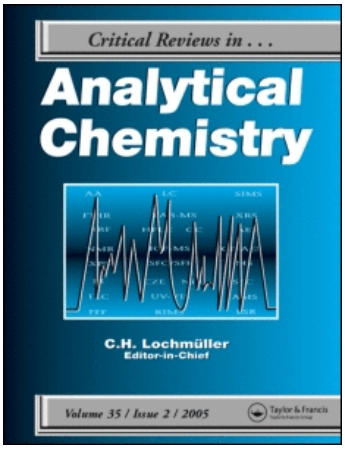
This article was downloaded by:
On: *17 January 2011*
Access details: *Access Details: Free Access*
Publisher *Taylor & Francis*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Critical Reviews in Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400837

Molecular Chemometrics

E. L. Willighagen[a]; R. Wehrens[a]; L. M. C. Buydens[a]
[a] Institute for Molecules and Materials, Radboud University Nijmegen, Toernooiveld 1, Nijmegen, The Netherlands

To cite this Article Willighagen, E. L. , Wehrens, R. and Buydens, L. M. C.(2006) 'Molecular Chemometrics', Critical Reviews in Analytical Chemistry, 36: 3, 189 — 198
To link to this Article: DOI: 10.1080/10408340600969601
URL: http://dx.doi.org/10.1080/10408340600969601

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Critical Reviews in Analytical Chemistry, 36:189–198, 2006
Copyright © Taylor and Francis Group, LLC
ISSN: 1040-8347 print / 1547-6510 online
DOI: 10.1080/10408340600969601

Molecular Chemometrics

E. L. Willighagen, R. Wehrens, and L. M. C. Buydens
Institute for Molecules and Materials, Radboud University Nijmegen , Toernooiveld 1, Nijmegen, The Netherlands

This paper reviews literature from the past 5 years in the field of molecular chemometrics, which applies modeling and data analysis to molecular data. It discusses advances and standing challenges in the fields of molecular representation, similarity and diversity analysis, quantitative structure-activity and structure-property relationship modeling, and library searching.

Keywords chemometrics, chemoinformatics, molecular representation

Downloaded At: 12:17 17 January 2011

INTRODUCTION

Molecular chemometrics is the subsection of chemometrics which applies modeling and data analysis to molecular data, such as found in diversity analysis, property or activity relationship modeling and descriptors calculation. The field includes topics as molecular representation, similarity measures for molecular data, database and diversity analysis quantitative structure activity and property relationship (QSAR/QSPR) modeling, including feature selection and model validation, and methods to automate finding and processing molecular data. We have chosen not to incorporate data of molecular mixtures, e.g., found in proteomics, image analysis, sensor data and multivariate calibration in this review, and also to exclude methods, such as quantum mechanical and force field calculations, because these do not include statistic analysis. The topic is not restricted to just data about the molecule itself, but also includes topics in which the molecule interacts with an environment. Examples of this are the studies where the binding affinity is modeled by representation of both the molecule and the binding site.

The field shows a large overlap with chemoinformatics, though chemometrics tends to prefer to work with numerical data, and in chemoinformatics other fields of mathematics are strongly represented in chemoinformatics too, such as graph theory. There is also overlap with other informatics topics like data mining in general. Library searching is becoming more important again, now that more and more information is available. Increasingly, this information is available in such formats that machines can process the information, and integrate information from different sources. Extensible Markup Language (XML) applications and ontologies play a major role here, and use cases are found in the representation of molecular structures, as well as other representations of molecular information.

Problems intrinsic to this field originate from a few causes. First, the field has to deal with the huge amount of molecules in chemical space; an often cited estimation of the size of chemical space is 1060 unique molecules for structures with a molecular mass up to 500 atomic units (1). Moreover, many molecular properties do not solely depend on the molecular structure itself, but may critically depend on influences from outside the molecule. For example, in binding affinity or toxicity modeling, the activities depend on the protein structure and metabolic pathways, respectively. Additionally, the discrete nature of matter at the molecular level is complicating modeling and analysis even further; we no longer deal with macroscopic properties, and simple physical laws, like the Beer-Lambert equation, get much more complicated when dealing with individual molecules.

This review discusses molecular chemometrics and touches other fields, like chemoinformatics and data mining, focusing on multivariate data analysis of molecular data. Developments reported in literature in the last five years are presented, grouped in four topics: molecular representation; chemical space, similarly and diversity; activity and property modeling; model validation; and library searching. Publications can be found in a diverse set of journals, covering many research fields including machine learning, chemometrics, analytical chemistry, bioinformatics, pharmacy and chemical information. While reviews tend to be biased towards the authors preferred journals, we believe it is safe to say that the *Journal of Chemical Information and Modeling* (formerly, the *Journal of Chemical Information and Computing Sciences*) stands out as a source of related literature. General reading can be found in, for example, *Chemogenomics in Drug Design* edited by Kubinyi and Müller (2), Handbook of u Molecular Descriptors by Todeschini and Consonni (3), and *Handbook of Chemoinformatics* edited by Gasteiger (4).

Address correspondence to L. M. C. Buydens, Institute of Molecules and Materials, Radboud University Nijmegen, Toernooiveld 1, Nijmegen, NL-6525 ED, The Netherlands. E-mail: L.Buydens@science.ru.nl

MOLECULAR REPRESENTATION

Central to molecular chemometrics is molecular data. The data describe chemical facts using a scientifically accepted representation. While in the eighteenth century scientists believed matter to be combinations of elements, it is now accepted that molecules are combinations of atoms held together in specific bonding patterns, governed by quantum mechanics (5). Hence, molecular compounds are no longer identified by a pseudo molecular formula, but more detailed molecular representations are used. Molecular representations are, however, often not suitable to be use directly in data analysis and modeling; instead, descriptors derived from these representations are used which match the data analysis and modeling process. Figure 1 shows the relation between these representations and data analysis and modeling, and the role of descriptors in that relation.

Several basic approaches now exist to describe (small) molecular structures, each with specific function and characteristics: the first is a set of three-dimensional atomic coordinates, for example, used in crystallography. The second is the quantum mechanical representation, where molecular structure is, in principle, a linear combination of atomic orbitals represented with three-dimensional equations, called basis functions. Deriving properties based on this representation require a lot of computation time and scales badly with a growing number of atoms. It is, therefore, not much used in molecular chemometrics; its use is outside the scope of this article. Third is the graph based representation, where atoms are nodes and bonds are edges.

