Published in final edited form as:
Trends Biotechnol. 2010 April ; 28(4): 161–170.

Novel opportunities for computational biology and sociology in drug discovery☆

Lixia Yao[1], **James A. Evans**[2,3], and **Andrey Rzhetsky**[3,4]

[1]Department of Biomedical Informatics, Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA

[2]Department of Sociology, University of Chicago, Chicago, IL 60637, USA

[3]Computation Institute, University of Chicago, Chicago, IL 60637, USA

[4]Department of Medicine, Department of Human Genetics, Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL 60637, USA

Abstract

Current drug discovery is impossible without sophisticated modeling and computation. In this review we outline previous advances in computational biology and, by tracing the steps involved in pharmaceutical development, explore a range of novel, high-value opportunities for computational innovation in modeling the biological process of disease and the social process of drug discovery. These opportunities include text mining for new drug leads, modeling molecular pathways and predicting the efficacy of drug cocktails, analyzing genetic overlap between diseases and predicting alternative drug use. Computation can also be used to model research teams and innovative regions and to estimate the value of academy–industry links for scientific and human benefit. Attention to these opportunities could promise punctuated advance and will complement the well-established computational work on which drug discovery currently relies.

Introduction

The identification of chemical agents to enhance the human physiological state – drug discovery – involves coordination of highly complex chemical, biological and social systems and requires staggering capital investment, estimated at between $100 million and $1.7 billion per drug [1,2].

In the search for new drugs there are numerous sources of error stemming from our limited understanding of the biology of drug action and the sociology of innovation. Biologically, the bottleneck is our poor knowledge of molecular mechanisms underlying complex human phenotypes [3,4]. Socially, we lack models that accurately capture the link between successful discovery and the dynamic organization of researchers and resources that underpins it.

Computational approaches, if applied wisely, hold the potential to substantially reduce the cost of drug development by broadening the set of viable targets and by identifying novel

☆The editorial office apologizes for a production error, which has led to the misplacement of several references in the original Review article that was published in the September 2009 issue of Trends in Biotechnology. The full article with the correct references is reprinted. Trends in Biotechnology apologizes to the authors and readers for the confusion and inconvenience this error has caused

Corresponding authors: Evans, J.A. (jevans@uchicago.edu); Rzhetsky, A. (arzhetsk@medicine.bsd.uchicago.edu).

therapeutic strategies and institutional approaches to drug discovery. Here we provide an overview of what computational biology and sociology have to offer and what problems need to be solved so that these approaches can support drug discovery.

Computational biology methods for drug discovery

A number of computational methods have been successfully applied throughout the drug discovery process, from mining textual, experimental and clinical data to building network models of molecular processes, to statistical and causal analysis of promising relationships, as summarized in Figure 1 and Box 1.

Molecular modeling or *structural biology* is the most established domain of computational biology. The aim is to predict and model properties of biological targets using chemistry, quantum and bio-physics, experimental crystallography and computer science. Alongside established tools of computational chemistry, which model properties of drug leads and their interaction with targets, molecular models form the centerpiece of computational drug discovery and development.

The aim of *sequence analysis* is to efficiently compare nucleotide and amino acid sequences, thereby allowing researchers to impute a gene's function by considering evidence from homologous genes, often from different biological systems. Sequence analysis has now become an indispensable part of target identification in early-stage drug discovery (Figure 1).

Over the last several years, molecular models and sequence analysis have become productive and mature. Marginal improvements have diminished in value and there is less room for radical innovation. Thus, we focus this review on emerging computational approaches such as the simulation of molecular networks, probabilistic data integration, and development of drug cocktails. These and related approaches point to broader biological systems underlying disease, systems that extend beyond single-gene inheritance and pairwise molecular interactions. As such, they promise to extend the repertoire of molecules suitable as drugs and suggest novel therapeutic strategies likely to yield big returns in drug discovery.

Text mining (http://people.ischool.berkeley.edu/~hearst/text-mining.html) is one such emerging discipline that aims to extract phrases and statements from electronically stored texts [5,6]. This information is then assembled and analyzed to generate new knowledge [7]. Text mining techniques draw from computational linguistics, natural language processing, data mining and artificial intelligence. In the biomedical sciences, text mining is currently used to screen articles for biological terms including molecule names and biological statements such as molecular interactions.

