ELSEVIER

feature

Exploiting personalized information for reagent selection in drug design

Jonas Boström*, jonas.bostrom@astrazeneca.com, Niklas Falk and Christian Tyrchan

Drug discovery is currently being industrialized. This fact is confusing, given that it is happening in times when the rest of the world has entered the subsequent information age. Here, we introduce a concept and an infrastructure for the now popular and well-known recommender systems in the context of exploiting one of the cornerstones of drug design: chemical reagent selection. The goal is to create and transfer information openly to facilitate intuition and serendipity in drug design. The system is tailored to highlight reagents from our corporate reagent database; reagents that a chemist might not have considered based purely on their own experience.

Introduction

The past decade has seen drug discovery attempt to rely less on the unpredictable nature of creative invention. Instead, systematic processes from other industries, such as car manufacturing, have been adapted [1,2]. Such process optimizations combined with timelines, 'metrics' and the abundant use of rules are now the commonly accepted approach to reducing attrition in drug discovery. The frequent use of rules primarily comes from the concept of rational design, and one might speculate that the seminal rule-of-five publication by Lipinski *et al.* [3] gave extra fuel to this school of thought. Lipinski *et al.* successfully correlated simple molecular properties, such as molecular weight (MW) and *c* Log *P*, with the increased risk of clinical failure [4,5]. Many other derivative studies followed and there have been numerous publications over the past decade on rules that use simple molecular descriptions that attempt to define drug-like chemical space [6–17]. One of the consequences of these rules is that the synthesis of molecules not fulfilling certain criteria is at best ill advised, and often completely restricted in most pharmaceutical companies. Despite this profound paradigm shift, the pharmaceutical industry shows no sign of reversing the decline in productivity, in terms of numbers of launched drugs. In truth, there is no evidence that introduction of rules (or any other technique for that matter) has increased pharmaceutical productivity. An ironic look at the number of approved drugs per year against the number of *Journal of Medicinal Chemistry* articles containing the word 'rule' (or 'rules') per year reveals a striking correlation (Fig. 1). Figure 1 should remind readers that correlation does not imply causation; that is, correlation between two variables does not automatically imply that one causes the other.

Despite their simplicity and the lack of statistical rigor underlying most drug-likeness rules, they can be helpful. In essence, the construction of most rules follows the medicinal chemistry mantra 'reduce lipophilicity', which is the near-universal response to any evidence of a problem. Lipophilic compounds, presumably owing to inadvertent effects on other biological targets, are more prone to side-effects than are hydrophilic compounds [17,18]. However, there are also drawbacks to following rules [19,20]. It increases the likelihood of rejecting the essential along with the inessential, or 'throwing out the baby with the bathwater'. A different type of rule hysteria is the use of results from experimental *in vitro* assays to predict complex clinical side-effects. For example, regulatory authorities require hERG (human Ether-a-go-go Related Gene) screening as a predictor of a compound producing cardiac arrhythmic side-effects. However, there are drugs that exhibit these side-effects with no evidence for any interaction with the hERG receptor and, conversely, there are many compounds known to interact with hERG and yet there is no evidence that they

