J Comput Aided Mol Des (2012) 26:107–112
DOI 10.1007/s10822-011-9535-9

PERSPECTIVE

Of possible cheminformatics futures

Tudor I. Oprea · Olivier Taboureau · Cristian G. Bologa

Received: 12 December 2011 / Accepted: 14 December 2011 / Published online: 30 December 2011

Abstract For over a decade, cheminformatics has contributed to a wide array of scientific tasks from analytical chemistry and biochemistry to pharmacology and drug discovery; and although its contributions to decision making are recognized, the challenge is how it would contribute to faster development of novel, better products. Here we address the future of cheminformatics with primary focus on innovation. Cheminformatics developers often need to choose between "mainstream" (i.e., accepted, expected) and novel, leading-edge tools, with an increasing trend for open science. Possible futures for cheminformatics include the worst case scenario (lack of funding, no creative usage), as well as the best case scenario (complete integration, from systems biology to virtual physiology). As "-omics" technologies advance, and computer hardware improves, compounds will no longer be profiled at the molecular level, but also in terms of genetic and clinical effects. Among potentially novel tools, we anticipate machine learning models based on free text processing, an increased performance in environmental cheminformatics, significant decision-making support, as well as the emergence of robot scientists conducting automated drug discovery research. Furthermore, cheminformatics is anticipated to expand the frontiers of knowledge and evolve in an open-ended, extensible manner, allowing us to explore multiple research scenarios in order to avoid epistemological "local information minimum trap".

Keywords Machine learning · Drug discovery · Data mining · Semantic web technologies · Forecast support · Decision-making tools

T. I. Oprea (✉) · C. G. Bologa
Division of Biocomputing, Department of Biochemistry and Molecular Biology, University of New Mexico School of Medicine, MSC11 6145, Albuquerque, NM 87131, USA
e-mail: toprea@salud.unm.edu

T. I. Oprea
Sunset Molecular Discovery LLC, 1704 B Llano Street, Suite 324, Santa Fe, NM 87505, USA

T. I. Oprea · O. Taboureau
Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Kemitorvet, Building 208, 2800 Lyngby, Denmark

Introduction

In his 1669 opus, *Ars Magna Sciendi,* the seven-teenth century scientist (and alchemist) Athanasius Kircher [1] explored combinatorics in a non-mathematical way, by devising an "universal diagram for the formation of questions about every possible subject" (Fig. 1, retrieved from [2]). Fast forward about 330 years, and we're in the process of combinatorial exploration of large numbers of chemicals as devices to query proteins. Indeed, series of small molecules binding to a macromolecule represent a chemist's way to interrogate that target and extract information about its binding site(s) and mode of action. The demand for computation in chemistry informatics increased by three to five orders of magnitude between the late 1980s and late 1990s, which makes combinatorial chemistry and high throughput screening the strongest driving forces between increased innovation and funding in cheminformatics [3]. This was paralleled by significant improvements in computational power, algorithmic developments, and increased inter-personal communication (fueled by the internet). The magnitude and complexity of these changes are likely to render today (late 2011) any

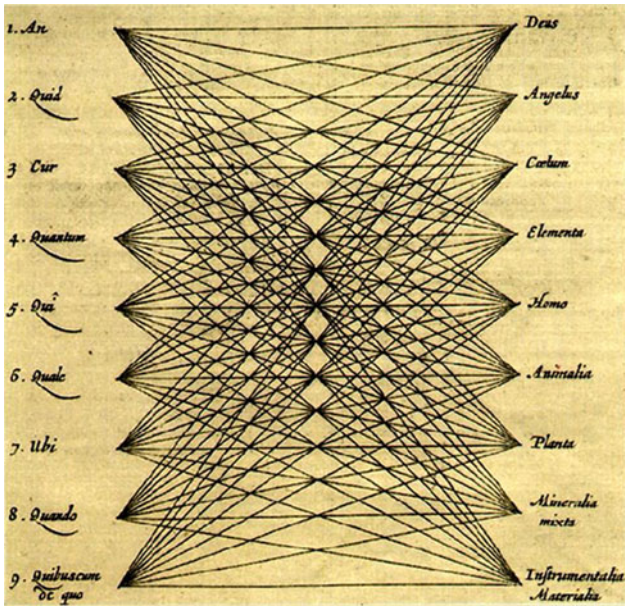

Fig. 1 Combinatorial diagram from *Artis Magna Sciendi*, Chapter IV, Amsterdam 1669, by Athanasius Kircher, page 18. The text of the diagram (in Latin) is "*Typus universalis, omnibus de quacunque re proposita questionibus formandis, aptus*" [2]

predictions about the future of computer-aided molecular design made in 1985, if such predictions are on record, seem naïve.

