ELSEVIER

Bayesian versus Frequentist statistical modeling: A debate for hit selection from HTS campaigns

L. Martin Cloutier[1] and Suzanne Sirois[2]

[1] Department of Management and Technology, Room R-3570, School of Management, University of Quebec at Montreal, 315 Ste. Catherine East, Montreal, QC H2X 3X2, Canada
[2] Department of Chemistry, University of Quebec at Montreal, P.O. Box. 8888, Succ. Centre-Ville, Montreal, QC H3C 3P8, Canada

The existing literature suggests the Bayesian–Frequentist debate could soon be involved in the prioritization of hits from HTS campaigns. The Bayesian–Frequentist debate reflects two archetypical attitudes regarding the process of conducting scientific and technological research. This review article covers recent advances in statistical analyses, currently in use, for hit selection in the drug discovery process. The impact of decisions (e.g. attrition) executed at early stages in the drug discovery process influences HTS performance in later development stages. It shows that, as the high content value of the information from HTS campaigns increases over time, the two statistical approaches aim to provide similar answers, but they might not succeed.

Often, there are archetypically opposing perspectives to perceive and understand the world: Origene versus Tertullian; Baconian versus Newtonian; Bayesian versus Frequentist. Alternative perspectives do not always provide the same answer to inform the object of inquiry. Eflon [1] was the first to introduce the Bayesian–Frequentist debate in the biological scientific community by explaining how these perspectives reflect relevant alternative attitudes for conducting science. Eflon suggested data analyses may contribute to the Bayesian–Frequentist debate. Perhaps this suggestion could be extended to help identify a workable compromise in the interest of hit selection in HTS campaigns.

The Bayesian–Frequentist debate exemplifies that HTS is still a young innovative technology and history offers limited help to researchers on this journey because hit selection was conducted mostly using basic statistics.[3] An analysis of literature and software for hit selection in HTS campaigns reveal that researchers either adopt a Bayesian or a Frequentist perspective. This article discusses the impact of using a Bayesian versus a Frequentist view on analyzing HTS data and its implications. The *Z*-score method and its variants are commonly used for hit selection in HTS assays [2–6] whereas Bayesian statistics have been introduced in the literature for hit selection more recently [7–12], although they exist for a longer time [13,14].[4] With an emerging use of Bayesian methods in every nook-and-cranny of the drug discovery process (DDP) and clinical research, more attention should be brought to highlight distinctions between these two perspectives to understand why it matters. The Frequentist perspective attributes an equal probability to each member of a set of events, simply counting how many of them are 'favorable' as a proportion of the total. By contrast, the subjective interpretation, also known as subjective Bayesianism, identifies probabilities with degrees of confidence, or 'partial' beliefs.

Resolving this debate can inform the decision-making framework employed by scientists in hit selection. The objective of this paper is to sort out how and why, within the DDP, Frequentists and Bayesian perspectives for hit selection and prioritization are key ingredients of a larger system that can impact its performance significantly. Indeed, it remains unclear which statistical perspec-

Corresponding author: Sirois, S. (suzanne.sirois@gmail.com)

[3] A review of the current statistical practice in HTS data analysis was conducted by Malo *et al.* [2].

[4] An extensive empirical evaluation of machine-learning techniques on HTS data such as support vector machine, naive-Bayesian (NB) classification, *k*-nearest-neighbor classification, memory-based learning, random forests, artificial neural networks, genetic function approximation, standard and partial least squares, principal component analysis regression, recursive partitioning, among others, would be extremely useful.

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2008.03.022

tive (with a high degree of confidence) is more appropriate for hit selection from HTS campaigns. But, expectations and results on the basis of a particular statistical perspective continue to widen. A retrospective of the past ten years could help explain the productivity gap between expectations about HTS platforms and available empirical evidence. The expectation was that biotechnology could transform pharmaceutical innovation by increasing the number and effectiveness of drugs and diagnostics [15,16]. Instead, according to the FDA: 'the applied sciences needed for medical product development have not kept pace with tremendous advances in the basic sciences' [17,18]. There could be a disjuncture between technological complexity in drug development and its gigantic data production, and the fundamental hypotheses in basic science from which drugs are discovered and developed (i.e. target-based or function-based approach[5] [19–22]). HTS technology has drifted away from theory-driven rational drug design, back towards empirically driven screening of large libraries of compounds, and now is back again to a theoretical understanding of chemical and biological space with the integration of data mining and machine learning techniques [16,38]. The latter is evidenced by the surge in methods and the increasing number of scientists bringing statisticians and cheminformaticans millions of data and thousands of parameters to consider all at once [1].