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2011.01.006

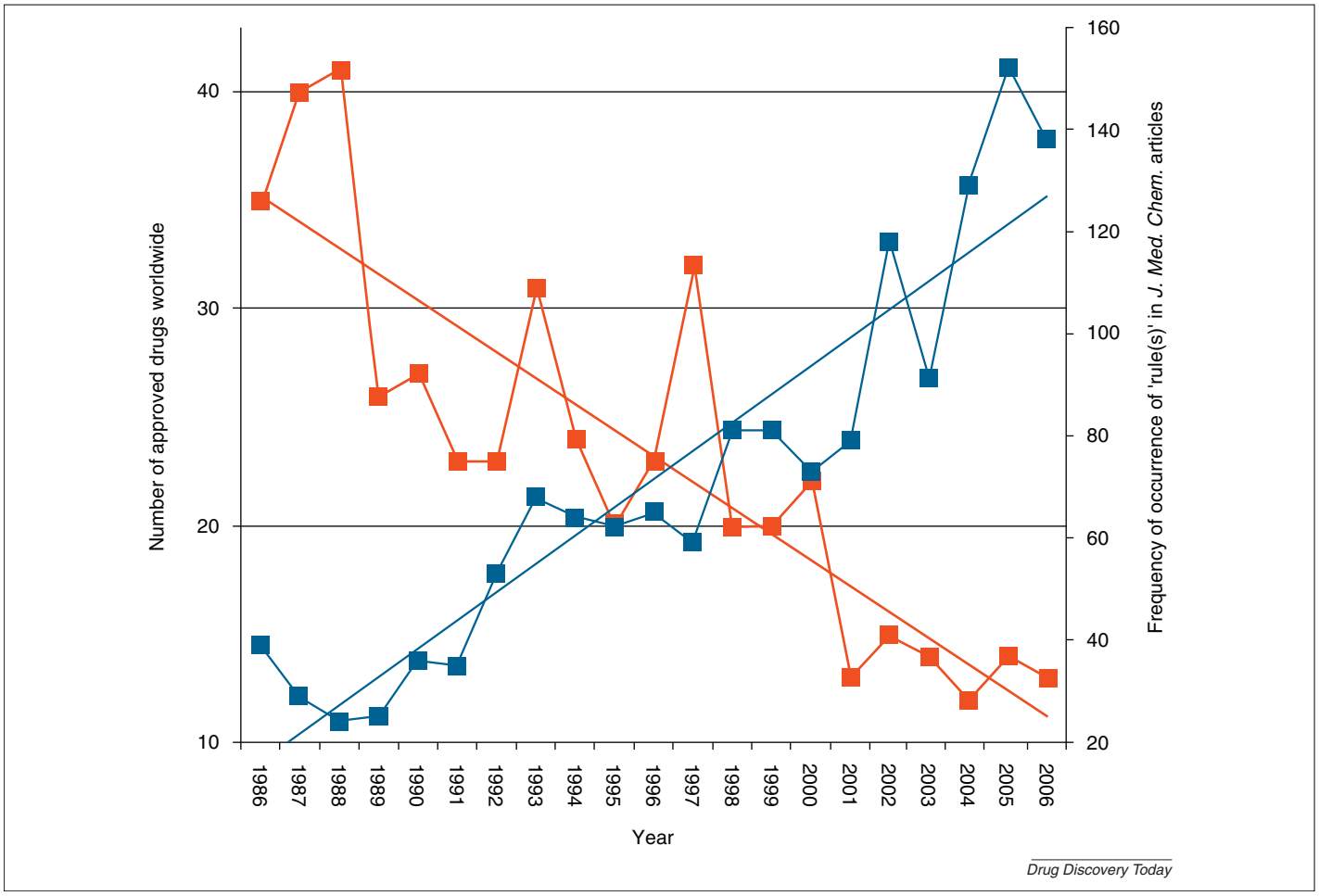


FIGURE 1

The number of oral drugs approved per year worldwide from 1986 to 2006 (red squares) and the number of *Journal of Medicinal Chemistry* articles containing the word 'rule(s)' per year over the same period (blue squares) reveals a striking correlation. The oral drugs data were taken from [9]. The advanced search option on the American Chemical Society Publications homepage (http://pubs.acs.org/search/advanced/) was used to search for the frequency of occurrence for the singular and plural form of the word 'rule' per year (the yearly intervals were specified by 'Print Publication Date', only *Journal of Medicinal Chemistry* articles were searched and the option field 'Anywhere in Content/Website' was used.

lead to cardiac complications [21]. The obvious consequence of these rules is that entire structural classes of chemistry have been labeled as hERG active when, in reality, such interactions tend to be compound specific. Hence, these types of rule-based approach can lead to detrimental restrictions on compound design for drug discovery.

Perhaps even more importantly, the mere existence of rules can stifle innovation and creativity, leaving medicinal chemists with the unlucky feeling of being stuck in corners of chemical space. This new way of working is in stark contrast to how drug discovery was performed in the past, where chance and accidents had important roles. For example, serendipity had a role in the discovery of many drugs, such as LSD, Thorazine®, Tofranil® and Viagra®, to name a few [22–24]. Serendipity can be taken to mean 'the effect by which one accidentally stumbles upon something fortunate, especially while looking for something entirely unrelated'. It should be stressed that this does not mean discovering something by sheer luck. Rather, it is the ability to see significances and find values in the chance discoveries. Without this ability, accidents will not lead to discoveries, as Louis Pasteur understood: 'in the fields of observation, chance favors the prepared mind'.

Our primary aim here is to show a possible way to increase the 'odds for serendipity' happening in drug discovery. This is done by using a form of the familiar recommender systems [25]. We apply the Amazon.com philosophy to develop a novel approach for recommending chemical reagents. The technical term used is 'item-to-item collaborative filtering' [26], but it is also commonly known as the 'people who bought . . .also bought. . .' of Amazon.com. Our system is designed to enhance discovery, and highlight chemical reagents in our corporate reagent database; reagents that medicinal chemists, using only their own judgment, might not have considered.