The abundant mathematical literature on graph theory makes this representation historically successful, and even now used a lot in new research. This representation is unable to represent electron systems that cover more than two atoms, e.g., delocalization and multi atom bonds. These features are important in, for example, organometallic compounds where metals bond to electron systems, instead of atoms directly. Modifications have been proposed that allow electron systems with more than two atoms (6, 7), but application of this representation is not yet common in chemoinformatics.

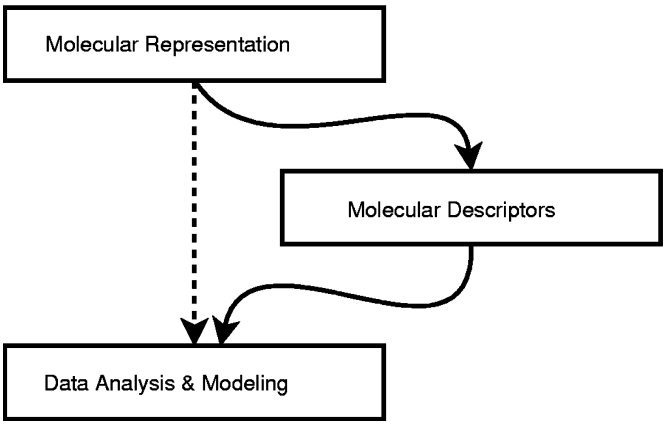


FIG. 1. Common molecular representations, such as the quantum mechanical and the graph representation, are not well suited for direct use in statistical data analysis and modeling; instead, descriptors derived from these representations are used that match the data analysis and modeling process.

Many chemometrical modeling methods, however, require a numerical, fixed length, vectorial representation of the molecular structure (8); the preceding representations do not fulfill this requirement, and hence derived descriptors have been and still are being developed to bridge the gap between those representations and the mathematical modeling methods. These descriptors allow statistical modeling and analysis with, for example, classical methods like principle component analysis (PCA), partial least squares (PLS), neural networks (NN), and classification methods like linear discriminant analysis (LDA). Only very few methods, such as classification and regression trees (CART), do not require a numerical representation. Using distance-based clustering, is another example, and, for example, the distance measure based on the maximal common substructure: the more substructures two molecules have in common, the smaller the distance.

The currently used representations each have a specific field of application, and seem sufficiently adequate to solve a diverse set of chemical problems. However, all of them have limitations that restrict the applicability. With quantum mechanics on one side and the graph(-like) approaches on the other side, one might suggest there is room for an intermediate representation, allowing new types of derived derived descriptors for use in data analysis and modeling.

Molecular Descriptions

Todeschini published the *Handbook of Molecular Descriptors* in 2000 (3), giving a broad overview of known molecular descriptors at that time, but a universal descriptor has not been found (9), and the search for new descriptors has not stopped. Depending on the information content, descriptors are usually classified as 0D, 1D, 2D and 3D descriptors. The first category encompasses descriptors that do not take into account the molecular structure, e.g., the molecular mass and atom type counts. Where 2D descriptors are derived from the molecular connectivity, 1D descriptors can be considered a substructure list representation. The last category, 3D, additionally takes into account the three-dimensional geometry of the molecule. Recently, a fifth category has been proposed, 4D descriptors, but several different definitions have been given. Todeshini defines the 4th dimension to describe the interaction field of the molecule (3), while others reserve this dimension to describe conformations of the molecule (10).

That insight in the three-dimensional interaction of ligand with protein cavities is important in the modeling biochemical endpoints, such as binding affinity, became apparent, and computationally feasible, in the last decade. Comparative Molecular Field analysis (CoMFA) is the primary example of this concept (11). The CoMFA method studies molecule-environment interaction by putting the molecules in an equidistant grid of

Downloaded At: 12:17 17 January 2011

points in three-dimensional space. At each point, the interaction energy is calculated using a hypothetical probe, for example, using the Lennard–Jones potential function and the Coulomb potential energy function. It is important to note that, because the molecules are aligned, the interaction similarities of the ligands can be compared, by comparing the interaction energies of the same grid point for all molecules. Then, PLS is used to correlate the matrix expansion of the grid with the activity or property, though CoMFA is mostly applied to ligand-target binding properties (12, 13).

CoMFA requires, however, geometrical alignment of the molecules and only considers one conformation for each molecule, which is only a simplification of reality. Therefore, focus moved on to descriptors that are independent of the orientation of the molecules in its reference frame, and possibly even include information of multiple conformations. This was already acknowledged in 1997, for example, by Hopfinger who made a scheme that incorporated some ideas from CoMFA, but which also was alignment independent, and took into account multiple conformations (14). Based on this concept Senese developed a 4D-fingerprint (15), which uses single value decomposition to transform the aforementioned representation. This vector representation still contains geometrical information about the possible conformers of the molecule.

In the last 5 years, several new descriptors have been published. Bursi proposed the use of experimental or simulated infrared and 1D NMR spectra as molecular descriptor (16). Most commonly used is the whole spectrum approach (16–20), which takes the whole spectrum as descriptor. Alternatively, the chemical shift of an atom present in all compounds can be used (21, 22). The advantage of this method is that it explicitly focuses on information relevant to the problem; for example, when modeling chemical reactivity, one can take the chemical shift of an atom close to the reactive center. Spectrum-derived descriptors are also used, such as the accumulative differences in peak shifts of nuclei in octanol and in water to model the partition coefficient between those solvents (23). The whole-spectrum approach was recently shown not to be suitable for all applications (24).

Not only new 3D descriptors have been introduced, but also new 2D connectivity-derived descriptors have been developed. For example, Faulon introduced the molecular signature which describes molecules by a vector of integers (25). The length of the vector is determined by the number of unique fragments in the data set. Each integer, then, indicates the number of occurrences of the fragment in the molecule. The fragments themselves are called atomic signatures, and are line notions of the connectivity of that atom, very much like the HOSE code (26). While substructure-based fingerprints only list the occurrence of a number of fragments, the signature describes fragments for all atoms, and therefore, the full connectivity of the molecule. Randic used a specific counts paths of length three descriptor to include the model the boiling points of alcohols (27). The include the steric hindrance around the oxygen he counted paths that included this oxygen; more path counts indicates more hindrance. Use of counts paths descriptors is not novel, but this example shows how descriptors can be customized for a specific problem.

