Bioorganic & Medicinal Chemistry 20 (2012) 5317–5323

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Review

Chemoinformatics: A view of the field and current trends in method development

Martin Vogt, Jürgen Bajorath *

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

ARTICLE INFO

Article history:
Available online 23 March 2012

Keywords:
Chemoinformatics
Computational methods
Algorithms
Molecular similarity
Molecular representations
Machine learning
SAR data mining
Computational efficiency
Drug discovery

ABSTRACT

The chemoinformatics field continues to evolve at the interface between computer science and chemistry. Chemical information and computational approaches in pharmaceutical research are major focal points of chemoinformatics. However, the boundaries of this discipline are rather fluid and the chemoinformatics spectrum is difficult to delineate. The field is in flux, which also provides opportunities for further developments. As a lead-in to the Chemoinformatics Symposium-in-Print, we present a brief view of this discipline (with a little anecdotal touch), highlight current trends in method development, and discuss a number of representative examples.

Contents

1. Introduction . . . 5317
2. Terminology . . . 5318
3. General trends . . . 5318
4. Design of molecular fingerprints . . . 5318
5. Similarity measures . . . 5318
6. Clustering . . . 5320
7. Support vector machines . . . 5320
8. Other trends in machine learning . . . 5320
9. SAR elucidation and visualization . . . 5321
10. Statistical approaches . . . 5321
11. New methodologies . . . 5321
12. Computational efficiency . . . 5322
13. Concluding remarks . . . 5322
References and notes . . . 5322

1. Introduction

Providing a detailed description (or definition) of chemoinformatics as a scientific discipline is far from being a trivial task. No one would question that chemical data processing, mining, and archiving as well as the design, implementation, and integration of chemical information systems belong to the spectrum of chemoinformatics approaches. However, so do development and

Abbreviations: MCS, maximum common subgraph; MMP, matched molecular pair; QSAR, quantitative structure–activity relationship; ROC, receiver operating characteristic; SMILES, simplified molecular input line entry specification; SVM, support vector machine.

* Corresponding author. Tel.: +49 228 2699 306; fax: +49 228 2699 341.
E-mail address: bajorath@bit.uni-bonn.de (J. Bajorath).

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmc.2012.03.030

application of computational models for property/bioactivity analysis and prediction. This is where the boundaries of the chemoinformatics field are still rather fluid. There is considerable overlap with drug design and other areas of computational chemistry. It is often not evident where chemoinformatics begins or ends. One would not question either that virtual screening methods are an integral part of chemoinformatics.[1] However, what about QSAR, pharmacophore modeling, or other established computational chemistry or molecular modeling methods that have long been standing on their own? Views might differ here. Since the original definition of 'chemoinformatics' by Frank Brown in 1998, proposing that this emerging discipline should include 'all the information resources that a scientist needs to optimize the properties of a ligand to become a drug',[2] a number of other (often generic) definitions have been put forward. It is also worth noting that the term 'chemical informatics' has already been used much earlier, essentially referring to all applications of information science to chemistry,[3] although an original literature reference for its introduction appears to be lacking.

In 2004, one of us proposed an 'umbrella' function for chemoinformatics to cover a broad spectrum of scientific efforts ranging from chemical data collection and analysis to the exploration of structure–activity relationships and prediction of in vivo compound effects.[3] Hence, on the basis of this proposal, QSAR would be a part of the chemoinformatics spectrum (although this might often be matter of debate) and so would be the application of statistical methods to chemistry (thereby 'invading' the 'chemometrics' territory). The interested reader is referred to a comprehensive description of different QSAR methods[4] and a review of chemometrics.[5] Since the time of our 2004 proposal, the chemoinformatics discipline has done little to sharpen its focus or further (re-)define itself, as nicely illustrated by a very comprehensive recent review.[6] If anything, the coverage is becoming even broader.

2. Terminology

This heterogeneity of the field is also reflected by the absence of a generally accepted name. On an anecdotal note, it has on occasions been vehemently argued by insiders (and probably still is) whether the term 'chemoinformatics' or 'cheminformatics' would be more appropriate. At least, increasing focus can be observed here because the 'name game' is by now essentially reduced to the 'o' issue. There are some apparent regional preferences for using the 'o' or not (e.g., Europe vs the United States). Clear is that the additional vocal in chem(o)informatics is not a phonetic requirement; clear is also that phonetic resemblance might make it smoother to speak about bio- and chemo-informatics in the same context. Be this as it may, in January 2012, Google searches yielded ~456,000 and ~513,000 entries for 'chemoinformatics' and 'cheminformatics', respectively, and hence the 'o' issue might not be settled any time soon (and this is also not too important). On a personal note, we preferentially adhere to the term chemoinformatics to ensure consistency with previous publications (and not for philosophical, let alone scientific reasons).

3. General trends

However one might understand the chemoinformatics field as a whole, there are a few aspects that are quite characteristic of its status quo. First, a key feature of many chemoinformatics investigations is their large-scale character, irrespective of the questions studied and the methods used. This means, for example, that millions of compounds are analyzed and compared (and not small sets) and that large bodies of SAR data are gathered and organized (rather than only data associated with individual series).[7,8] This large-scale character of its applications sets chemoinformatics increasingly apart from other areas of computational chemistry. Second, although chemoinformatics approaches are applicable to many, if not all areas of chemistry, most activities continue to focus on pharmaceutical research and drug discovery-related questions. Hence, analyzing, rationalizing, and predicting biological activity of small molecules continues to be the central theme of chemoinformatics.

