering of the rows and columns. For clusters as rows, these are presented in the same correlation order as in the CorScape. The columns can also be ordered (e.g., correlation) to add structure to the view.

Alternatively, the association of attributes with other attributes can be done by selecting the rows to be other attributes - for example, in a preclinical trial, the association of outcome (categorical attributes) with treatment (a separate categorical value). In this case, each cell in the matrix represents how many records contain the attribute in the row and the attribute in the column, with color-coding using the same statistical options as above.

As implemented in OmniViz Pro, the CoMet visualization is also fully interactive, allowing ready access to the underlying information and the relevant analytical tools.

Data Type Specific Visualizations

For some data types, there are specific visualizations that are needed to convey aspects of the information space. In the case of text, we have created the ThemeMap visualization. The landscape visualization metaphor for the major themes within the text provides a rapid means for getting oriented in the two-dimensional Galaxy projection. To this visualization, we have added a suite of tools that facilitate analysis, discovery, and presentation.

Integration

Each of the visualizations described above provides unique value, but should not be viewed in a vacuum. In the course of data exploration, the complementary views need to be linked together so that assessment across separate analyses, different experiments, or even different data types is facilitated. This linkage must essentially be universal within the information space defined by the data set so that examination of subsets of data (e.g., in progressive disclosure) or different subsets of the data attributes can be fully integrated.

Our method for implementing this unified approach is to provide active linkage of records throughout the visualizations and tools. Using an event-driven model, each visualization and each interactive tool

displays the selected records from any other visualization. Thus, records selected in a CorScape view are immediately highlighted in the Galaxy view to link the data overview with the better presentation of record-record relationships. Similarly, records clustered by one set of attributes (e.g., chemical structure descriptors) in one visualization are automatically linked to records in another view clustered by another set of attributes (e.g., biological activity). Linkage from experimental data sets with literature analysis is also possible, through integrated query capabilities. The integration across data sets and data types is facilitated by the common visualization schemes and interactive tools used for all data. This is made possible by the common data table concept; most visualizations and tools access record information through the same underlying data structures.

SUMMARY

As the methods being employed in chemical and life sciences continue to evolve and produce even greater volumes of information, exploratory data analysis will become increasingly dependent on visualization methods. In addition to analysis of specific high-throughput experiments, the integration of multiple experiments across the discovery and development process can be approached. This integration extends across data types to analysis of internal and external data repositories, including historical information such as literature and patents, bringing a new level of continuity to the data mining process.

REFERENCES AND NOTES

[1] van Heel, M. *J Mol Biol.* **1991**, 220, 877.

[2] Hobohm, U.; Sander, C. *J Mol Biol.* **1995**, 251, 390.

[3] Holm, L.; Sander, C. *Science* **1996**, 273, 595.

[4] Holm, L.; Sander, C. *Nucleic Acids Res.* **1998**, 26, 316.

[5] Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. *Proc Natl Acad Sci USA* **1998**, 95, 14863.

[6] Weinstein, J. M.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace Jr., A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W.W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zahaevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D.; *Science* **1997**, 275, 343.

[7] Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S.; Weinstein, J. N. *Mol. Pharma*. **1998**, 52, 241.

[8] Sammon, J. W. *IEEE Trans Comp C* **1969**, 18, 401.

[9] Agrafotis, D. K. *Protein Science* **1977**, 6, 287.

[10] Kohonen, T., Self-organizing Maps, *Series in Information Science, vol. 30*, Springer-Verlag, Heidelberg, **1997**.

[11] Tamayo, P. ; Slonim, D. ; Mesirov, J. ; Zshu, Q. ; Kitareewan, S. ; Dmitrovsky, E. ; Lander, E. S. ; Golub, T. R *Proc Natl Acad Sci USA* **1999**, 96, 2907.

¶ OmniViz Pro, CorScape, Galaxy, CoMet, and ThemeMap are registered trademarks of OmniViz Inc.

INDEX

C

D

N