Mansfield proposed a new class of 84 shape descriptors based on the volume distribution in three-dimensional space (28). In itself, these descriptors are dependent on the alignment, which can be used to align molecules such that their volume distribution shows the best overlap. Tuppurainen introduced the electronic eigenvalue descriptor, which describes molecules by a smoothed function of the orbital eigenvalues put on an electron volt scale (29). This descriptor is alignment independent and represents electronic substituent effects. The article shows the application in three QSAR studies for phenyl containing molecules. Stiefl developed a descriptor that represents molecules by a one-dimensional transform of a property mapped on the molecular surface (30). The one dimensional mapping is to make the descriptor orientation independent, and shows the number of interactions between surface points, split up by distance between points, and the type of interaction, e.g., an interaction between a hydrophobic and a hydrophylic surface point.

Descriptors are a projection of the information in the molecular representation onto a lower dimension space, and, therefore, they inherently focus on a specific part of that information. Though many have theorized about this, and ideas and assumptions are present, it is generally still difficult to pick the right descriptor (set) for a given chemical problem. Should global information be taken in account, or local information, or the right mix of both? This limited understanding of how things work at a molecular level even becomes worse if the molecule starts to interact with an environment. Is only one specific conformer important, and if so, which one? Or is the system much more dynamic and should this be taken into account too, for example, when modeling the binding affinity of a ligand with a protein when this protein is also receptor at a different location, affecting the binding cavity of this ligand and thus the affinity? Much of this is unexplored territory.

Beyond the Molecule

The use of descriptors to model molecular properties is not restricted to QSAR and QSPR. Other fields are picking up these descriptors too, e.g. in the field of proteomics, where Lapinsh and Prusis developed an extension of QSAR coined proteochemometrics (31, 32). In this method, binding affinities are modeled on both the structure of the ligand, small peptides in this case, and the structure of the receptor. They found that including cross terms of the ligand and receptor descriptor blocks significantly improved the models, not surprising for binding affinity modeling, and stresses that binding affinity is a close interplay between both actors.

Crystallography is picking up statistical modeling methods too. For example, Habershon used a neural network to predict unit cell parameters from a powder diffraction patterns (33), using a pattern recognition approach. Willighagen used a novel representation for molecular crystal structures and hierarchical clustering methods to classify experimental and simulated

Downloaded At: 12:17 17 January 2011

crystal structures (34). Wehrens used powder diffraction patterns to classify molecular crystal structures using a self-organizing map (35).

Molecular reactions are another area where data analysis and modeling is taking off. While rule based analysis and classification of reactions has been around for quite some time (see e.g. (4)), the use of numerical representation in computational classification and modeling of reactions took off in the nineties, when Chen and Gasteiger used neural networks to classify a number of organic reaction types (36). A year later they published a method that used a self organizing map to classify organic reactions (37), where reactions were represented by a set of physical parameters for common atom in the reaction center on the reactant as well as the product side. The system was unable to classify reactions sets which did not have an atom in common. Zhang and Aires-de-Sousa addressed this problem by using a second self-organizing map, to map the reactant and product sides onto a fixed length representation (38). The reaction itself can then be represented as the difference vector of the reactant and product sides. Using this approach, they were able to classify the metabolic reactions on a genome scale (39) and match this with the empirical EC numbering scheme (40).

As more and more experimental data becomes available, data analysis and modeling will become more important in a whole range of new scientific fields. The few mentioned in this section are just examples of what we can expect in the coming years.

CHEMICAL SPACE, SIMILARITY AND DIVERSITY

Chemical space is the term used to indicate the set of all possible connection tables, given a molecular formula (41). The more atoms in this formula, the larger the number of possible connection tables. It is generally impossible to count the actual number of possible isomers, though attempts have been made to enumerate subsets of chemical space. Applications of the chemical space concept are found in many parts of molecular chemometrics, such as clustering of molecules, diversity analyses, subset selection and structure enumeration.

Especially in structure-activity modeling it is a well-established assumption that structurally similar compounds are likely to exhibit similar properties (42). However, biochemical activity does not just depend on the molecular structures, e.g., acknowledged in the earlier discussed proteochemometrics, and structures can bind in different ways to binding sites. The similarity paradox states that small changes in the structure can lead to large difference in activity (43). The study of structural similarity is an extensive field; this review will not give a full view on this subject, and readers are recommended to read one of the comprehensive reviews available in literature, e.g., by Nikolova (43), Bender (42) or Maldonado (44). Instead, it will highlight the function of similarity measures and show interesting developments in this field.

A similarity measure, the quantifier of similarity, is made up of two components: a representation of the relevant molecular information, discussed earlier, and an index or coefficient suitable for this representation. Well-known similarity measures include the Euclidean distance for continuous-valued representations, and the Tanimoto coefficient for binary representations, such as fingerprints. An example which shows that a proper coefficient should be used, is the use of the weighted cross correlation when comparing crystal structures on the basis of an electronic radial distribution function (34). The representation resembles a peak-like spectrum in which small peak shift indicates a small structural change; a Euclidean distance measure would fail to properly describe the structural similarity. The same problem is encountered when crystal structures are represented by their powder diffraction pattern (45).

An important application of similarity is diversity analysis. Especially in library design and subset selection, diversity is regarded an important feature, for similar reasons as those in experimental design. The goal of library design is to set up a library of molecular structures with the highest possible diversity, to achieve the largest coverage of chemical space. Another application is subset selection which can be used to define independent test sets in activity and property modeling. Again, coverage of chemical space is the goal. The analogy with experimental design is confirmed by the overlap in methods used; for example, D-optimal designs have also been used for subset selection (46). Olsson introduced an improvement on this design, which addresses the occasional redundancy and replication (47).

Though several diversity and similarity measures have been developed in the past, the applicability all depends on the descriptors used, and the chemical problem for which they are used. For example, one can use the similarity in chemical space, but for biochemical activities this might not be the right measure. The same holds for diversity, and both will continue to evolve together with the use of new descriptors and the use in new fields of research.