The remainder of the paper is organized as follows. Following this introduction, there is a series of sections on the role and relevance, respectively, of Frequentist and Bayesian perspectives to the DDP. Key findings from papers using these statistical perspectives are surveyed and commented. A discussion section establishes connections between the information generated from statistical analyses placed within the decision-making context scientists exercise for hit selections in HTS. A conclusion follows.

Type I and type II errors

Statisticians are concerned by two significant sorts of statistical error: false positive (FP) or type I error; and false negative (FN) or type II error [2,23]. A null hypothesis is set up to be nullified or refuted to support an alternate hypothesis. The FP is the error of the first kind (i.e. it is the error of rejecting a null hypothesis (H_0) when it is actually true). The FN is the error of the second kind (i.e. the error of failing to reject the null hypothesis (H_0) given that the alternative hypothesis (H_1) is true). For example, H_0: no hit, this is the null hypothesis; H_1: hit, this is the alternative hypothesis. Type I error rejects the null hypothesis, which is to say that the compound is presumed active (a FP), and similarly for the type II error.

Systematic versus random errors

The first step in any HTS experiment is to optimize the technical and procedural efficiencies and to determine the level of statistical error. In contrast to a random error, a systematic error results from a degree of bias in the measurement process and is not because of chance. HTS systematic errors encompass assay artifacts, compounds that are reactive, fluorescent, quenchers, aggregators, chelators and reducing agents. These compounds tend to read as FPs or FNs they affect cells, the target protein or the detection method [12,24]. Other sources of systematic errors are reagent and temperature errors, errors in liquid handling, reading and time, and errors related to storage conditions such as compound precipitation from DMSO stock solutions [12]. These systematic errors and artifacts can often be minimized by quality control procedures. Other factors, however, can escape the filters typically applied to quality control procedures but could introduce exogenous variations (e.g. variability in compound potency, row and column biases, across and within place column and row) [2].

HTS pre-processing data analysis

Hit identification is affected by several factors such as assay format and quality and screening measurement variability. After detection and elimination of systematic errors three methods are widely used during HTS data pre-processing. They are percent of control, normalized percent inhibition and *Z*-score [3,6] and its variants [4]. They implicitly assume a random error distribution that is common to all measurements within a single plate [2]. But differences in variability raise questions about the constant error assumption because variability among replicates can differ at various concentrations [6]. More robust data analysis procedures such as *B*-score [4] have been adopted, since these methods rely on mean and standard deviations, and are influenced by statistical outliers. After a score is obtained for each compound, a cutoff has to be chosen to classify the tested compounds into potential hits or non-hits.

Frequentist methods

Following HTS pre-processing on assay quality, subsequent analyses for hit profiling with highly sophisticated statistical and/or cheminformatics methods are key. Frequentist approaches try to define confirmation rate (CR), false-positive rate (FPR), and false-negative rate (FNR) at different hit threshold limits to select hits. The threshold (the *hit limit*) for declaring a hit is expressed as a certain activity value or several standard deviations away from the mean: three SDs (standard deviation) of the library population in a primary screen [3,6] or the robust version of three MADs (median absolute deviations) away from the median [4]. HTS data are often noisy, contain FPs and FNs [2,6,12,23,25], and cell-based assays are considered noisier than enzymatic screens [12]. This implies that a correct selection of the primary hit list is paramount. In a typical screen, inactive compounds vastly outnumber active compounds. This is illustrated in Diller *et al.*'s example [25]: '... on average from a single screen we would expect $0.999 \times 1000 = 999$ true positives and $999000 \times 0.001 = 999$ false positives. Thus, even with a nearly perfect assay, false positives are still equal in number to true positives .... Furthermore, if we assume a more realistic true hit rate of 1 in 10 000 compounds, we would on average find 10 FPs for every true positive'. But, according to Malo *et al.* [2], it remains unclear whether current inefficiencies are mostly because of the generation of too many FPs, too many FNs or both. A large number of FPs early on could overwhelm the number of true positives, and FNs would be overlooked. As a result, the true active compounds may not be discovered. For example, in a screen with a large number of FPs the activity threshold may be increased to obtain a manageable number of compounds to follow up. As a result, a

[5] The target-based approach aims to develop a target-selective drug and the function-based approach aims to produce a specific biological effect irrespective of its mode of action. The target-based success approach goes alone with better target selection on the basis of druggability through computational algorithms.

large number of FPs will influence the number of FNs from a given assay, hence reaching higher sensitivity at the expanse of lower specificity.