The development of *terminologies* and *ontologies* – consistent systems of terminologically bounded statements – has enabled standardization of text-mined information so that it can be used to model wider biological processes and inspire new drug targets and strategies. Conceptual standardization also facilitates the consistent organization of experimental data. This solves a perennial problem in biomedical research: Data produced with disparate methods from distinct institutions are heterogeneous. For example, drug information provided by Medline Plus (http://www.nlm.nih.gov/medlineplus/) is aimed at the general public and is linguistically incompatible with the drug handbook used by physicians. Standardization enables the merger of diverse resources, increasing the total volume of data for analysis and the number and subtlety of patterns discoverable within it. Together, these steps constitute *knowledge engineering*. Drawing on philosophy and computer science, knowledge engineering aims to create machine-organized systems of knowledge that enable

artificial intelligence applications to make inferences that would be difficult or impossible to achieve for human scientists. Knowledge engineering relates to text mining in two distinct ways. Texts are first mined to discover terminologies such as a lexicon of gene names; these terminologies, when organized, enable more consistent extraction of relations (such as activation of one gene by another) in subsequent text searches. We illustrate this in Figure 2 with an ontology containing the most common concepts in drug discovery and exemplify it in Box 2 with a case study of the calibration of warfarin dosage. With this type of structured knowledge representation, computers can simulate the reasoning process of human scientists on a large scale. For example, the protein BCR-ABL is involved in the apoptosis pathway that results in cell death, which, in turn, is related to chronic myelogenous leukemia. Reasoning across this chain, we might hypothesize that a BCR-ABL inhibitor could cure chronic myelogenous leukemia, as confirmed by researchers with the discovery of Gleevec [8].

Data integration deals with the challenges arising from fusion of multiple data types, such as relationships established from literature mining, gene expression data from microarrays, protein–protein interactions from yeast two-hybrid experiments and even clinical patient records. Data integration involves not only an accurate ontological representation of relevant factors, but also an estimate of uncertainty and bias in the data from each source. As a field, data integration is still in its infancy, a state of ‘productive anarchy’ in which approaches abound but no universal standards exist. We believe that data integration will become increasingly important for drug discovery as available data for biosystems that are relevant to human disease increase (Figure 3).

Data mining takes these large, complex arrays of data as input and uses a range of statistical and mathematical techniques to discover unanticipated regularities. Data mining has become important in diverse commercial applications including the mining of movie review databases, credit card purchase data and stock exchange transactions. In biomedicine, the mining of databases of molecular interactions has already proven successful in suggesting novel disease-related processes [9] (note the hypothetical association between cellular process and molecular pathway in Figure 3).

Data mining has also been used to identify patterns in clinical patient records that point toward novel therapeutic interventions (Figure 3). In combination with patient-specific genomic data, such as single nucleotide polymorphisms and copy number variation, clinical data mining could lead to a better understanding of drug action in terms of the mechanisms underlying drug successes and failures. This could ultimately increase the match between patients and drugs by inspiring the development of customized drugs. There are, however, many challenges associated with the use of clinical data. Physicians are not consistent in entering data and, for billing purposes, it is acceptable to encode diseases with their ‘equivalents’, so that diabetes mellitus I is often coded diabetes mellitus II and male breast cancer as its female counterpart. Moreover, patient records are often incomplete as patients move across hospitals and regions. Even a perfect data series often does not contain other relevant patient data, such as diet and over-the-counter drug use. Concerns about patient privacy also limit access to clinical data. Nevertheless, we believe that mining medical records will become increasingly important for drug discovery and development.

There are several approaches to identify patterns within the data mining toolbox. *Pattern recognition and classification* are methods that mimic human perception of real-world phenomena, such as recognizing faces and handwritten letters, detecting contours of objects and threatening sounds and separating relevant from irrelevant text in search queries. These methods have been used in biomedicine to discover multi-molecular processes in complex databases containing biological and clinical information [10]. *Information theory* and *signal*

processing are other approaches to pattern identification. They originated in response to practical problems of communication [11] in which the analyst attempts to separate a communicated signal from the noise that masks it. For example, a sender might emit an encoded message over a noisy channel such as radio waves in the open air. A receiver records the signal, separates it from noise and decodes the message. The corresponding mathematical toolbox allows estimation of mutual information from several variables to detect non-linear dependence among them. These approaches are commonly used in computational biology and drug discovery [9,12–14] to identify pathways involving multiple genes, proteins and other molecules.

Uncovering relationships through data mining naturally reveals spurious associations derived from noise and artifacts of the method or data source. Statistical models are required to rigorously test these associations as hypotheses. Perhaps the most common family of statistical models used in drug discovery and development is *regression analysis* (see the warfarin example in Box 2). Regressions characterize the stochastic relationship between a response variable, such as drug toxicity, and input or causally independent variables, such as the presence of certain structural motifs in the drug or its molecular weight or solubility. Although regression is one of the oldest approaches to statistical modeling, it permeates virtually every stage of drug discovery, from ranking prospective targets and leads to analyzing clinical trial data.

As statistical modeling has evolved, increased care has been taken to isolate empirical measurements for separating causal from coincidental associations. Analysts have also worked to systematically compare competing causal models. *Causality analysis* is a paradigm within statistics and hypothesis testing devoted to these concerns. Its methods have been used in biomedicine to determine causal relations between drugs and dangerous side effects [15], to link genetic variations and phenotype [16] (Figure 3 and Text Box 2) and to separate genetic regulatory relationships from the co-occurrence of molecular events [17].