Recommender systems and item-to-item collaborative filtering

Recommender systems commonly attempt to recommend information about music, films, books and so on that are likely to be of interest. The most low-tech way to obtain recommendations is to ask a small group of people; your friends and family, for example. Statistical conclusions about the group can be drawn by collecting their answers. However, the advantage of the internet is that it enables large statistical sampling. This is why personalization and recommendation algorithms such as collaborative filtering were developed [27]. The two biggest internet successes, Google and

Amazon.com, serve as examples. There were already several other search engines available when Google started in 1998. However, Google took a completely new approach in that it ranked search results by using links on millions of websites to decide which pages to recommend. Google search results were so much better than its competitors that, by 2004, it was handling 85% of the searches on the internet. The other success company, Amazon.com, claims to know what you need. It 'knows' by using cleverly designed recommendation algorithms [26].

Recommender systems are usually classified into two main categories: content-based recommendations and collaborative recommendations. Content-based recommendations resemble conventional similarity searches to identify recommended items. By contrast, collaborative filtering makes recommendations by using learning algorithms based on the collective choices of individuals. Amazon.com introduced item-to-item collaborative filtering matching a given user's purchase of items to those purchased by other users. This analysis can be used to build a dynamic similarity measure. However, the idea of recommending items existed long before Amazon.com; for example, grocery stores have long put impulse buys on the checkout racks. Linden et al. at Amazon.com saw an opportunity to personalize impulse buys [26]. As a metaphor, it is as if the rack examined the grocery cart and magically rearranged the checkout rack based on what was in the cart.

To the best of our knowledge, this work represents the first example of using personalized information in drug design. The technical details of the implementation are described in Box 1, and the generation of the data set of reagents is explained in Box 2.

The dynamic nature of the recommendations

Although reagents are selected for many reasons, the motivation here is to present chemists with options related to a given reagent that might have proved a useful choice for molecular design reasons. Clearly, some 'expert' review of these choices is required, because not all recommendations will be based on problems relevant in making the current choice. This is analogous to E-commerce recommendation systems, where only a small fraction of recommendations are acted upon. The expert review ensures that recommendations are not treated as rules, but rather uses the inter-relationships built up between reagent selections, as a source of serendipity in reagent selection.

BOX 1

Computational details

The reagent recommendation is based upon the Amazon.com implementation of item-to-item collaborative filtering, and was implemented as follows. First, for each reagent, a vector is defined whose length is determined by the number of chemists using the reagent access system. In this case, the length was 193. The components of this vector represent whether a given chemist has checked out the reagent, in which case the component is set to a value of one, otherwise it is zero. It is then possible to compute a similarity matrix from these vectors, representing whether reagents are alike based on their usage by chemists. In principle, this matrix is large (13,292 × 13,292, Box 2). However, in practice, it is sparse, owing to the functional groupings of reagents (Table 1).

We used the original cosine approach to measure similarity [26]. Thus, the elements of the similarity matrix were computed using the cosine formula applied to vectors of arbitrary length. Given two vectors of attributes (*A* and *B*) the cosine similarity (θ) is represented using a dot product and magnitude, as shown in Eqn (I):

$$similarity(\vec{A},\vec{B}) = \cos(\vec{A},\vec{B}) = \frac{\vec{A}\cdot\vec{B}}{||\vec{A}|| \times ||\vec{B}||} \quad \text{(I)}$$

For a given reagent selected by a chemist, a set of recommendations is made by choosing the top ten most similar reagents from consideration of the similarity matrix. The recommended list of reagents all belong to the same functional class as the query reagent.

There is a risk that collaborative approaches simply capture only trivially popular reagent selections, sometimes referred to as the 'herd behavior' [28]. However, we take the more positive view that medicinal chemists are continuously solving a relatively narrow range of related problems (e.g. modifying lipophilicity, and solubility, avoiding off-target interactions, such as hERG and cytochrome P450s, and optimizing pharmacological processes, such as clearance), by their choice of reagents. Solutions to such issues are usually carefully considered, and often represent skillful medicinal chemistry. It is this type of thinking that the collaborative filtering model is trying to capture. Although there are anecdotal rules well known to medicinal chemists, we also anticipate that our implementation of collaborative learning captures herd behaviors beyond the obvious. One practical possibility for further avoiding bias in our data produced by a herd-like effect is to consider some type of normalization to be applied to the rank of a reagent based on its existing 'popularity'. For example, popularity could be derived from the historical use of the reagent, such that the collaborative methods seek reagents that are used with an unexpected high frequency by medicinal chemists to solve problems.