It should also be noted that many of the algorithms and methods that are being used in chemoinformatics have originated from computer science. For example, this applies to most machine learning methods that are currently popular in the chemoinformatics field. However, it would certainly be an oversimplification to conclude that chemoinformatics would be firmly rooted in computer science. In fact, in a discipline evolving at the interface between computer/information science and chemistry, we also observe many methodological developments that differ from standard algorithms adapted from computer science. For example, this applies to a number of numerical SAR analysis functions.[4,7] Hence, a high degree of methodological diversity is a characteristic of the chemoinformatics arena.

In the following, we highlight a number of current trends in chemoinformatics method development. Figure 1 illustrates exemplary topics that are, among others, discussed below. We especially (but not exclusively) focus on recent developments reported over the past two to three years and begin the discussion with mainstays of chemoinformatics.

4. Design of molecular fingerprints

Molecular fingerprints have for long been popular descriptors and similarity search tools.[1] However, while fingerprints are widely applied, in recent years, only a limited number of new fingerprint designs have been reported. For some period of time, protein–ligand interaction fingerprints have been intensely investigated,[9] but there have only been few new developments since 2009. Bridging between interaction fingerprints and conventional 2D fingerprints, a fragment-based fingerprint has been introduced that implicitly accounts for ligand–target interactions.[10] This is facilitated by encoding in an atom-centered format those fragments of ligands in complex X-ray structures that directly interact with protein atoms.[10] A few new 2D fingerprints have also been designed including a generally applicable bonded atom pair fingerprint that specifically focuses on short-range atom environment information[11] and thereby conceptually departs from topological fingerprints that systematically detect atom/bond pathways in molecules. Furthermore, a new topological fingerprint design captures all systematically generated subgraphs up to a predefined size and separately encodes connected atoms and bonds.[12] In addition, extended connectivity fingerprints that capture layered atom environments and are among the currently most widely applied fingerprints (partly due to their availability in a popular commercial software package) have now been described in detail.[13]

5. Similarity measures

Considerably more efforts than in fingerprint design have recently been spent to better understand and further improve measures for molecular similarity evaluation in the context of similarity searching where the Tanimoto coefficient continues to be the most popular measure. Considering the ever increasing size of compound databases, there is continuing interest in efficient strategies for database searching. A number of studies have been published that aim at establishing bounds for the Tanimoto coefficient, hence making it possible to efficiently prune dat-