Promising directions in computational biology for drug discovery

If we conceive the molecules inside a human cell – genes, proteins, RNAs and exogenous small molecules – as vertices of a graph and interactions between pairs of molecules as directed edges of this graph, we can represent the cell as a *molecular network*, as illustrated in the center scheme of Figure 3. Drugs are external substances that enter and interrupt this complex network of molecular processes. A robust, dynamic model of human cellular machinery could explain clinical outcomes for existing drugs and predict the effect of new leads. This is, of course, the Holy Grail of computational biology, a dream rather than reality. Nevertheless, recent advances in data integration have enabled the construction of complex ‘hairball’ networks based on text mining and high-throughput experiments [18] (Figure 3). Most of these networks are static, but as they become more precise and incorporate the temporality of biomolecular processes they are likely to enable identification and systematic screening of novel drug targets. Even current static networks suggest design principles and metrics that could accelerate the discovery of successful drugs. For example, population variability in the genes underlying molecular networks, especially in the region of the drug target, could pinpoint patient subgroups for which a drug might be most (and least) effective. The occurrence of a rare variation in patient subgroups might even warrant undertaking clinical trials for a personalized drug. On the other hand, molecules that are highly connected in the network of cellular processes or that are present in a number of human tissues are likely to be poor drug targets: their interruption will have broad and often unintended consequences. (See the fragile X case study in Box 2, for which the drug-responsive protein was distinct from the molecular cause of the condition). High-quality

targets will instead have the property of high 'betweenness centrality', i.e. they will ideally link normal and pathological pathways in such a way that interruption of their function will eliminate only disease-related processes, leaving healthy processes intact [19,20].

It is difficult, however, to isolate exclusive links between vicious and virtuous molecular processes. More often, multiple targets exist for which concerted interruption would exert a greater therapeutic effect than the disturbance of any one in isolation. This is the goal of *multi-drug cocktails*, a growing strategy that applies compounds jointly to interact with one or more molecular targets, as illustrated in Figure 3. One successful drug of this type is Advair, a cocktail of fluticasone and salmeterol used to manage asthma and chronic obstructive pulmonary disease. Most available cocktails have been discovered serendipitously, leaving much room for research into the action of multiple drugs and their systematic identification [21]. The potential of this direction is highlighted by the profound link between the genetic basis of disease and the biology of drug action. Many common pathological phenotypes, such as diabetes and neurodegenerative disorders, appear to have multifactorial genetic bases that are confounded by environmental risk factors. For example, patients with very similar disease phenotypes can have distinct genetic variants (genetic heterogeneity). Alternately, multiple genetic aberrations can act synergistically to contribute to a specific disease phenotype (genetic epistasis). For these disorders, multiple genes or proteins would need to be targeted concomitantly to achieve the desired outcome with minimal side effects. Current approaches to identify cocktails range from high-throughput combinatory screening for drug leads [22] to large-scale modeling of molecular systems [23,24]. Because the combinatorial search space for drug cocktails increases exponentially with the number of ingredients, experimental design methodologies will be critical to efficiently sample this space.

As for single drug leads, computational methods can be harnessed to rank combinations of drug leads. To begin, historical publications could be mined for associations between currently underinvestigated molecules and their impact on human biological processes. Subsequently, clinical databases can be screened to detect serendipitous connections between approved drugs and harmful or beneficial interactions when taken by the same patient [15]. Finally, predictive models can be designed to relate the molecular structure of cocktail components with the clinical phenotype of treated patients. These models would incorporate information about cellular networks and patient genetic variability.

Even when associations within a molecular network do not prove causal, the persistent association of molecules with pathological pathways could be exploited as *biomarkers* (Figure 3). Biomarkers [25,26] are molecules present in human tissues, such as proteins and lipids, which can change state in response to fluctuations in normal or pathogenic cellular processes or even therapeutic intervention. Biomarkers are extremely useful for disease diagnosis [18] and prognosis, calibration of appropriate drug dosage [27], evaluation of drug efficacy and toxicity [28] and patient stratification in clinical trials.

Biomarkers might also help us understand the genetic heterogeneity associated with pathological phenotypes [29]. This could suggest patient-specific treatment recommendations [30] so that different drugs are administered to patients with a distinct genetic variation underlying the same phenotype. Despite this promise, the area of biomarker development has recently been shrouded by controversy [31] because a number of published associations between putative biomarkers and target phenotypes that were based on microarray and other genomic data now appear irreproducible, possibly because of flaws in statistical analysis and interpretation.