The calculated recommendations are dynamic and, consequently, will change over time. Every time a medicinal chemist uses (checks-out) a chemical reagent, it will have an effect on the similarity matrix, and thus provide the possibility

BOX 2

Data set

The data set of chemical reagents was generated as follows. All chemical reagents in the AstraZeneca R&D Mölndal stockroom flagged as being checked out at least once during the past five years were extracted. The reagents were then pretreated: covalently bonded salts were split; the smallest fragments were removed and canonical SMILES calculated. A check was made of whether the reagent samples were available (amount > 0.0), and unique IDs were assigned. This gave a number of 15,401 reagents. In a subsequent step, all the reagents were assigned into 19 common functional classes (amines, acids, alkylation reagents, etc) by SMARTS mapping (Table 1). Hence, by searching for amines, the user only obtains amine recommendations. That is, recommending, say, trifluoroborates would probably be too far off what the user was originally interested in (amines). Of the total reagents available, 13,292 could be mapped onto the 19 functional classes. A reagent with multiple functionalities will end up in multiple classes. Moreover, a significant fraction of reagents include a counter-ion. It was decided that entries having identical molecular structure, but with different counter-ions, should give the same results. Hence, the checkout data for such pairs were merged. All the data extraction, preparation and calculations were done using Pipeline Pilot v7.5.2 (Accelrys Software Inc.: http://www.accelrys.com/).

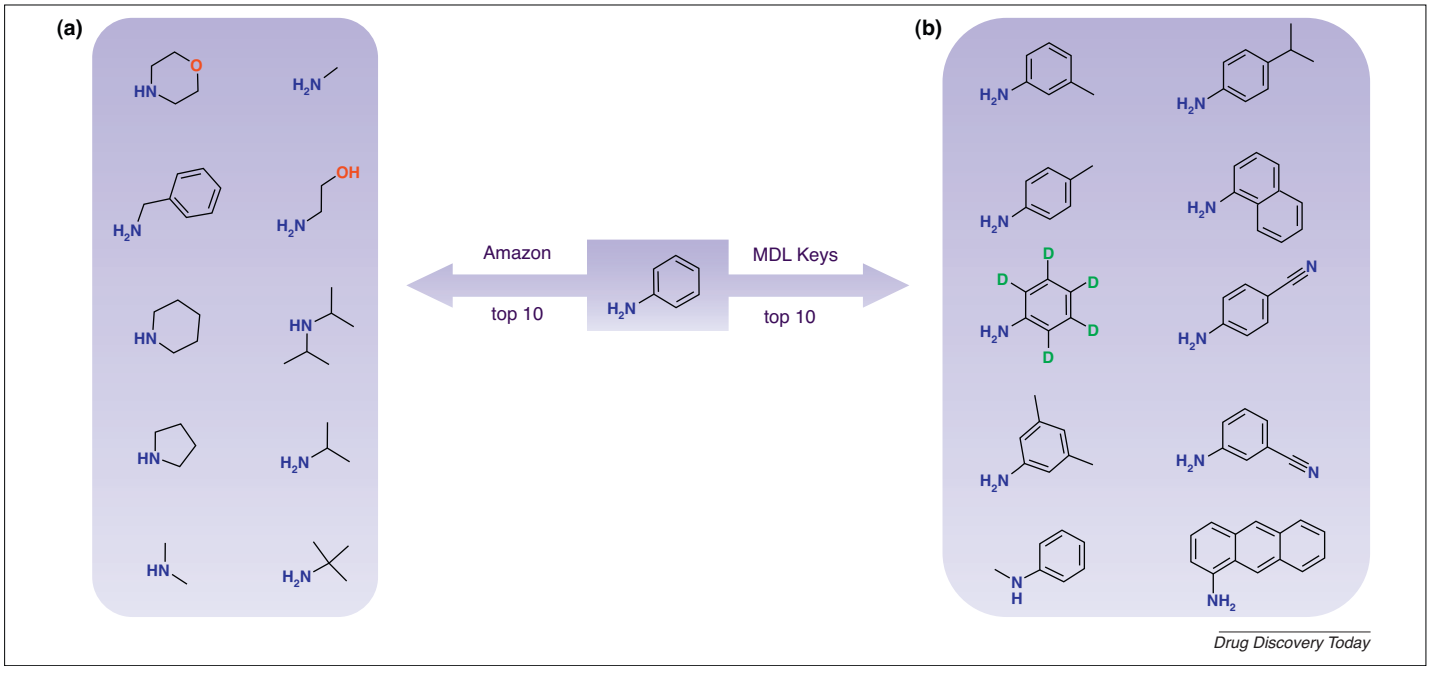


FIGURE 2

The top ten 'Amazon.com' recommendations are in all cases aliphatic amines **(a)**, whereas the ten most similar reagents as defined by MDL Public Keys are all aromatic amines **(b)**, when querying our internal stockroom with the structure of aniline.