ACTIVITY AND PROPERTY MODELING

Although the idea of relating physical properties to molecular structures dates from the 19th century (44), the first mathematical model was developed by Wiener for boiling points of paraffins (48) only in 1947. Hansch was the first to model a biological activity, when he correlated toxicity of benzoic acids to their structures (49), 17 years later. Modeling physical properties and biochemical activities is still a topic that receives a lot of attention. As discussed previously, the search for new descriptors is still ongoing, as is the search for new modeling methods. Common methods used include multilinear regression (MLR), principle component regression (PCR), partial least squares (PLS) and neural networks (NN) for regression, and k-Nearest Neighbors (kNN), classification and regression trees (CART), linear and quadratic discriminant analysis (LDA, QDA) and soft independent modeling of class analogy (SIMCA). Regression methods are also used for supervised classification. Often, these methods are combined with feature selection methods, discussed later.

A method that has received growing attention is Support Vector Machines (SVM), originally developed by Vapnik (50). Two

Downloaded At: 12:17 17 January 2011

types of SVM's have been developed: one that finds a hyper plane that separates two classes; and another one for regression, often referred to as SVR. While this hyper plane is linear in itself, the hyper plane can be sought in a space of higher dimension than the original data. The transformation of the data into this high-dimensional space is, and that is the elegance of the SVM method, equivalent to a formulation involving a so-called kernel function. Using such kernel functions makes SVM able to fit nonlinear behavior. Note that the use is not restricted to SVM, and can be used with partial least squares (PLS), too (51). While most SVM applications use the radial distribution function (RBF), other kernels are available too, like the polynomial, ANOVA (52) and Pearson IV (53) kernel. The latter is attractive as it can mimic both the RBF and the polynomial kernel. Any function that yields semi-definite kernel matrices can be used as kernel in SVM, allowing the use of chemoinformatics-specific kernels. For example, Lind et al. used a kernel based on the Tanimoto distance measure (54).

The number of articles that use SVM or SVR in QSAR and QSPR is steadily growing; this review cites a selection of the earlier studies. Serra found that SVM performed better than k-NN in classifying molecules according to their clastogenic behavior (55). Byvatov compared SVM with neural networks in a drug/non-drug classification problem and found that SVM was slightly better (56). Because SVM defines one hyper plane, it can only classify two classes. A common approach for dealing with more than two classes is to make an one-against-all model for each class. For example, Ivanciuc used SVM classification for a three class odor problem (52). Support Vector Regression (SVR) is an adaption of the SVM algorithm (57), and allows making regression models, where the hyper plane regresses through the data points. Burbidge was one of the first to apply SVR on structure-activity relationship data (58), and compared the performance of the method with C5.0 decision trees and neural networks, and found that SVM performed best. Bennett used SVR to model the retention times of proteins on an anion-exchange chromatography system (59).

Classification of molecules into two categories can also be performed by a method called substructure mining. The method uses subgraph searching to find molecular fragments that are specific for one of the classes. An important feature of these methods is that the resulting model is easily interpreted; substructures can directly be related to the modeled end point. Since the number of possible substructures is enormous, these graph mining methods start from the data set itself and only consider substructures found. Kazius used this approach to predict mutagenicity (60), and Borgelt developed an algorithm that performs such a search to predict anti-HIV activity (61). To reduce the number of substructures even further, only linear substructures, paths, may be considered, considerably speeding up the analysis (62).

While a lot of modeling methods have been tried in the past, it generally is still difficult to capture certain features of the data to model, including non-linearities and different modes of actions of the molecules itself. While the former can be addressed by using non-linear methods, or non-linear kernels as, for example, used in combination with PLS and SVR, the latter can be addressed by making local or sub models. Making physically and (bio-)chemically relevant local models, explaining different modes of action, is one of the challenges we face in the next years.

Dimension Reduction

Calculating hundreds, if not thousands, of descriptors has become feasible with the modern computing power, and the general lack of understanding which molecular information is important, makes feature selection a continuing challenge. Feature selection, or variable selection, is a popular way to reduce the number of variables to be used in a model and is an alternative to, for example, PCA where linear combinations of variables are sought to describe the data efficiently. Feature selection has the advantage that the selected features are easier to interpret than linear combinations. Selections can be made such that the variables are orthogonal, or such that they contain most additional information content, e.g., calculated using the Shannon entropy. More importantly, the number of possible selections increases more than exponential with the number of variables to choose from.

Feature selection is, in essence, an optimization problem in which the goal is to find a subset of features, or variables, the give the best performance, e.g., for building QSAR models. Reasons to do this include model interpretability and reducing chance correlation. Classical methods include forward selection, in which is started with zero variables, and the one variable gets added that improves the model performance most. Likewise, in backwards elimination one starts with all variables, and the one variable get deleted that reduces the model performance least. To reduce ending up in local minima, the stepwise method can be used, which starts by forward selection, but allows elimination of earlier added variables after each addition. However, these methods often end up in local optimal.

Because feature selection is, in essence, an optimization problem, global optimization methods can also be used. For example, genetic algorithms (GAs) have been used a lot for this purpose (63). Xu compared GAs with classical variable selection methods and found that the former performed better than the classical methods (64). Other optimization methods used for feature selection include tabu search (65) and simulated annealing (66).

Recently, other optimization methods have been applied too, including ant colony optimization (67), and particle swarm optimization (68). Like genetic algorithms, these methods evaluate variable subsets by making a new regression or classification model, using a prediction error measure, as discussed later. Alternatively, Byvatov used SVM to calculate the importance of features based on the support vectors, where features with a low importance were removed (69).

Downloaded At: 12:17 17 January 2011

Downloaded At: 12:17 17 January 2011

Model Validation

With a growing number of descriptors, modeling methods and feature selection methods, statistically sound performance estimation of classification or regression models is crucial. This section discusses new insight and validation approaches from recent literature. Modeling methods are designed to make the best fit, and are not concerned with underlying physical and chemical principles, nor do they care about the so-called combinatorial explosion with a growing number of dependent variables. Consequently, overfitting is a serious risk (70). Leave-one-out (LOO) and leave-more-out (LMO) cross-validation, also called k-fold cross-validation, have become increasingly popular. Baumann noted that while LOO performs well when selecting among a few alternatives, it yields overfitted models when used for feature selection (71). In all cases, an independent hold-out test set, not used in any step of the training process, should be used to estimate the final performance of the model, though it is noted that for small data sets one loses predictive power (72): Hawkins studied the behavior of the cross-validation q^2 and the R^2 for the independent test set, and proposed that with 100 or less compounds in the data set, only cross-validation should be used. Golbraikh further discusses the use of q^2, and argues that this statistic shows little correlation with predictive power (73).