We close this section with a discussion on *clinical data mining*, an area emerging with the growth of information technology in healthcare. There are many sources of clinical observation, from hospital billing records, patient discharge summaries and insurance archives to clinical trial statistics. All of these resources have been implemented in response to specific practical needs, usually related to billing for care. The richest of these resources, electronic health records, store information about tens of millions of patients and include demographic data, drug prescription history and side effects linked to the time series of diagnosed phenotypes and laboratory tests (Box 2). Despite the limitations described earlier, analysis of these records promises to be extremely useful. We anticipate using clinical databases for evaluating drug safety and the discovery of subtle adverse effects. Moreover, analogous to the rationale for using drug cocktails when multiple pathways lead to disease, knowledge of multiple diseases that result from the same pathway suggests opportunities to repurpose existing drugs, a considerably less expensive discovery strategy.

Modeling the social context of drug discovery

Drug discovery and development require expansive research teams that have increased in complexity over the past 25 years, spanning academy and industry and incorporating diverse expertise. Poorly orchestrated, disconnected or duplicative research programs can cause significant losses in research productivity and funds, leading to inflation of healthcare costs. The systematic study of innovation patterns in academia and industry has led to detailed models of the knowledge discovery process. We believe that improving these models will elucidate successful paths to drug discovery and highlight opportunities for institutional innovation that could increase the efficiency and public benefit of drug-related expenditure.

Based on analysis of research articles and patent bylines, social scientists have developed inventories of authors and inventors and of their evolving collaboration networks [32,33]. Social networks, like idealized molecular networks, provide a coarse view of complex social processes. They are nevertheless useful in capturing social patterns and trends. Recent findings point to the increasing importance of ‘big science’ as multi-author or team research has become more frequent and publications arising from such teams have the highest citation impact [34]. When team composition has been analyzed, those bridging different fields were more likely to achieve breakthroughs [35,36] and those mixing newcomers and incumbents resulted in the most referenced research [37]. By combining author and inventor information with text mining, it will become possible to build predictive models that guide scientists and organizations in assembling more successful teams for drug discovery.

Recent analyses of research articles and patents have revealed innovation patterns that vary considerably across the institutions in which biomedical research is carried out (Evans, unpublished). Analysis of regional biotechnology clusters suggests that as the number of institutional neighbors that fund research increases, the success of an institution’s own research also increases. These so-called spillovers have been measured in the form of successful patent applications [38,39], innovations described [40] and community recognition of innovations (e.g., article and patent citations) [41]. This suggests clear benefits of working in dense, well-funded research environments. The mobility of local labor is one factor that probably contributes to spillovers, as is the ease of informal data and idea sharing. Some ideas that spread locally through informal means never become widely disseminated because the review process for publications and patents systematically censors negative results, such as drug leads that do not inhibit certain targets. This highlights an opportunity for biomedical institutions to foster interpersonal and Internet venues where hypotheses, findings and negative results can spread.

Most notable in the field of drug discovery is the clear distinction between institutional sectors. Companies develop drugs for an anticipated market of patients, leaving fundamental investigation into treatments for small or impoverished populations underfunded. Governments and philanthropists aim to counteract these so-called *market failures* by selectively supporting such work in universities, research institutes and small companies. Fundamental research, however, has become more relevant for drug discovery since the detection of 'disease genes'. As a result, more scientists from academia and industry are converging to investigate the same diseases, leads and targets. Even though distinctions blur at the boundaries where researchers circulate between sectors [42], scientists in academia and industry have different incentives and resources that continue to inspire distinct activities and research cultures. In universities and research institutes, reputational credit is the reward scientists receive for priority in publishing [43] the most general scientific discoveries [44]. In companies, monetary rewards are obtained by controlling and appropriating benefits from the most powerful technological innovations [45]. As a result, university research is rapidly published to maximize dissemination, whereas company research is patented to maximize the ability to benefit from exclusive rights to the market [46]. Pharmaceutical companies surveyed in 1994 reported that secrecy was even more effective in protecting innovations than patenting, especially for techniques to manufacture drugs [46].

Other institutional differences exist. Because universities have a higher proportion of independent researchers per capital investment, at any given moment academic research will appear less focused than industry research, whereas companies channel money toward the exploitation of singular opportunities, such as an extensive molecular screen. Over time, however, industry shifts quickly between different research areas to identify and appropriate value from product-relevant discoveries, whereas academic researchers focus and swarm (Evans, unpublished). As a result, industry leaves behind many expensive draft genome sequences, partial screening experiments and incomplete collections of biomaterials that could be valuable for academic research.

These differences suggest unexploited complementarities between academic and industrial drug discovery efforts. Recent data [47] on the demography of illness, obtained from epidemiologists and researchers in the emerging field of population health, might enable us to improve models of existing and potential markets for drugs. This will, in turn, facilitate estimation of the potential market size for different treatments and answer important questions such as whether the afflicted are rich or poor, how many there are and how serious is the debilitating effect of the affliction. Models built with these data have recently been used to estimate the value that G8 states might be willing to prepay pharmaceutical and biotech companies for developing successful treatments afflicting impoverished populations unable to justify the commercial expense on their own [48].