Very often, information remains about the FPR after confirmation experiments. The longer FP hits remain in the DDP the greater time and financial resources are wasted. This highlights the importance of identifying FPs in the process, as well as compounds with undesirable ADMET [26] or PK profiles [24], as early on as possible to avoid costs incurred with time delay and financial resources that increase substantially in subsequent steps (i.e. try to adopt a 'fail-fast' and 'fail-cheap' management approach). Moreover, because each HTS run is very costly, the vast majority of primary screens provide only a single measurement of each compound's activity at one concentration. Researchers have demonstrated through sophisticated statistical analysis [2,6] that replicates reduce the number of FPs without increasing the number of FNs. Nowadays, the commonly used strategies for hit selection with high FPR is duplicate screening, and with high FNR is two-tier hit selection (in which the primary and confirmation hit limit settings are different) [6].

With the Frequentist approach there is no chemical information associated with hit selection, unless it is conducted sequentially with cheminformatics methods (e.g. QSAR, clustering and partitioning techniques). Whereas, Bayesian approaches calculate a probability distribution for active and inactive compounds in a training set from HTS data guided by the frequency of occurrence of various molecular descriptors or fingerprints. Thus, Bayesian machine learning methods embed molecular features within the probability distribution. Moreover, the handling of noisy screening data by Bayesian probabilistic binary QSAR has set up the stage for cheminformatics *in silico* statistical modeling of HTS data [13].

Bayesian methods

Bayes's rule of conditional probability developed by Bayes [27] is the prior probability of *A* being true without any knowledge of *B*. It is a widely used method of statistical inference applied to many real-world problems. The adoption of Bayesian statistical methods by researchers in the DDP is recent, and they are now widely used. These methods have emerged as potentially useful within the context of drug discovery and QSAR. Labute has pioneered the introduction of a Bayesian inference technique within HTS data analysis with the introduction of the Binary QSAR method (BQ) [13,14]. Following his work, many applications of Bayesian statistics in the DDP (see Table 1) now include prioritization of hits from HTS campaigns [7,9,12,25,28], prioritization of compounds for combinatorial synthesis and HTS, or to optimally select compounds from third party data collections [8,10,11,14], specific cheminformatics applications such as the creation of QSAR models [11,13,14], similarity searching [29], virtual screening [30–33], high-throughput docking [31,34–36], prediction of protein activity inhibition by small molecules [8], ADME prediction [26], among others.

The Bayesian inference technique, BQ, introduced by Labute is a nonlinear modeling method [13]. It is a method based on statistical probability estimation (and not regression) suited to the analysis of HTS data by correlating structural properties of compounds with a binary expression of biological activity (1 = active and 0 = inactive) and calculating a probability distribution for

TABLE 1

Naïve Bayesian classifier (NB), Laplacian-modified naïve Bayesian classifier (LMNB) and multiple-category Bayesian models (MCBN) and their applications

Application	NB	LMNB MCBM	Descriptor or fingerprint	PCA	Refs
Construct a probabilistic QSAR model from HTS data	NB		Chi, VSA	X	[13,14]
Similarity searching, Virtual screening and the elucidation of binding patterns	NB		Atom environments		[29,30]
Deriving knowledge through data mining HTS	NB		ROF[a]		[25]
Docking, structure-based virtual screening		LMNB	ECFP[b], FCFP		[31]
Enrichment of extremely noisy HTS data	NB		ECFP		[7]
Enrichment results from HTD[c] and combination with consensus scoring	NB		ECFP		[34,35]
Prioritize compounds for screening or select compounds from third party data collections		LMNB	ECFP		[8]
HTS follow-up		LMNB	ECFP, FCFP		[9]
Enrichment of HTD results		LMNB	ECFP		[36]
Modeling ADMET		LMNB			[26]
Virtual screening, prioritize libraries for combinatorial synthesis and HTS			MACCS[d]	X	[10]
Virtual screening		LMNB	ECFP		[48]
Triaging and prioritization of the primary hit list, QSAR model from HTS data, select a subset of the collection for screening		MCBM	ECFP		[11,12]

[a] Rule of five descriptors include hydrogen bond donor and acceptor, number of rotatable bonds and molecular weignt.
[b] The ECFPs are a new class of fingerprints for molecular characterization, developed by the Sci-Tegic group [9], that rely on the Morgan algorithm, see ref. [3] in [9].
[c] HTD, high throughput docking.
[d] MACCS descriptors includes 166 public MDL keys.