Cross-validation is not the only available method, and others include y-randomization (74), and bootstrapping (75). Mevik compared several prediction error estimators, among which were LOO cross-validation, k-fold cross-validation, and three bootstrap based methods, for situations where the number of variables exceeds the number of objects (76). Though differences are small, he recommends LOO cross-validation or the 0.632 bootstrap estimate, unless computational demand is too large, in which case the LMO is a viable alternative. Several groups have worked on general guidelines for building QSAR and QSPR models, often consisting of a combination of a few performance statistics, like those mentioned here. The reader is pointed to articles by Todeschini (77), Eriksson (78), and Tropsha (79).

Statistical and machine learning modeling methods do not try to understand underlying physical and (bio-)chemical concepts; instead, they try to make the best fit between the sets of data, often molecular structures and some property. With large numbers of variables to describe molecular information, the chance to find a combination of them that correlates with the modeled activity explodes, even if the are unrelated to the physical and chemical concepts. This danger is addressed by using cross-validation and test sets, and generally using data sets with more objects, which is becoming feasible with high-throughput experiments. Nevertheless, making scientifically sound and interpretable models is still an exciting challenge.

LIBRARY SEARCHING

Lavine identified Library Searching as one of the key areas of chemometrics in 1998 (80), though the topic did not return in his later reviews (81–83). Library searching is finding information in one or more libraries of data; with respect to molecular chemometrics, these libraries contain molecular information, such geometrical structures, spectra and physical and biochemical properties. Any of these can occur in literature, and sets of articles and abstracts are explicitly considered an (electronic) library in this review. To a large extent, these libraries are strictly formatted, for example, using relational databases from which extraction is easy. However, with the growing amount of diverse data produced in experimental work, a growing interest in sharing data on the Internet, and the trend towards a Semantic Web, data retrieval has become increasingly important.

Berners-Lee envisioned the Semantic Web in 2002, a future where information on the Internet is machine-readable, that is, where the information has semantic meaning (84, 85). For example, a client program would not just be able to retrieve safety information on a molecule, but it could also give suggestions where the compound could be bought, what biological processes it is involving, how it could be synthesized, etc. While most of this information is already available on the web, such client software is currently not generally available. Returning to our molecule query, the following problems exist: data bases do not use one unique identifier for a particular molecule; chemical information is not stored in a well documented format; information does not have clear semantic meaning; information is not freely available (86, 87).

The last problem is slowly being addressed by a growing number of open access databases (see Table 1). The information available from these data bases is diverse, and includes crystal structures, biological activities and binding information, metabolic relations, NMR spectra and reaction mechanisms of enzymatic reactions.

It is noteworthy that although these data bases are open access, not all of them allow the content to be replicated, modified and redistributed, like in open source software, but it shows at least a new trend compared with previous decades where chemical database were mostly proprietary and expensive. Library searching is, obviously, not restricted to open access data bases, but is applicable to proprietary data bases too (an overview of those is found in (4)). A bigger challenge is the lack of a uniform access to both types of data bases. Access is not always available

TABLE 1
Some examples of open access databases with molecular information

ChemDB (88)	http://cdb.ics.uci.edu/
KEGG (89)	http://www.genome.jp/kegg/
Ligand.info (90)	http://ligand.info/
MACiE (91)	http://www-mitchell.ch.cam.ac.uk/macie/
NMRShiftDB (92)	http://nmrshiftdb.org/
PubChem (93)	http://pubchem.ncbi.nlm.nih.gov/
RCSB PDB (94)	http://www.pdb.org/
ZINC (95)	http://blaster.docking.org/zinc/

other than via a web interface or a custom program, making it difficult for machines to retrieve information. The use of semantic markup languages, mostly using the XML syntax, should change this. For molecular information the Chemical Markup Language (96) is receiving growing interest. For example, it is used to distribute physical properties of isotopes and elements (87), storage of reaction mechanisms (97), and enrichment of blogs and news feeds with chemical content (98).

Instead of making access to the data uniform, using web services and XML languages, one could also take another approach: trying to write a computer parser algorithm that takes unformatted documents, and to extract information from text, tables and figures. Given the huge amount of electronic journal articles in PDF format available now, this, though difficult, might proof very fruitful (99). Townsend used this approach, and developed a system that uses regular expressions to extract information from experimental sections of articles (100). Karthikeyan developed a system for finding chemical information on the Internet (101), also using regular expressions.

Finding information on an individual molecular structure has become easier too, with the publication of the InChI (102). This unique molecular identifier will likely have an important function in the chemical semantic web (85). Because most of current literature does not use such a unique index, one has to rely in IUPAC names, trivial names, and other naming schemes (99), and finding literature related to a query compound on just such names is not optimal. Singh uses both textual as well as molecular descriptors to address this problem, and defined a similarity measure between the query molecule and articles based on both pieces of information (103).

Downloaded At: 12:17 17 January 2011

CONCLUSION

The research field of molecular chemometrics shows overlap with chemoinformatics, pharmaceutical studies, chemometrics and bioinformatics. Literature is scattered over a number of journals and the number of books in this area is also increasing. This review gives a view on the current trends in this field, and only a glimpse of the literature published in the past few years. The trends include the ongoing search for new ways to describe molecular structures, and, in a growing amount, molecules in some environment, for new multivariate modeling methods, and for new methods to deal with the ever growing amount of data in databases and on the Internet.

Though many of the problems have been addressed in literature, several important ones are still standing. For example, molecular chemometrics has to deal with an increasing amount of data to be analyzed and modeled. With the size of chemical space in mind, one cannot anticipate this amount to level off soon. The increase in computing power will not come close to what is needed. Consequently, more powerful searching, mining, feature selection, modeling and validation methods will become increasingly important.