Therefore, any discussion about the future of cheminformatics or the next 25 years of computer-aided molecular design has to take into consideration the multiplicity of driving forces behind innovation. Certain aspects, such as societal and economic factors, as well as hardware, computer science and nanoscience developments, are beyond the topic of this discussion. In the same way, we will not address the growing role of chemistry informatics in education, since the use of computers in chemistry, computer-based chemical structure manipulation and chemical synthesis planning via computers are becoming ubiquitous. Yet, the above factors are just as likely to shape the outcome as those that we, as community, are more likely to influence: the evolution of public and private sectors that are likely to demand chemistry services based on computational and algorithmic technologies, as well as our ability to influence the products emerging from such activities. So before envisioning the future, we will address innovation and its lasting role on the impact of cheminformatics on society.

Predictions are difficult, especially about the future

Take the above quote, for example. Although widely attributed to Niels Bohr, there are no less than 25 authors and four anonymous sources to which the quote has been attributed, from Confucius to Yogi Berra, and including Albert Einstein [4]. This illustrates how predictions can be difficult, even when they concern the past; more to the subject, they illustrate how (good) ideas are often likely to follow a diffuse path, one that makes it difficult to observe the "first author" principle. Even more to the point, most software tools in cheminformatics, and indeed innovation per se, are likely to follow the same algorithms (e.g., in calculating partial charges or chemical similarity), observe similar workflows (e.g., 1D database queries followed by structure–activity analyses and selection of novel candidates for experimental testing), and obey similar principles (e.g., natural products vs. "synthetics", the quest for novelty and intellectual property potential).

Though marketed by several, highly innovative companies, these tools are far from being on a divergent path. Rather, there is a significant cross-pollination trend in cheminformatics, for example the ability to compare molecules using molecular interaction fields and shape similarity, or the use of linear notation languages and derivative features to represent chemical structures and reaction queries are features present in several packages. *It is safe to state that the cheminformatics community observes the "best-in-class", rather than the "first-in-class" (first-author) principle, albeit confounded by the "free-to-use" vs. "for-fee" aspect.* Given the choice, the community is likely to migrate towards fast-and-efficient, open access cheminformatics tools—situation that is occasionally challenged by the existence of granted patents and by the emergence of completely novel, unique technologies.

Beyond inter-operability, data formatting and controlled vocabularies, this discussion on innovation centers on the ability of cheminformatics software to perform a highly diverse set of tasks. From analytical chemistry and biochemistry to immunology and toxicology, not to mention pharmacology and drug discovery, software designed for computer-aided molecular design needs to address a multitude of tasks that range from machine learning, database storage, archive and retrieval, to virtual screening, quantum mechanics, and biopolymer modeling—to mention a few (exhaustively covered [5]). Between profit and innovation, wide-spread vs. elitist usage, cheminformatics companies often walk a fine line amid "mainstream" (i.e., accepted, expected) and novel, leading-edge tools. While such innovation is observed in the academic sector, and often contributed as open-source, open-access software, there is often not enough reliability and stability within strictly academic software—although incentives and funding in this area are increasing. This latter aspect alone, availability of funding (or lack thereof) is potentially the single most restrictive factor to influence the future of cheminformatics.

Of possible futures

To speculate about the future of science in the current economic climate is to embrace a truly positive vision of the future, one that ignores the perils of diminishing funding across the entire industrial sector that employs cheminformatics practitioners, as well as the lack of clear-cut funding sources for cheminformatics at the U.S. National Institutes of Health and the E.U. Framework Programmes, to name a few.

Worst case scenario

Cheminformatics software will witness limited improvement over the next decades, and no disruptive, truly innovative events are in store. Related to this scenario, there is also the possibility that software and scientific tools develop in truly novel and unexpected directions, but the workforce that practices cheminformatics gradually diminishes, and is reduced to an unimaginative role.