Modeling potential drug markets, research institutions and investment alongside the biological systems described above will allow us to predict research areas for which cast-off information from industry could be most valuable. Some research suggests that cross-organization collaborations foster the most innovation in drug discovery [49]. Analyzing academia–industry complementarities explicitly will suggest where productive relationships could be formed and where governments could facilitate them by paying companies for: (i) leads or targets relevant for noncommercial diseases; (ii) rights to redirect drugs for noncommercial diseases that have a genetic overlap with commercially relevant diseases; and (iii) industrial 'draft' research made available for academic investigation and discovery.

Governments have already attempted to shape other aspects of the innovation landscape, which should now be reevaluated with available data. Because an increasing proportion of

biological research is medically relevant, the U.S. National Institutes of Health has begun to increase its sponsorship of 'translational' science to encourage the application of fundamental research (http://nihroadmap.nih.gov/clinicalresearch/overview-translational.asp). Governments have also expanded opportunities to obtain intellectual property (IP) for drug-relevant discoveries. Thirty years ago, patents were unavailable for many discoveries in basic science, but now it is possible to patent genes and altered organisms, as well as basic scientific research tools [50]. In 1976, after 9 years of trial and error, Eugene Goldwasser of the University of Chicago isolated erythropoietin [51], the human protein that stimulates red blood cell proliferation, and soon after identified the gene responsible, *rHuEpo*, publishing both discoveries [52]. These publications enabled Amgen and Johnson & Johnson to produce and to patent two synthetic erythropoietins, Epogen and Procrit, which are now blockbuster drugs prescribed for a variety of blood scarcity conditions. Today, this process would have been very different, probably involving academic patents with the help of a technological transfer office and pharmaceutical licensing or the formation of a spin-off company.

The choices that governments make to stimulate translational versus fundamental research might have major implications for the sustainability of pharmaceutical innovation. Translational research builds on existing biological knowledge and is most productive in short-term drug identification and optimization. Fundamental research into biological systems, however, is likely to be critical for the long-term success of drug discovery inasmuch as it proposes entirely new target families and therapeutic strategies. Researchers have begun to model this process and quantify the costs and benefits of research investment [45]. The importance of this approach is suggested by recent work that demonstrates 'the burden of knowledge' in modern biomedical science [53]. Much knowledge has already been accumulated and the diminishing success of established techniques suggests that 'low hanging fruit' on the tree of knowledge has been harvested.

Finally, many have observed the creation of patent 'fences' or 'thickets' surrounding fundamental biological techniques and resources [54]. Some have expressed concern that proliferating legal rights could result in an 'anti-commons' whereby fewer resources are available for researchers to build upon [55], slowing drug discovery and stifling life-saving innovations [56]. Computational methods can estimate the impact of IP rights on the accumulation of pharmaceutical knowledge and the speed of drug discovery. A recent analysis of paper–patent pairs covering identical biotechnologies suggested that once patents are granted, the dissemination of ideas slows slightly [57]. We believe that increasing resolution of data on published biomedical ideas and IP will enable realistic models of pharmaceutical innovation and legal protection that could become a powerful tool to advise economic and government policy.

Conclusions

We have discussed emerging computational approaches for modeling entire biological pathways relevant to health and disease. These hold great promise for drug discovery, especially building molecular network models of disease, mining the biomedical literature and patient records and harnessing computation to drive discovery of new targets, multi-drug cocktails and novel purposes for existing drugs. In parallel, electronic publication and patent data have facilitated the development of detailed models of the social process underlying biomedical innovation. We anticipate that advances in modeling the networks of scientists, inventors, institutions, resources and rights that drive drug discovery will reveal successful strategies and opportunities for institutional innovation to ensure the rate and social relevance of drug development.

Computation has traditionally been used *intensively* in biological and social science to solve difficult existing problems. Recent developments suggest that using computation *extensively* to link disparate data and support integrative models could broaden our vision of biological and social processes. We believe that this new view will fundamentally expand the possibilities of drug action and reorganize and reignite innovation in drug discovery.

Acknowledgments

We are grateful to Rajeev Aurora, Carolyn Cho, Murat Cokol, Preston Hensley, Ivan Iossifov, Aaron J. Mackey, and Nathan Siemers for insightful discussions and comments on earlier versions of the manuscript. This work was supported by the U.S. National Institutes of Health (GM61372 and U54 CA121852-01A1) and the National Science Foundation (0242971).