REFERENCES

1. R. S. Bohacek, C. McMartin, and W. C. Guida, The art and practice of structure based drug design: A molecular modeling perspective. *Medicinal Research Reviews* 16 (1996):3–50.
2. H. Kubinyi and G. Müller, *Chemogenomics in Drug Discovery*, volume 22 of Methods and Principles in Medicinal Chemistry (Weinheim: Wiley-VCH, 2004).
3. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, volume 11 of Methods and Principles in Medicinal Chemistry (Wiley-VCH, New York, 2000).
4. J. Gasteiger, editor. *Handbook of Chemoinformatics* (Wiley-VCS, Weinheim, 2003).
5. W. Brock. *The Fontana History of Chemistry* (Fontana Press, London, 1992).
6. A. Dietz. Yet another representation of molecular structure. *Journal of Chemical Information and Computer Sciences* 35 (1995):787–802.
7. S. Bauerschmidt and J. Gasteiger, Overcoming the limitations of a connection table description: A universal representation of chemical species. *Journal of Chemical Information and Computer Sciences* 37 (1997):705–714.
8. K. Baumann, Uniform-length molecular descriptors for quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR): classication studies and similarity searching. *Trends in Analytical Chemistry* 18 (1999):36–46.
9. D. Livingstone, The characterization of chemical structures using molecular properties. A survey. *Journal of Chemical Information and Computer Sciences* 40 (2000):195–209.
10. J. Duca and A. Hopfinger, Estimation of molecular similarity based on 4D-QSAR analysis: Formalism and validation. *Journal of Chemical Information and Modeling* 41 (2001):1367–1387.
11. R. Cramer III, D. Patterson, and J. Bunce, Comparitative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* 110 (1988):5959–5967.
12. K. Kim, List of CoMFA references, 1997. *Perspectives in Drug Discovery and Design*, 12–14 (1998):334–338.
13. K. Kim, G. Greco, and E. Novellino, A critical review of recent CoMFA applications. *Perspectives in Drug Discovery and Design* 12–14 (1998):257–315.
14. A. Hopfinger, S. Wang, J. Tokarski, B. Jin, M. Albuquerque, P. Madhav, and C. Duraiswami, Construction of 3D-QSAR Models using the 4D-QSAR analysis formalism. *Journal of the American Chemical Society* 119 (1997):10509–10524.
15. C. L. Senese, J. Duca, D. Pan, A. J. Hopfinger, and Y. J. Tseng, 4D-fingerprints, universal QSAR and QSPR descriptors. *Journal of Chemical Information and Computer Sciences* 44 (2004):1526–1539.
16. R. Bursi, Y. Dao, T. Van Wijk, M. De Gooyer, E. Kellenbach, and P. Verwer, Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *Journal of Chemical Information and Computer Sciences* 39 (1999):861–867.
17. R. Begigni, L. Passerini, D. Livingstone, M. Johnson, and A. Giuliani, Infrared spectra information and their correlation with QSAR descriptors. *Journal of Chemical Information and Computer Sciences* 39 (1999):558–562.

18. R. Beger, J. Freeman, J. Lay Jr., J. Wilkes, and D. Miller, Use of 13 C NMR spectrometric data to produce a predictive model of estrogen receptor binding activity. *Journal of Chemical Information and Computer Sciences* 41 (2001):219–224.
19. A. Asikainen, J. Ruuskanen, and K. Tuppurainen, Spectroscopic QSAR methods and self-organizing molecular field analysis for relating molecular structure and estrogenic activity. *Journal of Chemical Information and Computer Sciences* 43 (2003):1974–1981.
20. N. Bailey, Y. Wang, J. Sampson, W. Davis, I. Whitcombe, P. Hylands, S. Croft, and E. Holmes, Prediction of anti-plasmodial activity of Artemisia annua extracts: Application of ^{1}H NMR spectroscopy and chemometrics. *Journal of Pharmaceutical and Biomedical Analysis* 41 (2001):219–224.
21. S. Vanderhoeven, J. Troke, G. Tranter, I. Wilson, J. Nicholson, and J. Lindon, Nuclear magnetic resonance (NMR) and quantitative structure-activity relationship (QSAR) studies on the transaceylation reactivity of model 1-O-acyl glucuronides. II: QSAR modelling of the reaction using both computational and experimental NMR parameters. *Xenobiotica* 34 (2004):889–900.
22. P. Khadikar, V. Sharma, and R. Varma, Novel estimation of lipophilicity using ^{13}C NMR chemical shifts as molecular descriptor. *Bioorganic and Medicinal Chemistry Letters* 15 (2005):421–425.
23. L. K. Schnackenberg and R. D. Beger, Whole-molecule calculation of log p based on molar volume, hydrogen bonds, and simulated ^{13}C NMR spectra. *Journal of Chemical Information and Modeling* 45 (2005):360–365.
24. E. Willighagen, H. Denissen, R. Wehrens, and L. Buydens, On the use of ^{1}H and ^{13}C NMR spectra as QSPR descriptors. *Journal of Chemical Information and Computer Sciences* 46 (2006):487–494.
25. J.-L. Faulon, D. P. Visco, and R. S. Pophale, The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *Journal of Chemical Information and Computer Sciences* 43 (2003):707–720.
26. W. Bremser, HOSE—A novel substructure code. *Analytica Chimica Acta* 103 (1978):355–365.
27. M. Randi and S. C. Basak, A new descriptor for structure-property and structure-activity correlations. *Journal of Chemical Information and Computer Sciences* 41 (2001):650–656.
28. M. L. Mansfield, D. G. Covell, and R. L. Jernigan, A new class of molecular shape descriptors. 1. Theory and properties. *Journal of Chemical Information and Computer Sciences* 42 (2002):259–273.
29. K. Tuppurainen and J. Ruuskanen, Electronic eigenvalue (EEVA): a new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital energies. A QSAR approach to the Ah receptor binding affinity of polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs). *Chemosphere* 41 (2000):843–848.
30. N. Stiefl and K. Baumann, Mapping property distributions of molecular surfaces: Algorithm and evaluation of a novel 3D quantitative structure-activity relationship technique. *Journal of Medicinal Chemistry* 46 (2003):1390–1407.
31. M. Lapinsh, P. Prusis, A. Gutcaits, T. Lundstedt, and J. E. Wikberg. Development of proteo-chemometrics: A novel technology for the analysis of drug-receptor interactions. *Biochimica Biophysica Acta* 1525 (2001):180–190.
32. P. Prusis, R. Muceniece, P. Andersson, C. Post, T. Lundstedt, and J. E. Wikberg, PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions. *Biochimica Biophysica Acta* 1544 (2001):350–357.
33. S. Habershon, E. Cheung, K. Harris, and R. Johnston, Powder Diffraction Indexing as a pattern recognition problem: a new approach for unit cell determinination based on an artificial neural network. *Journal of Physical Chemistry A* 108 (2004):711–716.
34. E. L. Willighagen, R. Wehrens, P. Verwer, R. de Gelder, and L. M. C. Buydens, Method for the computational comparison of crystal structures. *Acta Crystallographica* B61 (2005):29–36.
35. R. Wehrens, W. Melssen, L. Buydens, and R. De Gelder, Representing structural databases in a self-organizing map. *Acta Crystallographica* B61 (2005):548–557.
36. L. Chen and J. Gasteiger, Organic reactions classified by neural networks: Michael Additions, Friedel-Crafts alkylations by alkenes, and related reactions. *Angewandte Chemie International Edition in English* 35 (1996):763–765.
37. L. Chen and J. Gasteiger, Knowledge discovery in reaction databases: Landscaping organic reactions by a self-organizing neural network. *Journal of the American Chemical Society* 119 (1997):4033–4042.
38. Q.-Y. Zhang and J. A. de Sousa, Structure-based classification of chemical reactions without assignment of reaction centers. *Journal of Chemical Information and Modeling* 45 (2005):1775–1783.
39. D. Latino and J. Aires-de Sousa, Genome-scale classification of metabolic reactions: A chemoinformatics approach. *Angewandte Chemie International Edition in English* 45 (2006):2066–2069.
40. K. Tipton and S. Boyce, History of the enzyme nomenclature system. *Bioinformatics* 16 (2000):34–40.
41. C. M. Dobson, Chemical space and biology. *Nature*, 432 (2004):824–828.
42. A. Bender and R. C. Glen, Molecular similarity: A key technique in molecular informatics. *Organic and Biomolecular Chemistry* 2 (2004):3204–3218.
43. N. Nikolova and J. Jaworska, Approaches to measure chemical similarity—A review. *QSAR & Combinatorial Science* 22 (2003):1006–1026.
44. A. G. Maldonado, J. P. Doucet, M. Petitjean, and B.-T. Fan, Molecular similarity and diversity in chemoinformatics: From theory to applications. *Molecular Diversity* 10 (2006):39–79.
45. R. De Gelder, R. Wehrens, and J. Hageman, A generalized expression for the similarity spectra: Application to powder diffraction pattern classification. *Journal of Computational Chemistry*, 22 (2001):273–289.
46. P. Gramatica, P. Pilutti, and E. Papa, Validated QSAR prediction of OH tropospheric degradation of VOCs: Splitting into training-test sets and consensus modeling. *Journal of Chemical Information and Computer Sciences* 44 (2004):1794–1802.
47. I.-M. Olsson, J. Gottfries, and S. Wold, Controlling coverage of D-optimal onion designs and selections. *Journal of Chemometrics* 18 (2004):548–557.
48. H. Wiener, Structural determination of paraffin boiling points. *Journal of the American Chemical Society* 69 (1947):17–20