Best case scenario

High value compound design is implemented in major industrial sectors, from agriculture to flavor and fragrance, from pharmaceutical to material science. As data accumulates and translates into knowledge, this in itself is likely to fuel the discovery of novel tools for visualization, integration, processing and knowledge. The roots of such a scenario rely on systems chemical biology [6], which aims to develop tools for integrated chemical–biological data acquisition, filtering and processing, by taking into account relevant information related to interactions between proteins and small molecules, possible metabolic transformations of small molecules, as well as associated information related to genes, networks, small molecules, and, where applicable, mutants and variants of those proteins [7]. Within the systems chemical biology framework, the user is expected to dynamically monitor the effects of perturbations on networked systems (pathways), whereby biochemical and pharmacological events are altered (and monitored) over time. Bioactivity and target integration [8], coupled with pathway simulations for biochemical and pharmacological processes and knowledge-based association of drug side effects to novel targets [9, 10] will enable the extension of systems chemical biology from cellular to organ and organism processes, whereby these tools will be integrated into the virtual physiological human [11]. Advances in proteomics, metabolomics, metagenomics and other –omics sciences, combined with "next generation" sequencing [12] are expected to benefit from cheminformatics support, quite likely combined with bioinformatics, computational chemical biology and other computational techniques. We will no longer evaluate the bioactivity profile of a chemical at the molecular level, but rather we will investigate biomedical knowledge with the integration of genetic polymorphisms [13] and clinical effects [14].

Tools we would like today, but are willing to wait for

On computational speed

By late 2012, the 64-bit GPUs from Nvidia will supersede current CPUs in terms of computing power by as much as two orders of magnitude. More cheminformatics software is being ported on the GPU platform, and our ability to perform cheminformatics research on highly-accurate electronic densities (pre-calculated with quantum chemistry software) will dramatically improve. Some of these GPU technologies are likely to result in marked improvements over current pharmacophore and fingerprint technologies. As quantum computing technologies evolve, the speed of computation is likely to improve even further. It is conceivable that "quantum fingerprints" or perhaps "dynamic quantum fingerprints" (taking into account relativistic effects, as well as multiple tautomeric and protomeric states) will greatly outperform existing (and near-future) technologies, for a net improvement in computing accuracy and precision.

Free text derived machine learning models

For now, it is rather difficult to process large bodies of free text (e.g., scientific publications, approved drug labels, blogs) and extract information that is relevant and pertinent to molecular discovery. Such technologies, presumably combining multiple controlled vocabularies and ontologies for e.g., biological assays [15] and chemical-biological interactions [16] in a semantic-web manner [17], are likely to enable machine learning processing and statistical model generation much the same way as we currently obtain from large numerical datasets. These technologies, likely to rely on human curation [18] as well as automation are expected to enable novel associations and conclusions. An EU-funded initiative, OpenPHACTS [19], is already exploring some of these aspects.

Environmental cheminformatics

It is of current interest for the society at large to examine the link between environmental exposure and human health. In the modern society, man is exposed to a large array of environmental chemicals, inherently present in food, cosmetics, medicines, as well as pollutants present in

air, water and foodstuff, among others. Although many of these chemicals are present at a low concentration (i.e., below the threshold of toxicological concern in risk assessment), their combination can potentially affect human health and explain non-inherent genetic diseases such as infertility [20], metabolic disorders [21] and lung cancer [22]. Perhaps using similar approaches to the prediction of drug–drug interactions [23], more general "chemical–chemical interaction" tools (here referring to their effects in living organisms, not to chemistry and physics) would be extremely valuable in risk assessment. One challenge for such studies is the integration of quantitative data monitored cautiously taking into account the concentration of chemicals but also time of exposure and period of life [24]. The field of predictive environmental cheminformatics is likely to grow within a collaborative framework [25], since current REACH legislation (regulation on registration, evaluation, authorization and restriction of chemicals [26]) mandates the in silico evaluation of chemicals within the EU. If successful, this type of regulation is likely to be mandated by other countries.

Forecast support

Extending the role of machine learning and chemistry-based informational systems, such technologies might become ubiquitous in our effort to reduce the synthesis cost of making approved drugs (i.e., computer-assisted chemical production) and to reduce the environmental impact of chemical reagents; they may assist with optimizing the selection of active pharmaceutical ingredients during the process of formulating new drug combinations; indeed, our capability to handle mixtures and perform cheminformatics at the "fixed" and "unspecified" mixture level is likely to evolve. Such forecast support systems could be used in evidence-based medicine, to evaluate hazard, environmental and chemical warfare agents, and are likely to assist decision makers to reduce risk of exposure, support and improve legislative and societal demands.