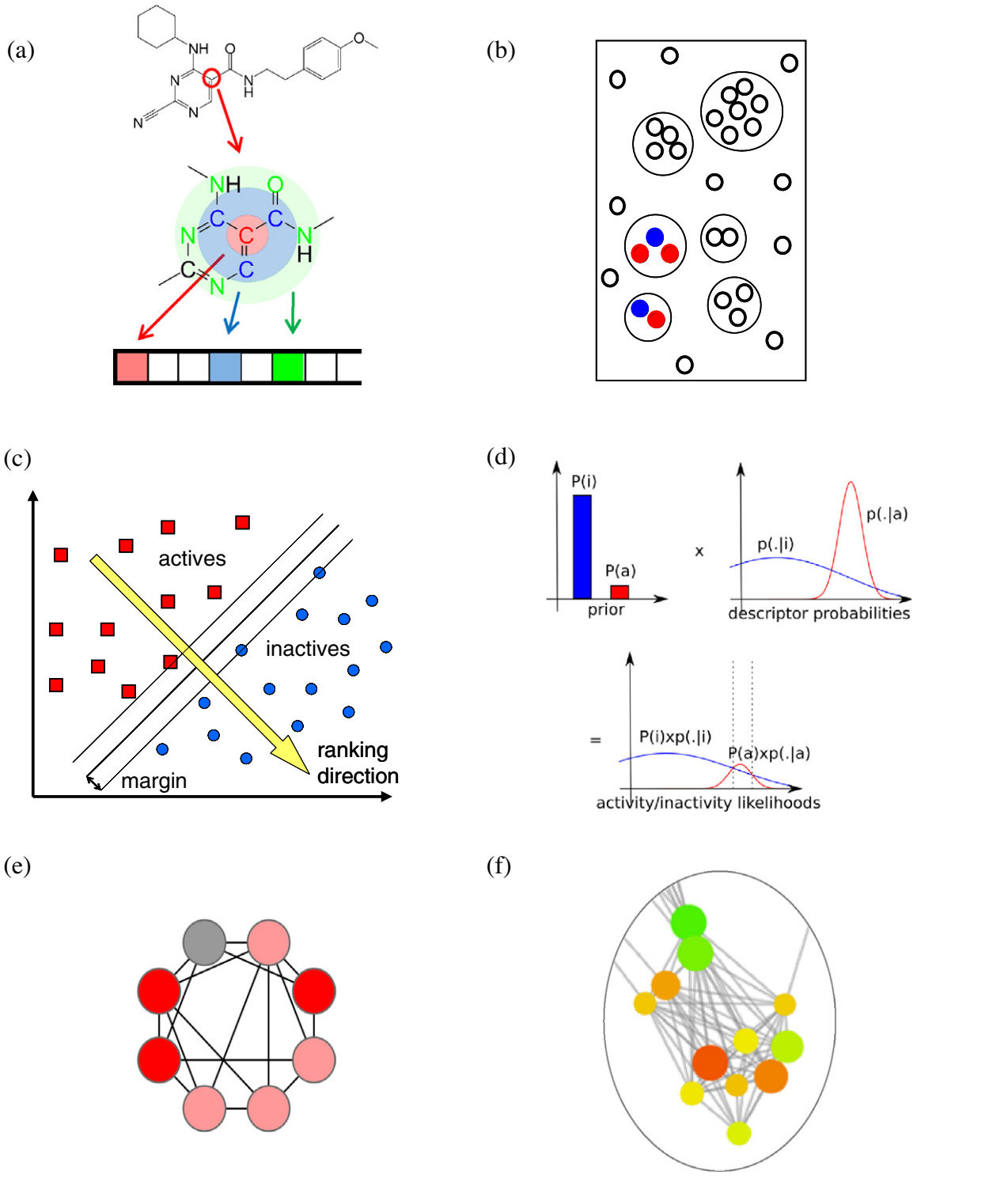

Figure 1. Method development areas. A few current topics of method development in chemoinformatics are schematically illustrated. Selected examples include (a) topological atom environments in small molecules that can be captured in a fingerprint format, compound classification techniques such as (b) advanced clustering methods taking activity information into account or (c) support vector machines, and (d) probabilistic models of biological activity such as Bayesian classifiers that yield likelihoods of activity as function of molecular descriptor settings. In addition, prototypic network representations are shown including (e) a scaffold-based target network (where nodes are targets that are connected by an edge if they share at least one active molecular scaffold) and (b) a similarity-based compound network (nodes represent active compounds and edges similarity relationships) annotated with activity information.

abases such that only a fraction of a database would need to be screened in order to identify molecules that exceed a given similarity threshold.[14–19] Usually, this would require pre-processing of a database. Traditionally, inverted indices have been used for fast structure and substructure searching using binary fingerprints. Recently, this approach has been extended to count or multi-set fingerprints.[20,21] These fingerprint representations were either generated by enumerating atoms, bonds, paths, and substructures[20] or from LINGO[22] profiles on the basis of SMILES subsequences.[21]

The complexity of Tanimoto calculations depends to a large extent on the underlying molecular representation. For instance, the Tanimoto coefficient has been applied to LINGO[22] profiles and the ROCS[23] methodology that compares molecules based on the overlap of their shapes. Especially for ROCS, the calculation of shape overlap of two molecules constitutes a significant bottleneck in similarity searching of large databases. Recently, the SCISSORS[24,25] approach has been introduced to efficiently approximate the Tanimoto coefficient. SCISSORS estimates similarities by representing compounds using a fixed set of basis molecules that are embedded in an n-dimensional Euclidean space. The Tanimoto coefficient of test compounds can then be estimated from the pre-calculated Tanimoto similarity using basis molecules. This reduces Tanimoto calculations to the determination of inner products in a low-dimensional space.

Especially in the context of virtual screening, the distribution of similarity scores across a compound database is of high interest. It has been shown that Tanimoto coefficients can be well approximated as the ratio of two correlated normal value distributions.[26] However, these distributions are highly dependent on the nature of the fingerprint.[26,27] Taking this into account, it has been demonstrated that the distributions of Tanimoto coefficients for a variety of different fingerprint types can be accurately modeled by estimating the feature frequencies and their correlations, hence enabling the accurate prediction of ROC curves for virtual screening.[27]

A different problem in chemoinformatics is scaffold similarity. The analysis and comparison of molecular scaffolds[28] and the assessment of scaffold hopping potential in virtual screening[1] have generally been complicated by the inability to reliably account for differences between scaffolds in a quantitative manner. For quantifying scaffold relationships, the calculation of Tanimoto similarity is often not of sufficiently high resolution. Recently, a function has been introduced to consistently quantify chemical distances between scaffolds (even if these differences are subtle).[29] It has been shown that this scaffold distance function is chemically more intuitive than Tanimoto similarity and suitable to quantitatively assess the degree of difficulty involved in scaffold hopping exercises.[29]

6. Clustering

Cluster analysis is another 'classical' approach in chemoinformatics for which many different algorithms have been adopted or developed over the years. Nevertheless, advanced clustering methods are being investigated. A characteristic feature of several new developments is the combination or integration of different types of approaches with cluster analysis. For example, molecular similarity-based clustering has been integrated with Bayesian models of bioactivity to include activity information in the structural organization of biological screening data.[30] Furthermore, clustering has also been integrated with principal component analysis to aid in the selection of structurally diverse compounds.[31] Also, molecular network analysis has been combined with network clustering to classify and predict active compounds.[32]

The selection of subsets of diverse compounds from large libraries also represents a traditional chemoinformatics task related to clustering. Recently, a method has been introduced for the selection of diverse compounds that combines diversity criteria with a scoring function to select diverse subsets of molecules having a high probability of activity.[33]

7. Support vector machines

One of the current growth areas in chemoinformatics includes the development of support vector machine (SVM)-based classification protocols and design of new kernel functions for SVM applications. In ligand-based virtual screening, SVMs have increasingly become a method of choice[1] over the past few years and this trend continues. In addition, SVMs are very popular for a variety of compound classification tasks and have, for instance, been applied to predict synthetic accessibility of compounds[34] or aqueous solubility.[35] Basically, SVMs derive a linear classifier and are generally used in conjunction with kernel functions that implicitly map input data into a high-dimensional space. A linear separation in high-dimensional space is then facilitated corresponding to a non-linear separation of the original data.