active and inactive compounds in a training set guided by the frequency of occurrence of various molecular descriptors [7]. The predictive capacity of BQ is not interpolative because data fitting is not used, but it is based on generalizations substantiated by experimental data [13,14]. As for the Bayesian classifier, it is called naive Bayesian (NB) because it naively assumes the descriptors in the training set are independent and equally important [37,38]. In the case of HTS experiments, *A* might refer to whether a given compound will show a desired biological activity, while *B* might refer to molecular descriptors [26]. Owing to these assumptions, the joint probability is obtained by multiplying individual probabilities. In practice, these two assumptions are often violated and the full independence of descriptors is rarely observed [7]. The distinction between BQ and NB approaches is the addition of a de-correlation step to perform a principal component analysis used as an approximation for statistical independence and not dimension reduction [14,39] which makes BQ robust and tolerant to noise. The assessment of NB noise tolerance has also been addressed by Glick *et al.* [7]. They showed that a statistical model built from the non de-convoluted primary data could prioritize the hits and identify the true positives. Many researchers consider NB to be one of the most efficient and effective inductive learning algorithms for machine learning and data mining. It appears that the justification for the surprisingly good performance of NB in classification arises only if the dependences distribute evenly in classes, or if the dependences cancel each other out, no matter how strong the dependences among attributes are [40]. According to Klon *et al*. [26] the simplified NB approach has been consistently shown to be comparable or even superior to other, more complicated, machine-learning techniques such as support vector machines, decision-tree induction or neural networks [26,41]. Bayesian classifiers have also been shown to be superior to logistic regression modeling [26,42]. 'However, the standard naive Bayesian formulation proved difficult to apply to high-dimensional fingerprint data, as the de-correlation step is expensive or impossible for a table that contains possibly millions of columns (one column for each unique feature in a library). Instead, the Bayesian method is modified by considering only the effect of the presence of a feature and not its absence; to correct for the differing sampling rates of different features' [26].

Currently, most of the recent *in silico* methods of triaging compounds from HTS campaigns include machine learning techniques NB and modified-NB statistical models built on molecular fingerprints [7,26,43], such as extended-connectivity fingerprints (ECFP) [9]. In this case, event A refers to the activity of a given compound, while event B refers to the presence of a certain ECFP bit [34]. These NB models include Laplacian-modified Naïve Bayesian (LMNB) classifiers [7,8] and multiple-category Bayesian models (MCBM) [12,37]. The two methods have been shown to be useful in creating models that can work well even with noisy data [7]. The derivation from LMNB models is described in reference [8]. The LMNB was developed to enrich noisy HTS data, to take advantage of the high-dimensional representation of molecules, and to address problems caused by modeling large data sets generated by HTS [9]. According to Xia *et al*. [8] the LMNB estimator is needed to adjust the uncorrected probability estimate of a feature to account for the different sampling frequencies of different features. Finally, according to Crisman *et al*. [11], a MCNB is more suitable for HTS data analysis because it can treat very frequent or very rare features, which normally either lead to overconfident or random prediction [11]. This MCNB model has also been trained by Nidhi *et al*. [37] ECFP of compounds in chemogenomics database.

Thus, Bayesian statistics use a set of descriptors such as ECFP (other descriptors can be used as well but ECFP are considered the most effective 2D molecular representation) as evidence for a given molecular structure, whereas hypotheses state whether (or not) a compound will have a particular biological activity [8]. In this context, statistical modeling with Bayesian methods tolerant to noise are employed to analyze HTS primary data to support hit prioritization, to identify of FPs and FNs, to maximize the number of chemotypes and build QSAR models. The performance reaches by Bayesian modeling on hit selection and enrichment over classical statistical modeling was demonstrated. Classical modeling approaches do not incorporate molecular information and need to rely subsequently on cheminformaticians expertise. However, the real core of the problem are differences in 'research philosophy' as it has been stated in Parker *et al*.: '... experimental and computational screeners have not necessarily the same mind set' [44]. Below, it is shown how different mindsets can impact on the productivity. This can be illustrated using simple feedback loops within which that decision-making process is imbedded.