of new and hopefully improved recommendations. The dynamic nature of the system facilitates its ability to pick up new trends, and one might speculate that the recommendations will converge at some point. That is, in a problem-solving manner this approach might pick up possible solutions, and could then serve as a guide to how to solve an existing problem.

To test the general advantage of using personalized information, and its inherent ability to aid knowledge transfer, we investigated whether our system could reveal solutions to the 'banned aniline' issue. To clarify, anilines have, during the past decade, been viewed as one of the most unwanted fragments in drug-like molecules. The reason for this obsession is that compounds including this moiety are occasionally found to induce genetic damage. Consequently, a common design strategy has been to steer clear of any aromatic amine, and simply use aliphatic amines as alternatives. This is probably the reason why only aliphatic amines are recommended using our method (Fig. 2). Interestingly, the aniline moiety is present in numerous prescribed drugs, notably Lipitor®; which is by far the most lucrative pharmaceutical of the past decade (Drug Information Online: http://www.drugs.com/top200.html). The preferred experiment to predict whether a compound is genotoxic is the AMES test [29,30], which is thought to have a high correlation with rodent carcinogenicity. Figure 3 shows seven aromatic amines that are structurally similar to aniline, and yet were found to be negative in our internal AMES assay, as curiously is aniline. None of these reagents was recommended by our method. In this respect, we acknowledge that this is presently a limitation of the collaborative learning approach, arising because none or too few chemists have made the association between these particular AMES data and reagent selection. This could be improved by the method being more sensitive to detecting 'emerging' herd behavior, by identifying reagents used with unexpected frequency (in this case, when a few medicinal chemists become aware of the negative AMES data for the subset of aromatic amines). By contrast, it also illustrates that, although the method is useful for identifying common reagent usage patterns, it is intended to complement expert and insightful treatment

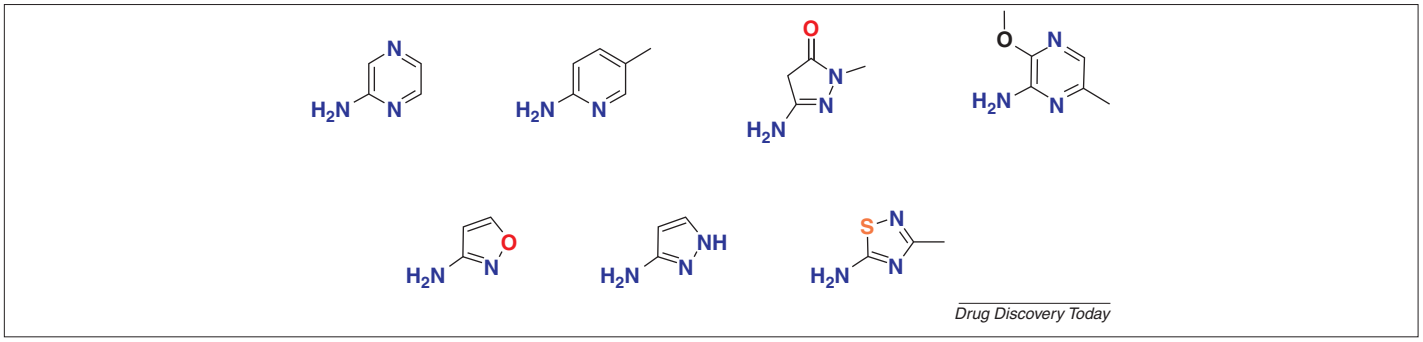


FIGURE 3

Seven aromatic amines, structurally similar to aniline, which have been determined to be AMES negative.

TABLE 1

The 13,292 reagents assigned onto the 19 functional classes by SMARTS mapping.