In ligand-based virtual screening, the classification function derived by an SVM has also been successfully utilized to rank database compounds according to their probability of activity.[36,37] Because SVMs attempt to optimize classification performance, which not necessarily coincides with optimal ranking, specialized ranking methods have been introduced for virtual screening that utilize optimization functions to minimize the ranking error.[38,39] Thus, these functions are designed to reduce as much as possible the number of inactive compounds that are erroneously ranked above an active molecule in a virtual screen.

For high-dimensional sparsely set fingerprints, the linear SVM approach (which does not utilize non-linear kernel functions) has yielded good performance in benchmark calculations.[40] By exploiting the relatively low computational complexity of linear SVMs, they have been successfully applied to large-scale learning problems.[40] The kernel approach represents a rather flexible way to assess the similarity of molecules and enables the integration of different types of similarity measures in machine learning. For instance, non-numerical molecular comparisons such as maximum common subgraph assessment[41,42] and other graph-based kernels[38] have been successfully applied. In a recent study, SVMs have also been combined with feature vectors derived from pharmacophores of active compounds, thus utilizing an activity class-specific representation.[43]

In virtual screening, the integration of target and ligand information through the use of combined target–ligand kernels has received considerable attention in recent years.[44,45] Many different protein kernel functions that account for sequence similarity, structural similarity, or an ontology were combined with ligand similarity measures to predict novel ligands for orphan targets.[45] However, in benchmark calculation, these combinations did not further improve the predictions of compared to standard ligand kernels.[45] Accordingly, these findings suggested that compound similarity dominated ligand predictions. Nevertheless, a recent study including protein sequence information in a kernel successfully identified novel ligands of four (non-orphan) targets.[46] Another investigation utilizing a protein kernel specifically encoding structural binding site similarity (relying on high-quality X-ray structures of protein–ligand complexes) has produced promising results in recognizing true protein–ligand pairings.[47] These findings suggest new ways how to potentially exploit combined protein–ligand information in SVM-based virtual screening.[47] In addition, SVMs have also been applied in the context of structure-based virtual screening to evaluate docking results.[48] For this purpose, SVMs have been trained to distinguish docked decoys from true actives. Similarly, a pharmacophore-based interaction fingerprint has been used as a descriptor for SVMs to evaluate docking poses.[49]

8. Other trends in machine learning

In addition to SVMs, new machine learning methods continue to enter the chemoinformatics field. Conceptually related to SVMs are so-called relevance vector machines[50] that estimate posterior probabilities of activity based upon a set of relevance vectors. This methodology has also been successfully applied to ligand-based

virtual screening.[51] Furthermore, the original potential function method,[52] another kernel-based methodology that is conceptually simpler than SVMs, has been adapted for compound classification and shown in at least one study to yield results comparable to SVMs.[53]

A further trend in machine learning is the application of methods that combine results of different algorithms. In general, the results are combined by training a 'meta-classifier' using the output of the individual methods. This approach has been utilized to combine different machine learning approaches such as SVMs, naïve Bayesian classifiers, neural networks, decision trees, or random forests.[54,55]

9. SAR elucidation and visualization

Another expanding area of research in chemoinformatics is systematic SAR analysis and visualization. Large-scale SAR exploration efforts have been much supported by the release of the public domain ChEMBL compound repository,[56] a well-curated collection of compound activity data from medicinal chemistry literature sources. For screening data, PubChem[57] continues to be the major public domain source. Different approaches have been adapted or developed for SAR mining of screening or compound optimization data including extensions of established methods such as activity-sensitive clustering[30] or 3D-QSAR complemented with molecular fragmentation analysis to process SAR tables.[58] Furthermore, the scaffold tree data structure has been extended to include additional substructure relationships[59] or systematic scaffold generation as well as statistical models for activity prediction.[60] The increasing popularity of SVMs and other kernel methods is also impacting activity predictions, for example, through the adaptation of the kernel dimensionality reduction approach for feature selection to predict ligand–target pairings.[61] Moreover, the matched molecular pair (MMP) concept[62] is becoming increasingly popular in SAR analysis, catalyzed by the development of an elegant and efficient algorithm for systematic MMP generation (vide infra).[63] An MMP is defined as a pair of compounds that are distinguished by the exchange of a defined substructure (chemical transformation).[62] The MMP formalism has provided the basis for different types of SAR analysis approaches. For example, bioisosteric replacements have systematically been determined for different target familes[64] as well as substitutions that lead to the formation of activity ciffs[65] or alter multi-target compound activities.[66] Furthermore, also utilizing the MMP concept, a method has been introduced to detect SAR transfer events in compound data.[67] SAR transfer occurs if two analog series with different molecular scaffolds (core structures) display corresponding substitutions with comparable potency progression.

The activity cliff concept[8] is also receiving increasing attention in chemoinformatics and medicinal chemistry. An activity cliff is defined as a pair of structurally similar or analogous compounds with a large difference in potency. This concept has recently been extended by confirming the presence of an 'activity ridge' structure in many different compound data sets.[68] A ridge consists of multiple activity cliffs consistently formed between sets of highly and weakly potent analogs. As such, activity ridges are a data structure that is particularly rich in SAR information. Activity cliffs represent the extreme form of SAR discontinuity. In order to systematically extract compounds forming discontinuous local SARs from large data sets, the particle swarm optimization method has recently been successfully adapted,[69] thereby introducing another systematic SAR data mining approach.