HTS decision-making and statistical perspectives

HTS technologies focus on the discovery of small molecules as promising drug leads to be turned into new molecular entities (NMEs[6]). An NME as defined by the FDA is 'a medication containing an active substance that has never before been approved for marketing in any form in the United States' [45]. Although there are more new drug products (i.e non-NMEs) on the market than NMEs, their R&D costs are lower than NMEs because they are incremental improvements on existing drugs. The productivity gap reported over the past few years impinges on the relationship between R&D spending per NME and the total number of NMEs approved by the FDA[7] [17]. The availability of HTS data from the NIH and the Broad Institutes stimulate the research community in developing new methods in the selection and prioritization of hit compounds and lead optimization. High quality information

[6] Non-NME encompasses 'me-too' or follow-on drugs that can be subdivided into two categories: (1) they may be innovative products, that is an NME that lost the race to be the first drug on the market in a given therapeutic class (such as antidepressants, antibiotics, or antihistamines), or (2) they may be incremental modification of an existing drug, that is a new drug entity with a similar chemical structure or the same mechanism of action as that of a drug already on the market. That is, a me-too drug is a new entrant to a therapeutic class that had already been defined by a separate drug entity that was the first in the class (sometimes referred to as the breakthrough drug) to obtain regulatory approval for marketing. Me-too drugs have also been characterized in a more value-neutral way as follow-on drugs.

[7] The perception by observers of a productivity gap has led the NIH to establish the NIH Road Map in 2003, and to provide funding for a network of HTS centers aimed at identifying and at hastening further developments of modulators and chemical probes of gene, pathway and cell functions [46,47]. The network has begun to provide publicly available data on a wide variety of biological assays tested against a common small molecule library. All assay descriptions, chemical structures and data are made publicly available through the NIH's Pubchem database [48].

is mandatory to achieve the objectives of the NIH road map. During the hit selection process, and further down the pipeline, two types of information are necessary: experimental information from HTS; and chemical information from cheminformatics-based profiling with descriptors or fingerprints as entailed by Bayesian methods.

However, to date, little is known about the appropriate selection and uses of statistical and cheminformatics methods for analyzing HTS data. Interestingly, both statistical and cheminformatics methods are increasingly intertwined into the HTS process as described in previous sections. Hit selection and/or profiling with cheminformatics methods highlight the similarities and differences between the two statistical perspectives.

High quality and high content information on primary screening data, counter screen and confirmation screen are essential for cheminformaticians and computational chemists analyzing HTS data to select promising hits for further consideration**.** The role of the theoretical chemists is to mine the wealth of biological and chemical activities for the development of QSAR hypotheses [46], to further lead optimization with the aim of developing NMEs on the basis of hits obtained from HTS campaigns, and in parallel, to eliminate unproductive compounds (i.e. those falsely identified as positive hits) [43]. Parallel to the integration of HTS technologies in the DDP, major advances in data integration, statistical analysis, hit and leads profiling with cheminformatics methods are becoming standards and a driving force for lowering the DDP productivity gap.

Expectations associated with an increased productivity include the value improvement of experimental and theoretical data. Hits selection and profiling embedded in a reinforcing feedback loop, virtuous cycle, enhance the theoretical understanding of chemical and biological space and also help the decision-making process in selecting chemotypes and individual compounds with, over time, greater and more favorable physicochemical properties, absence of toxophores and that are amenable to the rapid synthesis of chemical analogues. The implications of the interrelationship between technology and theory-driven hypotheses are well-understood using a conceptual representation of the dynamic decision-making process in research (see Figure 1) [49]. This could well be accomplished with cheminformatics methods on the basis of increasingly accurate and refined Bayesian statistics as high value content increases.

The answers are tied to practices and statistical methods and perspectives adopted by scientists and technologists *qua* decision-makers because information only arises in relationship with the technology employed in measurement, the measured object and methods of analysis. The disjunction between basic science and technologies has driven the emergence of new strategies to support the DDP through the integration of systems and data analysis [50].

Information feedback from *in silico* and experimental data do not only alter decisions within the context of the existing frames and applied decision rules. The structure of the DDP changes with the newer assimilation of statistical practice, creating different decision rules and new strategies. It follows that the same information processed and interpreted by an alternative statistical model leads to a set of alternative decisions on the basis of the approach employed.

The prioritization of resources (including the identification and termination of underperforming projects sooner) and effective decision-making between phases represent main steps in developing a drug. The DDP's dynamic process represents the interplay between *in silico* and experimentation that produces data that create novel practices, which, in turn, impact changes on the development of a new set of rules and strategies for the design of new experiments applied to the next iteration until convergence ensue between the physical and virtual worlds. This translates into reduced time-to-market for NMEs and lower financial burden by rejecting unproductive compounds upstream of the DDP. It becomes clearer that experimentalists prefer to use for HTS hit selection classical statistical modeling such as those derived from the frequentist approach, whereas cheminformaticians prefer Bayesian approaches. The reason for these preferences is clearly stated by Parker *et al.* [44,51]:'... the integration of experimental and computational methods is currently much less advanced than one might think (or perhaps wish for) despite these obvious synergies. Clearly, there are some technical hurdles to overcome, but equally, if not more, relevant for the current situation are differences in mind sets between experimental and computational screeners that are, at least in part, supported by different reward structures'.