Class	Reagents (no.)[a]	Functional groups
01	4240	Primary and secondary amines
02	4445	Acids, acid halides, esters and anhydrides
03	1721	Aldehydes, ketones and acetals
04	1056	Aromatic halides
05	326	Sulfonic acids and sulfonic acid chlorides
06	1376	Aliphatic alcohols
07	2487	Alkyl halides
08	410	Boronic acids, boronic esters and trifluoroborates
09	183	Isocyanates and isothiocyanates
10	81	Alpha halide ketones
11	884	Aromatic alcohols
12	34	Hydrazines
13	48	Epoxides
14	736	Amino acids
15	110	Alkynes
16	51	Isocyanides
17	224	Aliphatic nitriles
18	464	Aromatic nitriles
19	302	Nitros

[a] Reagents with multiple functionalities were assigned to multiple classes.

of data by medicinal chemists.It is not our intention that medicinal chemists follow the herd behavior, but by ensuring awareness of what it is, it can be a useful influence on their thinking.

Comparisons between recommendations and similarity-based fingerprint methods

To investigate whether our user-item (i.e. chemist-reagent) cosine-derived similarities differ from traditional similarity measures, they were compared to three frequently used fingerprints: functional circular fingerprint 6 (FCFP6) (Pipeline Pilot v7.5.2, Accelrys Software Inc.; http://www.accelrys.com/), extended connectivity fingerprint 6 (ECFP6) [31] and Molecular Design Limited (MDL) public keys [32]. This was done by generating ranked lists for each method, and subsequently counting the number of structures in common within the top ten and top 20 ranked lists. That is, the top *x* FP-hits compared with the top *x* recommendations (Tables 1 and 2). The results in Table 2 show that Amazon.com recommendations differ significantly from similarity searching fingerprint techniques (i.e. FCFP6, ECFP6 and MDL public keys). The average overlap among the top ten recommendations is no more than 12–13% when comparing Amazon.com recommendations with the corresponding results for the three fingerprint-based methods. The overlap is slightly higher when comparing the top 20 recommendations and/or hits (Table 2). The low degree of overlap is expected because most similarity-based methods will give different hits [33–35]. As a comparison, the top 20 overlap between the two fingerprint methods ECFP6 and MDL public keys is 60%. This is significantly higher compared with the overlap for Amazon.com recommendations. One might conclude that the degree of overlap with other methods is likely to be low, and the obtained recommendation will definitely add novel suggestions, probably suggestions that would not have been highly ranked by other methods.

TABLE 2

The percentage overlap between the top ten and top 20 recommended reagents and the top ten and top 20 fingerprint similarities (ECFP6, FCFP6 and MDL public keys)

Recommended reagents	Overlap (%) Fingerprint similarities		
	ECFP6	**FCFP6**	**MDL public keys**
10	12.3	12.7	13.4
20	21.3	21.9	23.1

To illustrate, Fig. 3 shows the top ten structures selected with the Amazon.com approach compared with the corresponding MDL public keys, using aniline as query reagent. This test case serves as an example of the different reagents recommended. It is striking that the recommendations are, in all cases, aliphatic amines, whereas the most similar reagents as defined by MDL public keys are all aromatic amines.

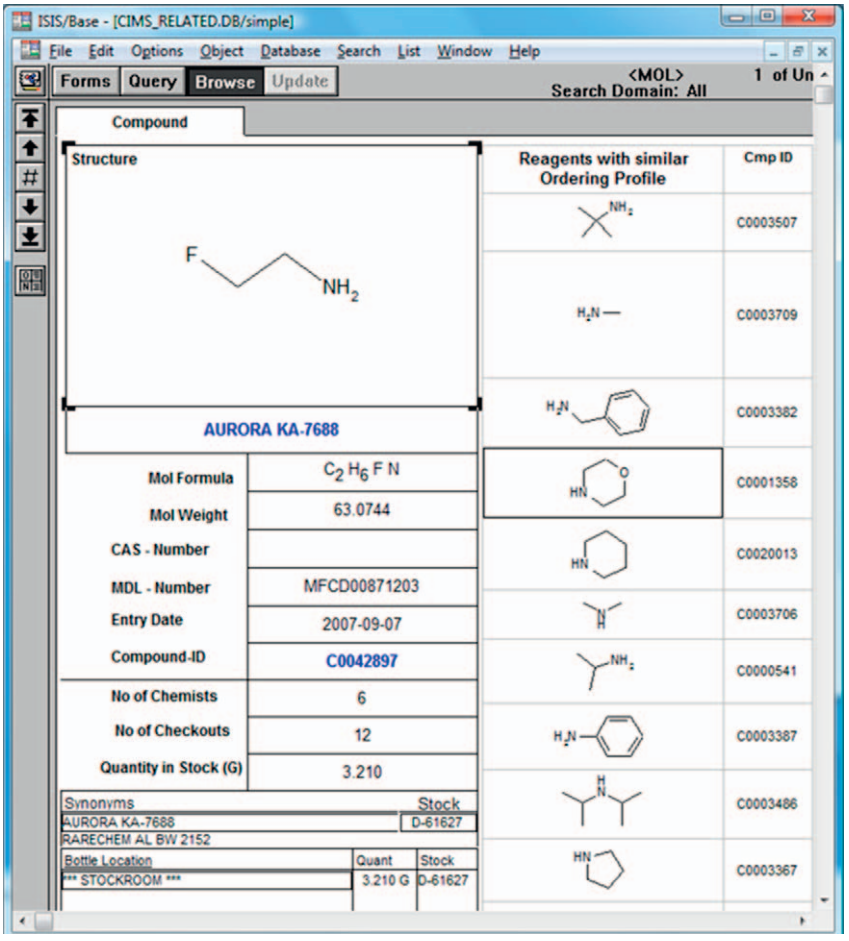
As an aside, the components of the vectors used to construct similarity are binary in our current implementation. Other fingerprint approaches for similarity, such as ECFP [31], suggest that there is an advantage of explicitly tracking the counts of reagent use in the vectors. Initial experiments show that using such counts does not provide appreciatively different recommendations. Nevertheless, this is an interesting track worth pursuing in future work.