For large-scale SAR analysis, visualization methods are also of high interest. Different types of SAR visualization tools have been reported including graphical representations of conventional SAR tables[70,71] and extensions of structure–activity similarity maps.[72,73] These maps report pair-wise comparisons of compound similarity and potency relationships in a data set. For a pair of compounds, calculated molecular similarity and activity similarity are plotted along two orthogonal axes, yielding a single data point per comparison. Moreover, first graphical representations of compound activity landscapes that capture multi-target SARs have been introduced.[74] For SAR visualization, molecular network representations are increasingly being used. In such networks, nodes represent compounds and edges similarity relationships.[75] These network representations are then annotated with potency and other SAR-relevant information. A number of different SAR network tools have been made publicly available.[75] Also, a first SAR network representation has been reported in which calculated similarity values were replaced with well-defined substructure relationships,[76] which further supports the chemical interpretability of such networks and the SAR information they capture. This approach has also utilized the MMP formalism to systematically identify substructure relationships between active compounds.

10. Statistical approaches

The application of statistical and information-theoretic methods also continues to be an active area of research in chemoinformatics. In addition to Bayesian classifiers that are widely used in chemoinformatics, Bayesian networks have recently been successfully applied in ligand-based virtual screening.[77,78] Furthermore, information-theoretic measures are particularly well suited for the analysis and evaluation of fingerprint similarity searching. The application of feature selection methods has made it possible to further improve fingerprint-based compound classification and similarity search performance.[79] Moreover, feature selection has identified fingerprint components that are relevant for detecting compounds belonging to a given class and also helped to rationalize why relatively simple search tools such as 2D fingerprints are able to recognize structurally diverse active compounds.[79]

11. New methodologies

In addition to developments in core areas of chemoinformatics, as discussed above, a number of new algorithms and approaches have been reported, in part to provide new solutions for long-standing problems.

Among the classical problems in chemoinformatics is the enumeration of chemical graphs. Recently, algorithms have been developed for the efficient enumeration of all stereo isomers of chemical graphs belonging to certain classes such acyclic and outerplanar graphs.[80,81] Also of general interest is the problem of enumerating all chemical graphs meeting pre-defined feature constraints, which can essentially be seen as the problem of reconstructing chemical graphs from a set of descriptors or features. For this purpose, algorithms have been introduced to enumerate tree-like chemical graphs based on the frequency of occurrence of paths up to a given length in the graph.[82,83] In addition, the maximum common substructure (MCS) concept is central to a variety of chemoinformatics applications. Given the computational complexity of determining an MCS, new algorithms designed for this purpose continue to be of interest to the community. For example, a fast incremental algorithm has recently been reported that heuristically determines the MCS.[84] In addition, for determining the MCS of more than two molecules, a novel method based on the correspondence graph has been introduced and applied to organize molecules into groups sharing a 'substantial MCS'.[85]

Recently, algorithms for the identification of MMPs[58] (vide supra) have become rather popular. Among these, an algorithm

developed by Hussain and Rea[63] has proven to be especially suited for systematic MMP generation and large-scale analysis of compound databases. This is due to the fact that the algorithm does not rely on pair-wise comparison of molecules. Instead, it constructs a library of systematically fragmented compounds obtained by systematically cutting acyclic single bonds. In contrast to this systematic approach, a recently reported alternative method determines the MCS of two molecules taking the immediate structural environment of replaced moieties into account.[86]

12. Computational efficiency

A number of recent studies have addressed computational efficiency of different methods, especially considering the use of graphics processing units (GPUs) that have also become popular in chemoinformatics. Given increasingly large compound source databases (e.g., the number of molecules available in PubChem[57] and ZINC[87] now exceeds 30 and 19 million, respectively), there is much interest in efficient implementations of classical chemoinformatics procedures such as similarity searching, clustering, or compound subset selection. These tasks have in common that they require pair-wise compound comparisons to determine molecular similarity. Indeed, calculation of similarity represents a major computational bottleneck when databases grow in size. This is especially the case for algorithms requiring the determination of full pair-wise similarity matrices such as, for example, standard clustering approaches. In addition to large database size, the development of efficient implementations is also motivated by at least two other factors. These include the popularity of sparsely populated high-dimensional fingerprints as descriptors such as extended connectivity fingerprints[13] or Molprint2D[88] and, on the more technical side, the availability of (cheap) graphical processing GPUs and efficient application programming interfaces like CUDA,[89] which enable the implementation of parallel code for GPUs in high level languages (CUDA C, OpenCL).[89,90]

GPUs are well suited for executing so-called single instruction multiple data (SIMD) algorithms, that is, algorithms that perform the same operations on different data in parallel. However, the optimization of implementations for GPUs is not without problems. For example, smart memory management is of prime importance; transferring data from the main CPU's memory to the GPU can become a bottleneck. Watson and co-workers implemented parallelized similarity calculations on GPUs for SVMs[91] and two compound selection algorithms[86] (i.e., the leader and spread algorithm) utilizing sparse[91] or dense[92] fingerprint representations. Another recent report focused on the GPU-based implementation of similarity search methods that require the determination of full similarity matrices like 1-NN searches.[93] Efficient implementations of Tanimoto coefficient calculations for dense and sparse fingerprint representations were provided, thus making 1-NN searches possible for databases of tens of millions of compounds against reference sets containing on the order of 10,000 molecules.[93] Due to memory and bandwidth constraints of GPUs, efficient encoding of fingerprints is an important factor for GPU-based adaptations of algorithms. A novel implementation[94] utilizes the Elia-coding[95] for count fingerprints that exhaustively enumerate subgraphs up to a given size such that large portions of a database can be kept in GPU RAM. Overall, for such GPU-based implementations, up to approximately 100-fold increases in search speed were reported compared to single-CPU implementations.[91–94]