References

1. Oberholzer-Gee F, Inamdar SN. Merck's recall of rofecoxib – a strategic perspective. N. Engl. J. Med. 2004; 351:2147–2149. [PubMed: 15548771]
2. Public Citizen Report. Rx R&D myths: the case against the drug industry's R&D "scare card". Public Citizen Congress Watch. 2001
3. Tobert JA. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. Nat. Rev. Drug Discov. 2003; 2:517–526. [PubMed: 12815379]
4. Apic G, et al. Illuminating drug discovery with biological pathways. FEBS Lett. 2005; 579:1872–1877. [PubMed: 15763566]
5. Ananiadou S, et al. Text mining and its potential applications in systems biology. Trends Biotechnol. 2006; 24:571–579. [PubMed: 17045684]
6. Rzhetsky A, et al. Seeking a new biology through text mining. Cell. 2008; 134:9–13. [PubMed: 18614002]
7. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics. 2009; 10(Suppl. 2):S6. [PubMed: 19208194]
8. Druker BJ, Lydon NB. Lessons learned from the development of an abl tyrosine kinase inhibitor for chronic myelogenous leukemia. J. Clin. Invest. 2000; 105:3–7. [PubMed: 10619854]
9. Anastassiou D. Computational analysis of the synergy among multiple interacting genes. Mol. Syst. Biol. 2007; 3:83. [PubMed: 17299419]
10. Campillos M, et al. Drug target identification using side-effect similarity. Science. 2008; 321:263–266. [PubMed: 18621671]
11. Shannon, CE.; Weaver, W. The Mathematical Theory of Communication. University of Illinois Press; 1949.
12. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac. Symp. Biocomput. 2000:418–429. [PubMed: 10902190]
13. Basso K, et al. Reverse engineering of regulatory networks in human B cells. Nat. Genet. 2005; 37:382–390. [PubMed: 15778709]
14. Watkinson J, et al. Identification of gene interactions associated with disease from gene expression data using synergy networks. BMC Syst. Biol. 2008; 2:10. [PubMed: 18234101]
15. Almenoff JS, et al. Novel statistical tools for monitoring the safety of marketed drugs. Clin. Pharmacol. Ther. 2007; 82:157–166. [PubMed: 17538548]
16. Chen Y, et al. Variations in DNA elucidate molecular networks that cause disease. Nature. 2008; 452:429–435. [PubMed: 18344982]
17. Chen LS, et al. Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biol. 2007; 8:R219. [PubMed: 17931418]
18. Liu J, et al. Analysis of *Drosophila* segmentation network identifies a JNK pathway factor overexpressed in kidney cancer. Science. 2009; 323:1218–1222. [PubMed: 19164706]
19. Yildirim MA, et al. Drug-target network. Nat. Biotechnol. 2007; 25:1119–1126. [PubMed: 17921997]

20. Yao L, Rzhetsky A. Quantitative systems-level determinants of human genes targeted by successful drugs. Genome Res. 2008; 18:206–213. [PubMed: 18083776]
21. Zimmermann GR, et al. Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug Discov. Today. 2007; 12:34–42. [PubMed: 17198971]
22. Borisy AA, et al. Systematic discovery of multicomponent therapeutics. Proc. Natl. Acad. Sci. U. S. A. 2003; 100:7977–7982. [PubMed: 12799470]
23. Qutub AA, et al. Integration of angiogenesis modules at multiple scales: from molecular to tissue. Pac. Symp. Biocomput. 2009:316–327. [PubMed: 19209711]
24. Klein ML, Shinoda W. Large-scale molecular dynamics simulations of self-assembling systems. Science. 2008; 321:798–800. [PubMed: 18687954]
25. Marrer E, Dieterle F. Promises of biomarkers in drug development – a reality check. Chem. Biol. Drug Des. 2007; 69:381–394. [PubMed: 17581232]
26. Marrer E, Dieterle F. Biomarkers in oncology drug development: rescuers or troublemakers? Expert Opin. Drug Metab. Toxicol. 2008; 4:1391–1402. [PubMed: 18950281]
27. Owen RP, et al. PharmGKB and the International Warfarin Pharmacogenetics Consortium: the changing role for pharmacogenomic databases and single-drug pharmacogenetics. Hum. Mutat. 2008; 29:456–460. [PubMed: 18330919]
28. Holly MK, et al. Biomarker and drug-target discovery using proteomics in a new rat model of sepsis-induced acute renal failure. Kidney Int. 2006; 70:496–506. [PubMed: 16760904]
29. Dudley JT, Butte AJ. Identification of discriminating biomarkers for human disease using integrative network biology. Pac. Symp. Biocomput. 2009:27–38. [PubMed: 19209693]
30. Chiang AP, Butte AJ. Data-driven methods to discover molecular determinants of serious adverse drug events. Clin. Pharmacol. Ther. 2009; 85:259–268. [PubMed: 19177064]
31. Nelson AE, et al. Failure of serum transforming growth factor-beta (TGF-β1) as a biomarker of radiographic osteoarthritis at the knee and hip: a cross-sectional analysis in the Johnston County Osteoarthritis Project. Osteoarthritis Cartilage. 2009; 17:772–776. [PubMed: 19091605]
32. Newman ME. Coauthorship networks and patterns of scientific collaboration. Proc. Natl. Acad. Sci. U. S. A. 2004; 101(Suppl 1):5200–5205. [PubMed: 14745042]
33. Fleming L, Frenken K. The evolution of inventor networks in the silicon valley and Boston regions. Adv. Complex Syst. 2007; 10:53–71.
34. Wuchty S, et al. The increasing dominance of teams in production of knowledge. Science. 2007; 316:1036–1039. [PubMed: 17431139]
35. Burt RS. Structural holes and good ideas. Am. J. Sociol. 2004; 110:349–399.
36. Fleming L, Sorenson O. Technology as a complex adaptive system: evidence from patent data. Res. Pol. 2001; 30:1019–1039.
37. Guimera R, et al. Team assembly mechanisms determine collaboration network structure and team performance. Science. 2005; 308:697–702. [PubMed: 15860629]
38. Owen-Smith J, Powell WW. Knowledge networks as channels and conduits: the effects of spillovers in the Boston biotechnology community. Organ Sci. 2004; 15:5–21.
39. Fritsch M, Franke G. Innovation, regional knowledge spillovers and R&D cooperation. Res. Pol. 2004; 33:245–255.
40. Acs ZJ, et al. Patents and innovation counts as measures of regional production of new knowledge. Res. Pol. 2002; 31:1069–1085.
41. Jaffe AB, et al. Geographic localization of knowledge spillovers as evidenced by patent citations. Q. J. Econ. 1993; 108:577–598.
42. National Research Council. Intellectual Property Rights and Research Tools in Molecular Biology. National Academy Press; 1997.
43. Merton RK. Priorities in scientific discovery: a chapter in the sociology of science. Am. Sociol. Rev. 1957; 22:635–659.
44. Latour, B. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press; 1987.
45. Partha D, David PA. Toward a new economics of science. Res. Pol. 1994; 23:487–521.