Downloaded At: 12:17 17 January 2011

49. C. Hansch and T. Fujita, Analysis—A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society* 86 (1964):1616–1626.
50. C. Cortes and V. Vapnik, Support-vector networks. *Machine Learning* 20 (1995):273–297.
51. B. Walczak and D. L. Massart, The radial basis functions—Partial least squares approach as a flexible non-linear regression technique. *Analytica Chimica Acta* 331 (1996):177–185.
52. O. Ivanciuc, Structure-odor relationships for pyrazines with support vector machines. *Internet Electronic Journal of Molecular Design* 1 (2002):269–284.
53. B. Üstün, W. Melssen, and L. Buydens, Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems* 81 (2006):29–40.
54. T. P. Lind, Support vector machines for the estimation of aqueous solubility. *Journal of Chemical Information and Computer Sciences* 43 (2003):1855–1859.
55. P. J. R. Serra and E. D. Thompson, Development of binary classification of structural chromosome aberrations for a diverse set of organic compounds from molecular structure. *Chemical Research in Toxicology* 16 (2003):153–163.
56. E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences* 43 (2003):1882–1889.
57. V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, Berlin, 1995).
58. R. Burbidge, M. Trotter, B. Buxton, and S. Holden, Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Computers and Chemistry* 26 (2001):5–14.
59. M. Song, C. Breneman, J. Bi, N. Sukumar, B. Bennett, S. Cramer, and N. Tugcu, Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *Journal of Chemical Information and Computer Sciences* 42 (2002):1347–1357.
60. J. Kazius, R. McGuire, and R. Bursi, Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry* 48 (2005):312–320.
61. C. Borgelt, T. Meinl, and M. Berthold, MoSS: A program for molecular substructure mining. In B. Goethals, S. Nijssen, and M. J. Zaki, ed., *Proceedings of OSDM 2005*, pages 6–15, 2005.
62. C. Helma, T. Cramer, S. Kramer, and L. D. Raedt, Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *Journal of Chemical Information and Computer Sciences* 44 (2004):1402–1411.
63. R. Leardi, Genetic algorithms in chemometrics and chemistry: A review. *Journal of Chemometrics* 15 (2001):559–569.
64. L. Xu and W.-J. Zhang, Comparison of different methods for variable selection. *Analytica Chimica Acta* 446 (2001):477–483.
65. J. Hageman, M. Streppel, R. Wehrens, and L. Buydens, Wavelength Selection with Tabu Search. *Journal of Chemometrics* 17 (2003):427–437.
66. P. Jurs and J. Sutter, Adaption of simulated annealing to chemical optimization problems, volume 15 of *Data Handling in Science and Technology*, chapter Selection of molecular descriptors for quantitative structure-activity relationships (Elsevier, Amsterdam, 1995).
67. Q. Shen, J.-H. Jiang, J.-C. Tao, G.-L. Shen, and R.-Q. Yu, Modified ant colony optimization algorithm for variable selection in QSAR modeling: QSAR studies of cyclooxygenase inhibitors. *Journal of Chemical Information and Modeling* 45 (2005):1024–1029.
68. Q. Shen, J.-H. Jiang, C.-X. Jiao, G.-L. Shen, and R.-Q. Yu, Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. *European Journal of Pharmaceutical Sciences* 22 (2004):145–152.
69. E. Byvatov and G. Schneider, SVM-based feature selection for characterization of focused compound collections. *Journal of Chemical Information and Computer Sciences* 44 (2004):993–999.
70. D. M. Hawkins, The problem of overfitting. *Journal of Chemical Information and Computer Sciences* 44 (2004):1–12.
71. K. Baumann, H. Albert, and M. Von Korff, A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Practical applications. *Journal of Chemometrics* 16 (2002):351–360.
72. D. Hawkins, S. Basak, and D. Mills, Accessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences* 43 (2003):579–586.
73. A. Golbraikh and A. Tropsha, Beware of q^2! *Journal of Molecular Graphics and Modelling* 20 (2002):269–276
74. S. Wold and L. Eriksson, Statistical validation of QSAR results, *Chemometrics Methods in Molecular Design* (VCH, Weinheim, Germany), 1995 309–318.
75. R. Wehrens and W. Van der Linden, Bootstrapping principal-component regression models. *Journal of Chemometrics* 11 (1997):157–171.
76. B.-H. Mevik and H. Cederkvist, Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics* 18 (2004):422–429.
77. R. Todeschini, V. Consonni, A. Mauri, and M. Pavan. Detecting "bad" regression models: Multicriteria fitness functions in regression analysis. *Analytica Chimica Acta* 515 (2003):199–208.
78. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, and P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives* 111 (2003):1361–1375.
79. A. Tropsha, P. Gramatica, and V. K. Gombar, The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science* 22 (2003):69–77.
80. B. Lavine, Chemometrics. *Analytical Chemistry* 70 (1998):209–228.
81. B. Lavine, Chemometrics. *Analytical Chemistry* 72 (2000):91–98.
82. B. Lavine and J. Workman, Chemometrtics. *Analytical Chemistry* 74 (2002):2763–2770.
83. B. Lavine and J. Workman, Chemometrics. *Analytical Chemistry* 76 (2004):3365–3372.