Different from GPU-based approaches, Haque et al.[96] have shown that optimized implementations of similarity calculations on state-of-the-art multiple-core CPUs utilizing streaming SIMD extensions yield significant acceleration. Depending on the algorithmic problem at hand and the sizes of data sets under consideration, performance levels can be achieved that are only two to five times slower than GPU implementations; for some applications GPU-performance can be reached or even surpassed. Thus, in addition to GPUs, multi-core CPUs also offer much room for the improvement of computational efficiency over conventional implementations of chemoinformatics methods.

13. Concluding remarks

Herein we have provided a brief overview of the current status of the chemoinformatics field and discussed recent trends in method development. The chemoinformatics discipline is still in the process of defining itself and would benefit from community-wide efforts to further refine its scientific structure. Also, generally accepted standards for method evaluation are still lacking (which is of course also the case in other fields).[1,97] Our methodological discussion reflects the diversity of chemoinformatics and highlights the continued introduction of new methodologies for different types of applications. These also include new solutions to classical algorithmic problems. In addition, computational efficiency is still on the agenda, as one should expect. Furthermore, our survey identifies current growth areas such as support vector machine methodologies, systematic SAR analysis and visualization, or large-scale compound data mining, which have different methodological focal points. In SAR analysis, the matched molecular pair concept is increasingly being applied and has provided the basis for different developments. SAR analysis and compound data mining have been further supported by the introduction of the public domain ChEMBL compound collection, which represents another milestone for the chemoinformatics field, similar to the release of PubChem and also ZINC a few years ago. For chemoinformatics methods, the *Journal of Chemical Information and Modeling* continues to represent the central and by far most popular publication venue. We also note that another characteristic feature of chemoinformatics research is its strong theoretical focus, with still relatively little immediate experimental applications. Of course, in core areas of chemoinformatics such as the development of database and information systems, experimental applicability is naturally limited. An exception is the virtual screening area where increasing numbers of prospective applications are reported,[98–101] in addition to large numbers of (often questionable) benchmarking studies.[1,102] The impact of prospective virtual screening applications is often limited because the majority of newly identified hits are only weakly potent and their structural novelty might also be a matter of debate on occasions. In addition, the relevance of complex computational screening protocols for hit identification often remains obscure. Nevertheless, computational hit identification is without doubt an attractive area within the chemoinformatics spectrum with viable interfaces to experimental work. However, key issues in chemoinformatics that remain to be solved include, among others, the rationalization and calibration of relationships between calculated molecular similarity and observed activity similarity. The absence of methodological concepts to quantify and accurately predict such relationships has consequences. For example, it correlates with the still limited specificity of ligand-based virtual screening calculations, despite their popularity. Hence, in this and other areas, there remains much room for further innovation and scientific developments.

References and notes

1. Geppert, H.; Vogt, M.; Bajorath, J. *J. Chem. Inf. Model.* **2010**, *50*, 205.
2. Brown, F. K. *Ann. Rep. Med. Chem.* **1998**, *33*, 375.
3. Bajorath, J. *Drug Discovery Today* **2004**, *9*, 13.
4. Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. *Methods Mol. Biol.* **2004**, *275*, 131.
5. Lavine, B.; Workman, J. *Anal. Chem.* **2008**, *80*, 4519.
6. Warr, W. A. *Methods Mol. Biol.* **2011**, *672*, 1.