46. Cohen WM, et al. Protecting their intellectual assets: appropriability conditions and why U.S. manufacturing firms patent (or not). NBER Working Paper 7552. 2000
47. Anonymous. Disease and demography. Health Aff. (Millwood). 2008; 27:1051. [PubMed: 18607040]
48. Berndt ER, et al. Advance market commitments for vaccines against neglected diseases: estimating costs and effectiveness. Health Econ. 2006; 16:491–511. [PubMed: 17013993]
49. Powell WW. Interorganizational collaboration in the biotechnology industry. J. Inst. Theor. Econ. 1996; 120:197–215.
50. Heller MA, Eisenberg RS. Can patents deter innovation? The anticommons in biomedical research. Science. 1998; 280:698–701. [PubMed: 9563938]
51. Goldwasser E. Erythropoietin: a somewhat personal history. Perspect. Biol. Med. 1996; 40:18–32. [PubMed: 8946758]
52. Jelkmann W. Erythropoietin after a century of research: younger than ever. Eur. J. Haematol. 2007; 78:183–205. [PubMed: 17253966]
53. Jones BF. The burden of knowledge and the “death of the Renaissance man”: is innovation getting harder? Rev. Econ. Stud. 2009; 76:283–317.
54. Jensen K, Murray F. Intellectual property. Enhanced: intellectual property landscape of the human genome. Science. 2005; 310:239–240. [PubMed: 16224006]
55. Heller MA, Eisenberg RS. The tragedy of the anticommons: property in the transition from Marx to markets. Harv. Law Rev. 1998; 111:621–688.
56. Turner JS. The nonmanufacturing patent owner: toward a theory of efficient infringement. Calif. Law Rev. 1998; 86:179–210.
57. Murray F, Stern S. Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis. J. Econ. Behav. Organiz. 2007; 63:648–687.
58. Lipkus AH, et al. Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. J. Org. Chem. 2008; 73:4443–4451. [PubMed: 18505297]
59. Klein TE, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. N. Engl. J. Med. 2009; 360:753–764. [PubMed: 19228618]
60. Eichelbaum M, et al. New feature: pathways and important genes from PharmGKB. Pharmacogenet. Genomics. 2009; 19:403. [PubMed: 20161212]
61. Garber KB, et al. Fragile X syndrome. Eur. J. Hum. Genet. 2008; 16:666–672. [PubMed: 18398441]
62. Mulligan LM, et al. Genetic mapping of DNA segments relative to the locus for the fragile-X syndrome at Xq27.3. Am. J. Hum. Genet. 1985; 37:463–472. [PubMed: 2988332]
63. Gedeon AK, et al. Fragile X syndrome without CCG amplification has an FMR1 deletion. Nat. Genet. 1992; 1:341–344. [PubMed: 1302032]
64. Bakker CE, et al. *Fmr1* knockout mice: a model to study fragile X mental retardation. The Dutch–Belgian Fragile X Consortium. Cell. 1994; 78:23–33. [PubMed: 8033209]
65. de Vrij FM, et al. Rescue of behavioral phenotype and neuronal protrusion morphology in *Fmr1* KO mice. Neurobiol. Dis. 2008; 31:127–132. [PubMed: 18571098]
66. Pecknold JC, et al. Treatment of anxiety using fenobam (a nonbenzodiazepine) in a double-blind standard (diazepam) placebo-controlled study. J. Clin. Psychopharmacol. 1982; 2:129–133. [PubMed: 7042771]
67. Jacob W, et al. The anxiolytic and analgesic properties of fenobam, a potent mGlu5 receptor antagonist, in relation to the impairment of learning. Neuropharmacology. 2009; 57:97–108. [PubMed: 19426746]
68. Berry-Kravis E, et al. A pilot open label, single dose trial of fenobam in adults with fragile X syndrome. J. Med. Genet. 2009; 46:266–271. [PubMed: 19126569]