The automation of drug discovery

Robot scientists [27] will use computer-aided molecular and synthesis design, coupled for example with flow-based micro-reactor systems, plan and conduct biomolecular screening and toxicity experiments, to identify novel, safe chemicals—no doubt supported by systems chemical biology within the virtual human context. These robots would also process the medicinal chemistry and toxicology literature corpus, seeking supplementary evidence to prioritize the "most likely" chemicals, and propose these for proof-of-concept (first-in-man) experiments.

Personalized drug discovery

In this embodiment, robot scientists would process "23andMe" [28] genomic information for particular individuals, comb through the biomedical literature corpus as well as other on-line and for-fee resources (e.g., evidence-based medicine databases, clinical trial and drug label information, etc.), to find the most appropriate pathways that are susceptible to modulation via small molecules, then conduct cheminformatics experiments and identify the optimal combination of approved drugs (i.e., "drug repurposing" [29]) for any given ailment. Cheminformatics can become a valuable tool in integrating personalized medicine [30] with systems medicine [31] at the chemical level.

Open-ended science

In an aptly named book, Horgan [32] postulates, quite convincingly, that we are currently witnessing the end of science: the frontiers of knowledge are harder to reach, and according to Horgan most scientific disciplines will become stagnant as their knowledge domain expands from the "known known" to completely cover the "unknown unknown" (discussed by Taleb [33]; see also Fig. 2). To a great extent, "production cheminformatics" activities have become routine (e.g., running 2D-based similarity queries over the internet yields a number of hits for many input molecules). However, "research cheminformatics" will continue to expand, and will grow over the next decades. At the edge of knowledge, we anticipate that cheminformatics is one of several technologies that can assist with the expansion of the "bubble of knowledge" (Fig. 2). For the sake of brevity, we illustrate the "chemical-target" concept; however, as the frontiers of knowledge as defined by scientific activity expand, its exact meaning is likely to evolve. Since computer-based systems have, de facto, become the universal scientific support method and, implicitly, the main operational tool for chemistry, cheminformatics remains an active, open-ended area of science that is currently in expansion.

Faced with massive amounts of molecular property and activity data, the average scientist involved in molecular discovery is often confronted with the difficult choice of making a priori some assumptions about the data, which often implies that the research outcome is inherently related to the quality and type of data input. Akin to the "local minimum trap" where one seeks the global minimum on an energy landscape, this epistemological problem (if not properly tested, any theory becomes a limitation of what can be discovered), was recognized by Popper [34] among others. It is supported by anecdotal behavior: Most occasional users do not bother to scroll down past the top