Downloaded At: 12:17 17 January 2011

84. T. Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web. *Scientific American* (2001):28–37.
85. S. J. Coles, N. E. Day, P. Murray-Rust, H. S. Rzepa, and Y. Zhang, Enhancement of the chemical semantic web through the use of InChI identifiers. *Organic and Biomolecular Chemistry* 3 (2005):1832–1834.
86. P. Murray-Rust, H. S. Rzepa, S. M. Tyrrell, and Y. Zhang, Representation and use of chemistry in the global electronic age. *Organic & Biomolecular Chemistry* 2 (2004):3192–3203.
87. R. Guha, M. Howard, G. Hutchison, P. Murray-Rust, R. Rzepa, S. Steinbeck, J. Wegner, and E. Willighagen, The blue obelisk-interoperability in chemical informatics. *Journal of Chemical Information and Modeling* 46 (2006):991–998.
88. J. Chen, S. J. Swamidass, Y. Dou, J. Bruand, and P. Baldi, ChemDB: A public database of small molecules and related chemoinformatics resources. *Bioinformatics* 21 (2005):4133–4139.
89. M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, The KEGG databases at GenomeNet. *Nucleic Acids Research* 30 (2002):42–46.
90. M. von Grotthuss, G. Koczyk, J. Pas, L. S. Wyrwicz, and L. Rychlewski, Ligand.Info small-molecule Meta-Database. *Combinatorial Chemistry and High Throughput Screening* 7 (2004):757–761.
91. G. L. Holliday, G. J. Bartlett, D. E. Almonacid, N. M. O'Boyle, P. Murray-Rust, J. M. Thornton, and J. B. O. Mitchell, MACiE: a database of enzyme reaction mechanisms. *Bioinformatics* 21 (2005):4315–4316.
92. C. Steinbeck, S. Kuhn, and S. Krause, NMRShiftDB—Constructing a chemical information system with open source components. *Journal of Chemical Information and Computer Sciences* 43 (2003):1733–1739.
93. C. P. Austin, L. S. Brady, T. R. Insel, and F. S. Collins, NIH Molecular Libraries Initiative. *Science* 306 (2004):1138–1139.
94. H. Berman, K. Henrick, and H. Nakamura, Announcing the worldwide Protein Data Bank. *Nature Structural Biology* 10 (2003):980.
95. J. Irwin and B. Shoichet, ZINC—A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* 45 (2005):177–182.
96. P. Murray-Rust and H. Rzepa. Chemical Markup XML, and the Worldwide Web. 1. Basic principles. *Journal of Chemical Information and Computer Sciences* 39 (1999):928–942.
97. G. L. Holliday, P. Murray-Rust, and H. S. Rzepa, Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for chemical reactions. *Journal of Chemical Information and Modeling* 46 (2006):145–157.
98. P. Murray-Rust, H. S. Rzepa, M. J. Williamson, and E. L. Willighagen, Chemical markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators. *Journal of Chemical Information and Computer Sciences* 44 (2004):462–469.
99. D. L. Banville, Mining chemical structural information from the drug literature. *Drug Discovery Today* 11 (2006):35–42.
100. J. A. Townsend, S. E. Adams, C. A. Waudby, V. K. de Souza, J. M. Goodman, and P. Murray-Rust, Chemical documents: Machine understanding and automated information extraction. *Organic & Biomolecular Chemistry* 2 (2004):3294–3300.
101. M. Karthikeyan, S. Krishnan, and A. Pandey, Harvesting chemical information from the internet using a distributed approach: ChemXtreme. *Journal of Chemical Information and Modeling* 46 (2006):452–461.
102. S. Stein, S. Heller, and D. Tchekhovski, An open standard for chemical structure representation —The IUPAC Chemical Identier. In *Nimes International Chemical Information Conference Proceedings*, pages 131–143, 2003.
103. S. B. Singh, R. D. Hull, and E. M. Fluder, Text Influenced Molecular Indexing TIMI): A literature database mining approach that handles text and chemistry. *Journal of Chemical Information and Computer Sciences* 43 (2003):743–752.

Downloaded At: 12:17 17 January 2011