Technical limitations with the current implementation

At AstraZeneca R&D Mölndal, the standard database management system for searching and displaying synthetic organic reaction structures and their associated data is still ISIS/Base (MDL ISIS™/Base 2.5 SP 4: http://www.symyx.com). ISIS has a client–server architecture that enables direct communication with Oracle databases. Given that the technical infrastructure is well known to us, and the query interfaces are easily customizable, it enabled us to build up quickly a prototype application for interfacing the Amazon.com-based reagent recommendations.

Given that we are familiar with the technical infrastructure, the threshold for creating the new ISIS/Base interface was low. The first working prototype was finished within a fortnight, and the production version is easy to maintain. Data are readily updated on a weekly basis, but can be updated more frequently if desired. Figure 4 depicts the stockroom ISIS/Base interface, where the recommended reagents are listed to the right.

Admittedly, the technical infrastructure and information content used are not ideal for fully exploiting the full force of recommender systems. For example, the strategy of tailored problem-specific approaches is difficult to



Drug Discovery Today

FIGURE 4

A view of the new ISIS/Base stockroom interface, including the recommended reagents.

implement in a straight-forward manner, if not impossible. In the current version, we cannot track the chemical reaction, or the purpose for why a chemical was checked out (e.g. target affinity increase, hERG decrease, solubility increase, etc). Compound quality improvements and demonstrating usage cannot be done unambiguously, if measured by numbers and metrics. Currently, we rely on soft information, such as word-of-mouth feedback. We cannot rule out tiresome prejudices, such as that the recommendation lists are driven by synthetic ease. The points just mentioned highlight the use and necessity of high-quality annotations and easy access to information in new and improved technical infrastructures.

Summary

Drug discovery is facing an era of industrialization. Without a doubt, drug research is currently being reduced to processes, and is subject to analyses more fit to a car assembly line. This fact is confusing, given that it is happening in times when the rest of the world has entered the subsequent information age. Certainly, a timelier strategy for drug discovery is to focus more on the ability of individuals to create and transfer information instantly and freely. This latter approach has advantages, such as instant access to knowledge (knowledge that previously would have been difficult to find) as well as being able to share information, for example using social media.

Here, we present a new application for the well-known recommender systems in the context of exploiting one of the cornerstones of drug design: chemical reagents. The system uses the item-to-item collaborative filtering technique of Amazon.com. The system is designed to assist medicinal chemists to discover chemical reagents they might not have found by themselves, but are potentially attractive based on the collective experience of many colleagues in AstraZeneca.

The method is novel in itself and differs fundamentally from common similarity measures, as it relies on user-item information and not descriptions of molecular structures. Results show that the recommendations are, more or less, orthogonal to conventional similarity methods.

One potential long-term advantage of using a recommender system lies within its dynamic nature, and the inherent possibility of transferring knowledge efficiently. At present, we cannot prove (with numbers) that this is the case. However, we were triggered to do this work in an unpretentious wish for drug discovery to leave industrialism behind, and enter the information age, as well as attempting to put previously successful approaches, such as intuition and serendipity, back on the drug discovery chart. If we can inspire one, and only one, medicinal chemist to think in a new direction, to focus less on processes, timelines and metrics and rely on the unpredictable

nature of creative invention, then we see that as a success.