Conclusion

The aim of this article was to stress the importance of examining the introduction and use of alternative statistical perspectives in the HTS process. This is important within the broader context of improving the performance of the DDP. Indeed, the inability to improve assessment and prediction of compound safety leads to failures and to an increased productivity gap with hundreds of millions of dollars spent, given the cost to develop NMEs falls well within the US$800 million to US$1.3 billion range. Over US$150 million of that amount are sunk into drug R&D and experiment failures, with a clinical trial failure rate of over 80%. This implies that some US$12 billion is annually spent worldwide on failed product development. Thus, any reductions in failure rates that can be generated before clinical trials can yield substantial savings. To predict which drugs will fail in development by just 10% could save US$100 million in R&D costs for the average drug.

One of the major issues in the DDP is to identify lead drug candidates from HTS campaigns, avoid unnecessary costs and resource wastage and bring new NMEs to the market as rapidly as possible. At a time when computers allow for the production and treatment of massive amounts of data, the task at hand in the DDP is no less complex than before. There is a strong possibility that Frenqentist and Bayesian analyses can contribute in a useful way. Figuring out ways to minimize the number of FPs and FNs is an important step in this direction, to flush out failed compounds that have yet to be properly identified early from the DDP. Scientists, as decision-makers, endowed with bounded rationality confront that difficult challenge. The decision set to the treatment of data is one big and important piece of the puzzle that drives the success of HTS, and where lies the importance of resolving the contribution of both the Bayesian and the Frequentist underpinnings in a meaningful fashion. It is important to recall that connected to that piece is the set of biological and chemical hypotheses, which form the basis of the knowledge stock. But