7. Peltason, L.; Bajorath, J. *Future Med. Chem.* **2009**, *1*, 451.
8. Wassermann, A. M.; Wawer, M.; Bajorath, J. *J. Med. Chem.* **2010**, *53*, 8209.
9. Weill, N.; Rognan, D. *J. Chem. Inf. Model.* **2009**, *49*, 1049.
10. Batista, J.; Tan, L.; Bajorath, J. *J. Chem. Inf. Model.* **2009**, *50*, 79–86.
11. Ahmed, H. E. A.; Vogt, M.; Bajorath, J. *J. Chem. Inf. Model.* **2010**, *50*, 487.
12. Liu, P.; Agrafiotis, D. K.; Rassokhin, D. N. *J. Chem. Inf. Model.* **2011**, *51*, 2843.
13. Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742.
14. Swamidass, S. J.; Baldi, P. *J. Chem. Inf. Model.* **2007**, *47*, 302.
15. Baldi, P.; Hirschberg, D. S.; Nasr, R. J. *J. Chem. Inf. Model.* **2008**, *48*, 1367.
16. Baldi, P.; Hirschberg, D. S. *J. Chem. Inf. Model.* **2009**, *49*, 1866.
17. Nasr, R.; Hirschberg, D. S.; Baldi, P. *J. Chem. Inf. Model.* **2010**, *50*, 1358.
18. Kristensen, T. G.; Nielsen, J.; Pedersen, C. N. S. *Algorithms Mol. Biol.* **2010**, *5*, 9.
19. Smellie, A. *J. Chem. Inf. Model.* **2009**, *49*, 257.
20. Agrafiotis, D. K.; Lobanov, V. S.; Shemanarev, M.; Rassokhin, D. N.; Izrailev, S.; Jaeger, E. P.; Alex, S.; Farnum, M. *J. Chem. Inf. Model.* **2011**, *51*, 3113.
21. Kristensen, T. G.; Nielsen, J.; Pedersen, C. N. S. *J. Chem. Inf. Model.* **2011**, *51*, 597.
22. Vidal, D.; Thormann, M.; Pons, M. *J. Chem. Inf. Model.* **2005**, *45*, 386.
23. Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. *J. Med. Chem.* **2005**, *48*, 1489.
24. Haque, I. S.; Pande, V. S. *J. Chem. Inf. Model.* **2010**, *50*, 1075.
25. Haque, I. S.; Pande, V. S. *J. Chem. Inf. Model.* **2011**, *51*, 2248.
26. Baldi, P.; Nasr, R. *J. Chem. Inf. Model.* **2010**, *50*, 1205.
27. Vogt, M.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 2496.
28. Hu, Y.; Stumpfe, D.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 1742.
29. Li, R.; Stumpfe, D.; Vogt, M.; Geppert, H.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 2507.
30. Lounkine, E.; Nigsch, F.; Jenkins, J. L.; Glick, M. *J. Chem. Inf. Model.* **2011**, *51*, 3158.
31. Rännar, S.; Andersson, P. L. *J. Chem. Inf. Model.* **2010**, *50*, 30.
32. Saito, S.; Hirokawa, T.; Horimoto, K. *J. Chem. Inf. Model.* **2010**, *51*, 61.
33. Meinl, T.; Ostermann, C.; Berthold, M. R. *J. Chem. Inf. Model.* **2011**, *51*, 237.
34. Podolyan, Y.; Walters, M. A.; Karypis, G. *J. Chem. Inf. Model.* **2010**, *50*, 979.
35. Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. *J. Chem. Inf. Model.* **2011**, *51*, 229.
36. Jorissen, R. N.; Gilson, M. K. *J. Chem. Inf. Model.* **2005**, *45*, 549.
37. Geppert, H.; Horvath, T.; Gärtner, T.; Wrobel, S.; Bajorath, J. *J. Chem. Inf. Model.* **2008**, *48*, 742.
38. Agarwal, S.; Dugar, D.; Sengupta, S. *J. Chem. Inf. Model.* **2010**, *50*, 716.
39. Rathke, F.; Hansen, K.; Brefeld, U.; Müller, K.-R. *J. Chem. Inf. Model.* **2010**, *51*, 83.
40. Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Ostermann, C.; Zell, A. *J. Chem. Inf. Model.* **2011**, *51*, 203.
41. Mohr, J.; Jain, B.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Obermayer, K. *J. Chem. Inf. Model.* **2010**, *50*, 1821.
42. Grave, K. D.; Costa, F. *J. Chem. Inf. Model.* **2010**, *50*, 1660.
43. Ranu, S.; Singh, A. K. *J. Chem. Inf. Model.* **2011**, *51*, 1106.
44. Jacob, L.; Vert, J.-P. *Bioinformatics* **2008**, *24*, 2149.
45. Wassermann, A. M.; Geppert, H.; Bajorath, J. *J. Chem. Inf. Model.* **2009**, *49*, 2155.
46. Wang, F.; Liu, D.; Wang, H.; Luo, C.; Zheng, M.; Liu, H.; Zhu, W.; Luo, X.; Zhang, J.; Jiang, H. *J. Chem. Inf. Model.* **2011**, *51*, 2821.
47. Meslamani, J.; Rognan, D. *J. Chem. Inf. Model.* **2011**, *51*, 1593.
48. Li, L.; Khanna, M.; Jo, I.; Wang, F.; Ashpole, N. M.; Hudmon, A.; Meroueh, S. O. *J. Chem. Inf. Model.* **2011**, *51*, 755.
49. Sato, T.; Honma, T.; Yokoyama, S. *J. Chem. Inf. Model.* **2009**, *50*, 170.
50. Tipping, M. E. *J. Mach. Learn. Res.* **2001**, *1*, 211.
51. Lowe, R.; Mussa, H. Y.; Mitchell, J. B. O.; Glen, R. C. *J. Chem. Inf. Model.* **2011**, *51*, 1539.
52. Aizerman, M.; Braverman, E. M.; Rozonoer, L. *Avtom. Telemekh.* **1964**, *25*, 917.
53. Mussa, H. Y.; Hawizy, L.; Nigsch, F.; Glen, R. C. *J. Chem. Inf. Model.* **2010**, *51*, 4.
54. Cheng, F.; Yu, Y.; Shen, J.; Yang, L.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. *J. Chem. Inf. Model.* **2011**, *51*, 996.
55. Kramer, C.; Beck, B.; Clark, T. *J. Chem. Inf. Model.* **2010**, *50*, 404.
56. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *Nucleic Acids Res.* **2012**, *40*, D1100.
57. Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. *Science* **2004**, *306*, 1138.
58. Wendt, B.; Uhrig, U.; Bös, F. *J. Chem. Inf. Model.* **2011**, *51*, 843.
59. Hu, Y.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 248.
60. Varin, T.; Schuffenhauer, A. Ertl. P; Renner, S. *J. Chem. Inf. Model.* **2011**, *51*, 1528.
61. Niijima, S.; Yabuuchi, H.; Okuno, Y. *J. Chem. Inf. Model.* **2011**, *51*, 15.
62. Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271–285.
63. Hussain, J.; Rea, C. *J. Chem. Inf. Model.* **2010**, *50*, 339.
64. Wassermann, A. M.; Bajorath, J. *Med. Chem. Commun.* **2011**, *2*, 601.
65. Wassermann, A. M.; Bajorath, J. *J. Chem. Inf. Model.* **2010**, *50*, 1248.
66. Hu, Y.; Bajorath, J. *ACS Med. Chem. Lett.* **2011**, *2*, 523.
67. Wassermann, A. M.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 1857.
68. Vogt, M.; Huang, Y.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 1848.
69. Namasivayam, V.; Iyer, P.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 1545.
70. Agrafiotis, D. K.; Wiener, J. J. M.; Skalkin, A.; Kolpak, J. *J. Chem. Inf. Model.* **2011**, *51*, 1122.
71. Agrafiotis, D. K.; Wiener, J. J. M. *J. Med. Chem.* **2010**, *53*, 5002.
72. Perez-Villanueva, J.; Santos, R.; Hernandez-Campos, A.; Giulianotti, M. A.; Castillo, R.; Medina-Franco, J. L. *Med. Chem. Commun.* **2011**, *2*, 44.
73. Yongye, A. B.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G. M.; Medina-Franco, J. L. *J. Chem. Inf. Model.* **2011**, *51*, 2427.
74. Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 256.
75. Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J. *J. Chem. Inf. Model.* **2010**, *50*, 68.
76. Wawer, M.; Bajorath, J. *J. Med. Chem.* **2011**, *54*, 2944.
77. Abdo, A.; Chen, B.; Mueller, C.; Salim, N.; Willett, P. *J. Chem. Inf. Model.* **2010**, *50*, 1012.
78. Abdo, A.; Salim, N. *J. Chem. Inf. Model.* **2010**, *51*, 25.
79. Heikamp, K.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 2254.
80. Imada, T.; Ota, S.; Nagamochi, H.; Akutsu, T. *J. Chem. Inf. Model.* **2011**, *51*, 2788.
81. Imada, T.; Ota, S.; Nagamochi, H.; Akutsu, T. *J. Math. Chem.* **2011**, *49*, 910.
82. Fujiwara, H.; Wang, J.; Zhao, L.; Nagamochi, H.; Akutsu, T. *J. Chem. Inf. Model.* **2008**, *48*, 1345.
83. Ishida, Y.; Kato, Y.; Zhao, L.; Nagamochi, H.; Akutsu, T. *J. Chem. Inf. Model.* **2010**, *50*, 934.
84. Kawabata, T. *J. Chem. Inf. Model.* **2011**, *51*, 1775.
85. Hariharan, R.; Janakiraman, A.; Nilakantan, R.; Singh, B.; Varghese, S.; Landrum, G.; Schuffenhauer, A. *J. Chem. Inf. Model.* **2011**, *51*, 788.
86. Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. *J. Chem. Inf. Model.* **2010**, *50*, 1350.
87. Irwin, J. J.; Shoichet, J. *J. Chem. Inf. Model.* **2005**, *45*, 177.
88. Bender, A.; Mussa, H. Y.; Glen, R. C. *J. Chem. Inf. Model.* **2004**, *44*, 1708.
89. NVIDIA CUDA C Programming Guide 4.1; NVidia: Santa Clara, CA, 2011.
90. Tsuchiyama, R.; Nakamura, T.; Lizuka, T.; Asahara, A. *The OpenCL Programming Book*; Oakland, CA: Fixstars, 2010.
91. Liao, Q.; Wang, J.; Webster, Y.; Watson, I. A. *J. Chem. Inf. Model.* **2009**, *49*, 2718.
92. Liao, Q.; Wang, J.; Watson, I. A. *J. Chem. Inf. Model.* **2011**, *51*, 1017.
93. Ma, C.; Wang, L.; Xie, X.-Q. *J. Chem. Inf. Model.* **2011**, *51*, 1521.
94. Liu, P.; Agrafiotis, D. K.; Rassokhin, D. N.; Yang, E. *J. Chem. Inf. Model.* **2011**, *51*, 1807.
95. Baldi, P.; Benz, R. W.; Hirschberg, D. S.; Swamidass, S. J. *J. Chem. Inf. Model.* **2007**, *47*, 2098.
96. Haque, I. S.; Pande, V. S.; Walters, W. P. *J. Chem. Inf. Model.* **2011**, *51*, 2345.
97. Jain, A.; Nicholls, A. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133.
98. Stumpfe, D.; Bajorath, J. Applied Virtual Screening: Strategies, Recommendations, and Caveats. In *Methods and Principles in Medicinal Chemistry. Virtual Screening, Principles, Challenges, and Practical Guidelines*; Sotriffer, C., Ed.; Wiley-VCH: Weinheim, 2011; pp 73–103.
99. Stumpfe, D.; Ripphausen, P.; Bajorath, J. *Future Med. Chem.*, in press.
100. Ripphausen, P.; Nisius, B.; Bajorath, J. *Drug Discovery Today* **2011**, *16*, 372.
101. Ripphausen, P.; Stumpfe, D.; Bajorath, J. *Future Med. Chem.*, in press.
102. Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. *J. Med. Chem.* **2010**, *53*, 5707.