Acknowledgement

The authors thank Dr J. Andrew Grant for many valuable discussions.

References

1 Andersson, S. *et al.* (2009) Making medicinal chemistry more effective: application of lean sigma to improve processes, speed and quality. *Drug Discov. Today* 14, 598–604
2 Petrillo, E. (2007) Lean thinking for drug discovery – better productivity for pharma. *Drug Discov. World* 8, 9–16
3 Lipinski, C. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25
4 Lipinski, C. *et al.* (2000) Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249
5 Lipinski, C. and Hopkins, A. (2004) Navigating chemical space for biology and medicine. *Nature* 432, 855–861
6 Wenlock, M.C. *et al.* (2003) A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* 46, 1250–1256
7 Veber, D.F. *et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623
8 Johnson, T.W. *et al.* (2009) Using the golden triangle to optimize clearance and oral absorption. *Bioorg. Med. Chem. Lett.* 19, 5560–5564
9 Leeson, P.D. and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* 6, 881–890
10 Leach, A.R. *et al.* (2006) Fragment screening: an introduction. *Mol. Biosyst.* 2, 430–446
11 Kelder, J. *et al.* (1999) Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* 16, 1514–1519
12 Bhal, S.K. *et al.* (2007) The rule of five revisited: applying log *D* in place of log *P* in drug-likeness filters. *Mol. Pharm.* 4, 556–560
13 Waring, M.J. (2009) Defining optimum lipophilicity and molecular weight ranges for drug candidates; molecular weight dependent lower log *D* limits based on permeability. *Bioorg. Med. Chem. Lett.* 19, 2844–2851
14 Lovering, F. *et al.* (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* 52, 6752–6756
15 Ritchie, T.J. and Macdonald, S.J. (2009) The impact of aromatic ring count on compound developability: are too many aromatic rings a liability in drug design? *Drug Discov. Today* 14, 1011–1020
16 Waring, M.J. and Johnstone, C. (2007) A quantitative assessment of hERG liability as a function of lipophilicity. *Bioorg. Med. Chem. Lett.* 17, 1759–1764
17 Hughes, J.D. *et al.* (2008) Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* 18, 4872–4875
18 Gleeson, M.P. (2008) Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* 51, 817–834
19 Zhang, M.Q. and Wilkinson, B. (2007) Drug discovery beyond the 'rule-of-five'. *Curr. Opin. Biotechnol.* 18, 478–488
20 Lajiness, M.S. *et al.* (2004) Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discov. Dev.* 7, 470–477
21 Ray, W.A. *et al.* (2009) Atypical antipsychotic drugs and the risk of sudden cardiac death. *N. Engl. J. Med.* 360, 225–235
22 Kubinyi, H. (1999) Chance favors the prepared mind. From serendipity to rational drug design. *J. Recept. Signal Transduct. Res.* 19, 15–39
23 Ban, T.A. (2006) The role of serendipity in drug discovery. *Dialogues Clin. Neurosci.* 8, 335–344
24 Terrett, N.K. *et al.* (1996) Sildenafil (VIAGRA™), a potent and selective inhibitor of type 5 cGMP phosphodiesterase with utility for the treatment of male erectile dysfunction. *Bioorg. Med. Chem. Lett.* 6, 1819–1824
25 Adomavicius, G. and Tuzhilin, A. (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* 17, 734–749
26 Linden, G. *et al.* (2003) Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Comput.* 7, 76–80
27 Goldberg, D. *et al.* (1992) Using collaborative filtering to weave an information tapestry. *Commun. ACM* 35, 61–70
28 Hamilton, W.D. (1971) Geometry for the selfish herd. *Theor. Biol.* 31, 295–311
29 Mortelmans, K. and Zeiger, E. (2000) The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.* 455, 29–60
30 Ames, B.N. *et al.* (1973) An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proc. Natl. Acad. Sci. U.S.A.* 70, 782–786
31 Rogers, D. and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754
32 Durant, J.L. *et al.* (2003) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comp. Sci.* 42, 1273–1280
33 Muchmore, S.W. *et al.* (2010) Cheminformatic tools for medicinal chemists. *J. Med. Chem.* 53, 4830–4841
34 Shanmugasundaram, V. *et al.* (2005) Hit-directed nearest-neighbor searching. *J. Med. Chem.* 48, 240–248
35 Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today* 7, 903–911

*Jonas Boström**
Niklas Falk
Christian Tyrchan
Lead Generation Department,
AstraZeneca R&D Mölndal,
S-431 83 Mölndal, Sweden