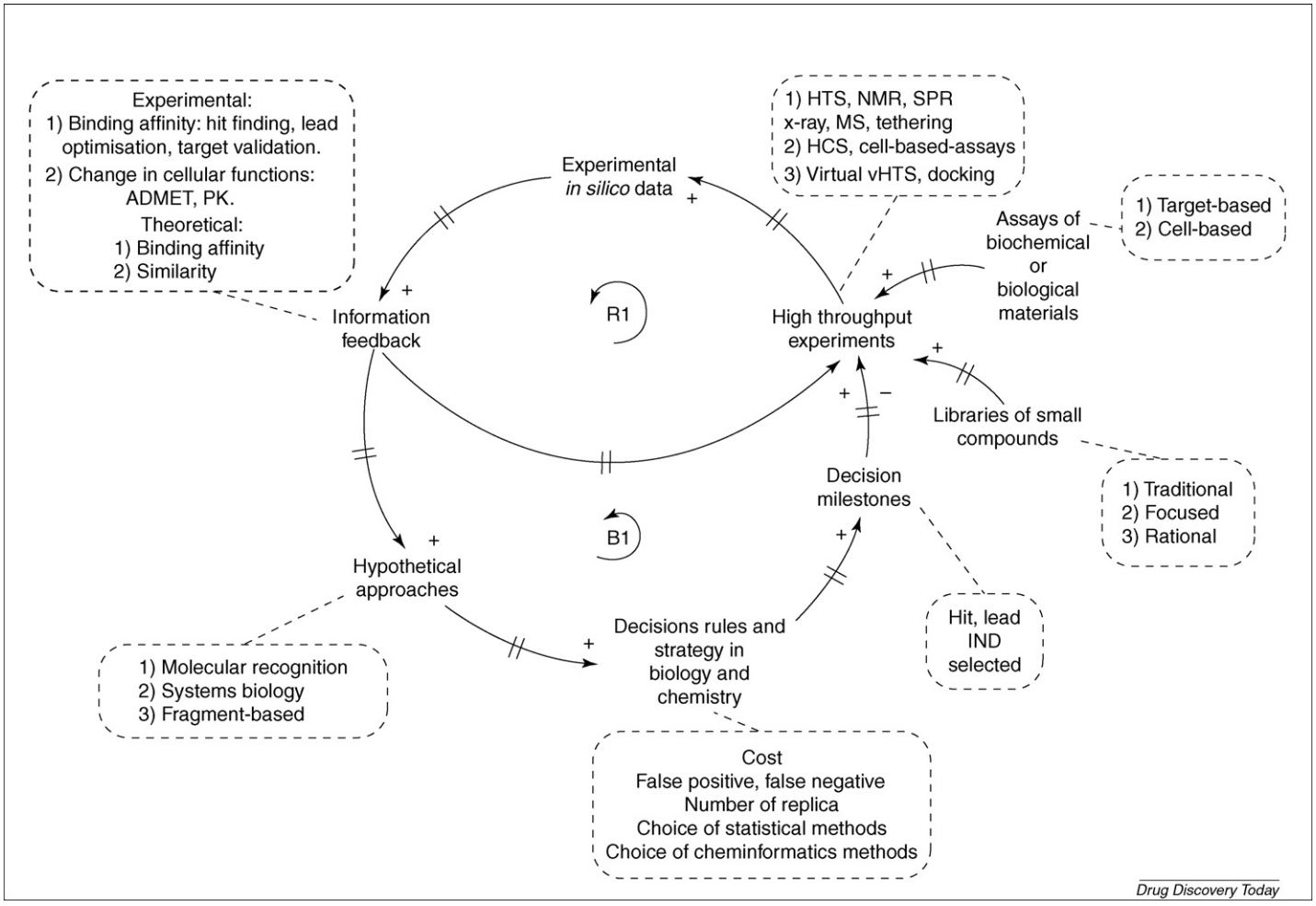

FIGURE 1

Influence diagram (ID) of information feedback from *in silico* and experimental screening data and the decision maker hypothesis formulation. The ID shows a simplified structure with key influences facing scientists and technologists in the DDP. There are three main sets of interrelationships represented within the structure. As part of reinforcing feedback loop R1, experimental and *in silico* data provide information feedback. This information influences hypothetical approaches in biology and chemistry. These approaches are framed as the decision rules employed by biologists and chemists (represented in the balancing loop B1). As seen, these have strong data implications in terms of costs, analytical outcomes and choice of statistical methods. Indeed, their outcomes influence decision milestones.

this stock cannot be forced into useful NMEs, at the end of the day, if the information produces more heat than light.

Acknowledgements

We would like to thank the four reviewers for insightful comments and suggestions on the manuscript of this article. Drs Robert Nadon (McGill University) and Christopher Williams (Chemical Computing Group, Inc.) are gratefully acknowledged for helpful discussions about some of the statistical issues discussed in this article. Any error and limitation in this article are the sole responsibility of the authors.

References

1 Eflon, B. (2003) Bayesians, Frequentists, and Physicists. In *PHYSTAT 2003*, SLAC
2 Malo, N. *et al.* (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* 24, 167–175
3 Zhang, J.H. *et al.* (1999) A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* 4, 67–73
4 Brideau, C. *et al.* (2003) Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* 8, 634–647
5 Kevorkov, D. and Makarenkov, V. (2005) Statistical analysis of systematic errors in high-throughput screening. *J. Biomol. Screen.* 10, 557–567
6 Zhang, J.H. *et al.* (2000) Confirmation of primary active substances from high throughput screening of chemical and biological populations: a statistical approach and practical considerations. *J. Comb. Chem.* 2, 258–265
7 Glick, M. *et al.* (2004) Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J. Biomol. Screen.* 9, 32–36
8 Xia, X. *et al.* (2004) Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* 47, 4463–4470
9 Rogers, D. *et al.* (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* 10, 682–686
10 Klon, A.E. and Diller, D.J. (2007) Library fingerprints: a novel approach to the screening of virtual libraries. *J. Chem. Inf. Model* 47, 1354–1365
11 Crisman, T.J. *et al.* (2007) Plate cherry picking': a novel semi-sequential screening paradigm for cheaper, faster, information-rich compound selection. *J. Biomol. Screen.* 12, 320–327

12 Crisman, T.J. *et al.* (2007) Understanding false positives in reporter gene assays: *in silico* chemogenomics approaches to prioritize cell-based HTS data. *J. Chem. Inf. Model* 47, 1319–1327

13 Labute, P. (1999) Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.* 444–455

14 Labute, P. *et al.* (2002) A probabilistic approach to high throughput drug discovery. *Comb. Chem. High Throughput Screen.* 5, 135–145

15 Nightingale, P. and Martin, P. (2004) The myth of the biotech revolution. *Trends Biotechnol.* 22, 564–569

16 Hopkins, M.M. *et al.* (2007) The myth of the biotech revolution: An assessment of technological, clinical and organisational change. *Res. Policy* 36, 566–589