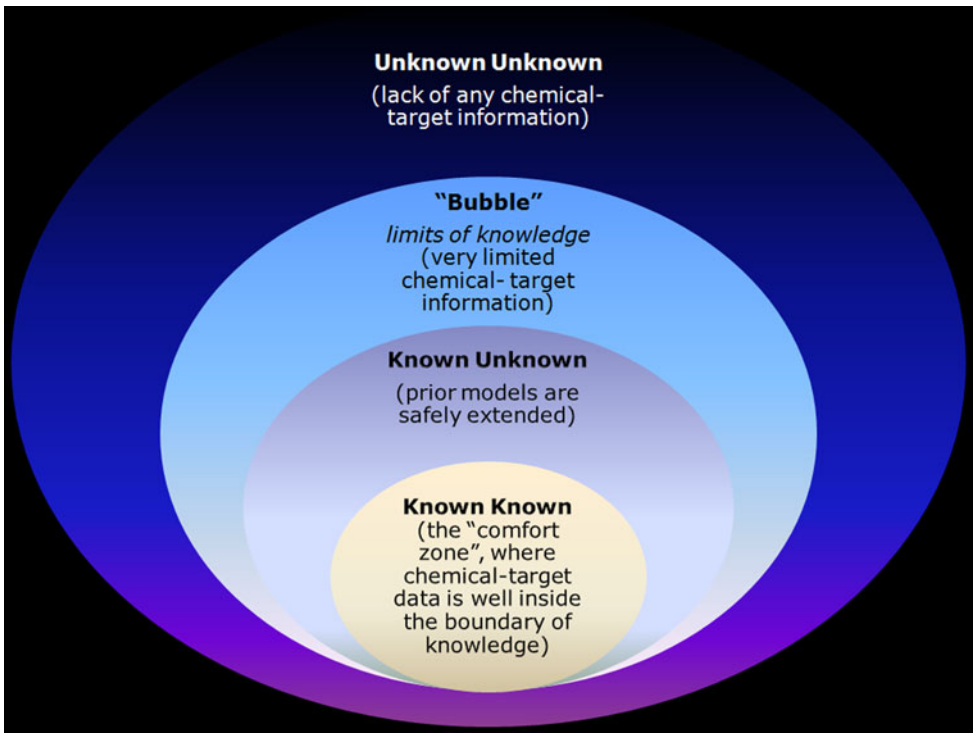

Fig. 2 Cheminformatics and the edge of knowledge. In this diagram, "target" includes all properties of potential interest (agro-chemicals, flavor and fragrance, pharmaceutical, material science, etc.)

twenty hits offered by search engines on the Internet. Likewise, *inexperienced scientists are less likely to thoroughly investigate multiple, diverging options while conducting biomedical research*. Such epistemological "local information minimum trap" situations are likely to occur when mining complex data. Using an integrative approach based on knowledge and scaffold-based mining, we are developing tools [35] within the Cytoscape [36] framework that will enable end-users to build bioactivity-chemical pattern networks, share them via a public setting (e.g., "crowdsourcing" [37]). This, and similar tools, will make possible the exploration of alternate hypotheses and multiple scenarios, given the constraints of available, as well as high-confidence prediction, data and models. Thus, the possible futures of cheminformatics are many, and their evolution is likely to help us avoid local epistemological information minima.

Acknowledgments This work was supported in part by NIH grants GM-095952, MH-084690, and CA-118100 (TIO, CGB), and by the Villum Foundation CDSB (TIO). This work was supported by the Innovative Medicines Initiative Joint Undertaking (OpenPHACTS).

References

1. Retrieved from Wikipedia, http://en.wikipedia.org/wiki/Athanasius_Kircher. Accessed 8 Dec 2011
2. Retrieved from ECHO (European Cultural Heritage Online), http://echo.mpiwg-berlin.mpg.de/ECHOdocuViewfull?pn=42&ws=2.5&cont=0..6&start=41&viewMode=images&mode=imagepath&url=/mpiwg/online/permanent/library/ZYCRR6CN/pageimg. Accessed 8 Dec 2011
3. Oprea TI (2002) Molecules 7:51
4. Retrieved from "Who first said "It is difficult to make predictions, especially about the future" (or one of its many variants)?" http://www.larry.denenberg.com/predictions.html. Accessed 8 Dec 2011
5. Gasteiger J (ed) (2003) Handbook of Chemoinform. Wiley-VCH, Weinheim, 1871 p
6. Oprea TI, Tropsha A, Faulon JL, Rintoul MD (2007) Nat Chem Biol 3:447
7. Oprea TI, May EE, Leitão A, Tropsha A (2011) Computational systems chemical biology. In: Bajorath J (ed) Chemoinformatics and computational chemical biology, methods in molecular biology, vol 672. Springer, Berlin, pp 459–488
8. Oprea TI, Tropsha A (2006) Drug Discov Today Technol 3:357
9. Mestres J, Seifert SA, Oprea TI (2011) Clin Pharmacol Ther 90:662
10. Abernethy DR, Bai JPF, Burkhart K, Xie HG, Zhichkin P (2011) Clin Pharmacol Ther 90:645
11. Fenner JW, Brook B, Clapworthy G, Coveney PV, Feipel V, Gregersen H, Hose DR, Kohl P, Lawford P, McCormack KM, Pinney D, Thomas SR, Van Sint Jan S, Waters S, Viceconti M (2008) Philos Trans A Math Phys Eng Sci 366:2979
12. Schuster SC (2008) Nat Methods 5:16
13. Wang L, Khankhanian P, Baranzini SE, Mousavi P (2011) BMC Bioinform 12:380
14. Oprea TI, Nielsen SK, Ursu O, Yang JJ, Taboureau O, Mathias SL, Kouskoumvekaki I, Sklar LA, Bologa CG (2011) Mol Inf 30:100
15. Abeyruwan S, Chung C, Datar N, Gayanilo F, Koleti A, Lemmon V, Mader C, Ogihara M, Puram D, Sakurai K, Smith R, Vempati U, Venkatapuram S, Visser U, Schürer S (2010) Semantic web challenge 9th international semantic web conference ISWC 1–12, http://www.cs.vu.nl/~pmika/swc/submissions/swc2010_submission_20.pdf

16. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, Wild DJ (2010) BMC Bioinform 11:255
17. Berners-Lee T, Hendler J, Lassila O (2001) Sci Am 284:34
18. Rosenbaum S (2011) Curation nation. McGraw-Hill, New York, 304 p
19. Retrieved from OpenPHACTS, http://www.openphacts.org. Accessed 8 Dec 2011
20. Krysiak-Baltyn K, Toppari J, Skakkebaek NE, Jensen TS, Virtalen HE, Schramm KW, Shen H, Variainen T, Kiviranta H, Taboureau O, Brunak S, Main KM (2010) Int J Androl 33:270
21. Ruzzin J, Petersen R, Meugnier E, Madsen L, Lock EJ, Lillefosse H, Ma T, Pesenti S, Sonne SB, Marstrand TT, Malde MK, Du ZY, Chavey C, Fajas L, Lundebye AK, Brand CL, Vidal H, Kristiansen K, Frøyland L (2010) Environ Health Perspect 118:465
22. Kovakic P, Somanathan R (2009) Rev Environ Contam Toxicol 201:41
23. Shardlow CE, Generaux GT, MacLaunchlin CC, Pons N, Skordos KW, Bloomer JC (2011) Drug Metab Dispos 39:2076
24. Wigle DT, Arbuckle TE, Walker M, Wade MG, Liu S, Krewski D (2007) J Toxicol Environ Health 10:3
25. Hardy B, Douglas N, Helma C, Rautenberg M, Jeliazkova N, Jeliazkov V, Nikolova I, Benigni R, Tcheremenskaia O, Kramer S, Girschick T, Buchwald F, Wicker J, Karwath A, Gütlein M, Maunz A, Sarimveis H, Melagraki G, Afantitis A, Sopasakis P, Gallagher D, Poroikov V, Filimonov D, Zakharov A, Lagunin A, Gloriozova T, Novikov S, Skvortsova N, Druzhilovsky D, Chawla S, Ghosh I, Ray S, Patel H, Escher S (2010) J Cheminform 2:7
26. Retrieved from the REACH website, http://ec.europa.eu/enterprise/sectors/chemicals/reach/index_en.htm. Accessed 8 Dec 2011
27. King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, Sparkes A, Whelan KE, Clare A (2009) Science 324:85
28. Retrieved from https://www.23andme.com. Accessed 8 Dec 2011
29. Oprea TI, Bauman JE, Bologa CG, Buranda T, Chigaev A, Edwards BS, Jarvik JW, Gresham HD, Haynes MK, Hjelle B, Hromas R, Hudson L, Mackenzie DA, Muller CY, Reed JC, Simons PC, Smagley Y, Strouse J, Surviladze Z, Thompson T, Ursu O, Waller A, Wandinger-Ness A, Winter SS, Wu Y, Young SM, Larson RS, Willman CL, Sklar LA (2011) Drug Discov Today: Therap Strategies, doi:10.1016/j.ddstr.2011.10.002
30. Gibson WM (1971) Can Fam Physician 17:29
31. Auffray C, Chen Z, Hood L (2009) Genome Med 1:2
32. Horgan J (1997) The end of science: facing the limits of science in the twilight of the scientific age. Broadway Books, New York, 322 p
33. Taleb NN (2007) The black swan. Random House, New York, 400 p
34. Popper KR (1935) The logic of scientific discovery. Routledge Classics (reprinted), New York, 510 p
35. Retrieved from http://hsc.unm.edu/som/biocomputing/carlsbad. Accessed 8 Dec 2011
36. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Genome Res 13:2498
37. Oprea TI, Bologa CG, Boyer S, Curpan RF, Glen RC, Hopkins AL, Lipinski CA, Marshall GR, Martin YC, Ostopovici-Halip L, Rishton G, Ursu O, Vaz RJ, Waller CL, Waldmann H, Sklar LA (2009) Nat Chem Biol 5:441