17 FDA. (2007) Food and Drug Administration: times for priority and standard NMEs and new BLAs calendar years 1993–2006, http://www.fda.gov/cder/rdmt/

18 FDA. (2004) Food and Drug Administration: challenge and opportunity on the critical path to new medical products, http://www.fda.gov/oc/initiatives/criticalpath/

19 Sams-Dodd, F. (2007) Research & market strategy: how choice of drug discovery approach can affect market position. *Drug Discov. Today* 12, 314–318

20 Schmid, E.F. and Smith, D.A. (2006) R&D technology investments: misguided and expensive or a better way to discover medicines? *Drug Discov. Today* 11, 775–784

21 Sams-Dodd, F. (2005) Target-based drug discovery: is something wrong? *Drug Discov. Today* 10, 139–147

22 Sams-Dodd, F. (2006) Drug discovery: selecting the optimal approach. *Drug Discov. Today* 11, 465–472

23 Miron, M. and Nadon, R. (2006) Inferential literacy for experimental high-throughput biology. *Trends Genet* 22, 84–89

24 Rishton, G.M. (2003) Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today* 8, 86–96

25 Diller, D.J. and Hobbs, D.W. (2004) Deriving knowledge through data mining high-throughput screening data. *J. Med. Chem.* 47, 6373–6383

26 Klon, A.E. *et al.* (2006) Improved naive Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *J. Chem. Inf. Model.* 46, 1945–1956

27 Bayes, T. (1763) Essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* 53, 370–418

28 Glick, M. *et al.* (2006) Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J. Chem. Inf. Model.* 46, 193–200

29 Bender, A. *et al.* (2004) Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* 44, 170–178

30 Bender, A. *et al.* (2004) Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT 3D). *J. Med. Chem.* 47, 6569–6583

31 Yoon, S. *et al.* (2005) Surrogate docking: structure-based virtual screening at high throughput speed. *J. Comput. Aided. Mol. Des.* 19, 483–497

32 Chen, B. *et al.* (2007) Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput. Aided. Mol. Des.* 21, 53–62

33 Vogt, M. and Bajorath, J. (2008) Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and molecular fingerprints. *Chem. Biol. Drug Des.* 71, 8–14

34 Klon, A.E. *et al.* (2004) Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results. *J. Med. Chem.* 47, 2743–2749

35 Klon, A.E. *et al.* (2004) Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J. Med. Chem.* 47, 4356–4359

36 Klon, A.E. *et al.* (2004) Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease. *J. Chem. Inf. Comput. Sci.* 44, 2216–2224

37 Nidhi, *et al.* (2006) Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. *J. Chem. Inform. Model.* 46, 1124–1133

38 Plewczynski, D. *et al.* (2006) Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* 46, 1098–1106

39 Glen, W.D. *et al.* (1989) Principal components analysis and partial least squares regression. *Tetrahedron. Comput. Methodol.* 2, 349–376

40 Zhang, J.H. (2004) The optimality of Naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004* (Vol. 2), pp. 562–567

41 John, G.H. and Langley, P. (1995) Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (Besnard, P. and Hanks, S., eds), Morgan Kaufmann

42 Ng, A.Y. and Jordan, M.I. (2002) MIT Press, Cambridge, Vol. 14

43 Davies, J.W. *et al.* (2006) Streamlining lead discovery by aligning *in silico* and high-throughput screening. *Curr. Opin. Chem. Biol.* 10, 343–351

44 Parker, C.N. *et al.* (2006) Measure, mine, model, and manipulate: the future for HTS and chemoinformatics? *Drug Discov. Today* 11, 863–865

45 FDA. (2001) www.fda.gov/cder/reports/reviewtimes/default.htm

46 Agrafiotis, D.K. *et al.* (2007) Recent advances in chemoinformatics. *J. Chem. Inf. Model.* 47, 1279–1293

47 Austin, C.P. *et al.* (2004) NIH molecular libraries initiative. *Science* 306, 1138–1139

48 Wheeler, D.L. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic. Acids Res.* 34 (Database Issue), D173–D180

49 Sterman, J.D. (2000) *Business Dynamics: Systems Thinking and Modeling for a Complex World.* Irwin/McGraw-Hill

50 Waller, C.L. *et al.* (2007) Strategies to support drug discovery through integration of systems and data. *Drug Discov. Today.* 12, 634–639

51 Parker, C.N. and Bajorath, J. (2006) Towards unified compound screening strategies: a critical evaluation of error sources in experimental and virtual high-throughput screening. *QSAR Comb. Sci.* 25, 1153–1161