Box 1. Drug discovery process

The traditional drug discovery workflow is shown in Figure 1 in red. It typically begins with target identification. The target is a human molecule that a drug recognizes and modifies to achieve an intended therapeutic effect. Alternatively, the target can be part of the cellular machinery of a pathogen; the role of the drug in this case is to kill the pathogen by interrupting the drug target. Most drug targets are proteins, historically drawn from a few families, such as enzymes, receptors and ion channels. Target identification is heavily dependent on: (1) analysis of disease mechanisms to locate the molecular system most likely to incorporate a promising target; (2) genomics to rank genes with respect to physiological function; and (3) experimental proteomics to identify candidate proteins and protein interactions that can be inhibited or enhanced by a drug.

The next stage is target validation. At this stage researchers use a battery of experimental techniques (genetic engineering, transgenic animal models, antisense DNA/RNA perturbation of pathways and structural biology) to better understand the molecular role of the prospective drug target and to determine whether an agonist or antagonist drug should be designed. It is not uncommon to discover that the initial target is inappropriate for a variety of reasons, in which case target identification must be repeated.

Following target validation, if the targeted molecule appears promising, it is time to identify and optimize a lead or prospective drug. Most frequently, the lead is a small molecule, but it can also be a peptide, antibody or other large substance. To appreciate the difficulty of this stage, consider the number of possible molecules. Although finite, because molecule size is naturally bounded, the number of potentially relevant compounds is greater than 10^{24} [58] even if we limit ourselves to small molecules. ‘Brute force’ search approaches, in which all leads are tested exhaustively, are clearly not feasible; intuition and serendipity are highly valued. There are numerous high-throughput techniques, such as synthetic and combinatorial chemistry, compound screening and design of compound libraries, and a variety of experimental design and sampling approaches (Figure 1) for relatively efficient searching of the space of lead compounds.

Lead optimization is the process of improving initial hits from primary screening. Medicinal chemists draw on principles of organic chemistry to tweak hit molecules and carry out batteries of tests that ensure the modified molecule is more specific to its target, more effective, less toxic, and has a long enough half-life in human tissue to achieve therapeutic effect. In the framework of lead optimization, pharmacodynamics measures the strength of drug effect as a function of drug dosage; pharmacokinetics studies the time course of drug absorption, distribution, metabolism, and excretion; and toxicology deals with any poisonous impact of the new lead.

Clinical trials are the pinnacle of drug development: The general public and patent agencies typically do not learn about the existence of leads until they reach trials. Clinical trials are divided into several stages. Phase I trials typically investigate the safety of a drug when administered to healthy humans. Given encouraging phase I results, phase II explores efficiency and safety of the drug in the class of patients expected to benefit from the proposed therapeutic effect. Phase III trials are similar to phase II, but the subject pool is larger to detect subtle influences of the drug and to better evaluate its efficiency. There are also optional phase IV trials that can provide improved estimates of drug efficiency and adverse effects. By the time the drug reaches phase IV trials, it has usually reached the market. Clinical trials can be followed by post-market research to discover and validate new applications for existing commercial drugs.

Box 2. Successful application of computational biology in drug discovery: two case studies

Case study 1. Determining patient-specific warfarin dosage based on clinical and genetic data

Warfarin is the most widely prescribed anticoagulant in North America. Delivering the right amount is critical. If the dose is too high, patients could bleed profusely; if it is too low, they could develop life-threatening clots. This task is difficult for doctors because ideal dosage varies from patient to patient. In a recent study [59], a consortium of researchers pooled data on 5052 patients who underwent warfarin treatment across several organizations. Information for 4043 patients was used to develop and fit a statistical model and data for the remaining 1,009 patients were used to validate the model. Parameters included traditional clinical variables such as demographic characteristics, primary justification for warfarin treatment, concomitant medications and genetic variables, primarily single nucleotide polymorphisms in the genes encoding VKORC1 (the therapeutic target of warfarin) and CYP2C9 (related to warfarin metabolism) for each patient.

The problem was then framed as a statistical regression: the outcome variable was the stable therapeutic dose of warfarin, and input variables included all clinical and genetic variables mentioned above. The researchers used a battery of regression methods: support vector regression, regression trees, model trees, multivariate adaptive regression splines, least-angle regression, Lasso and ordinary linear regression. The results show that for the 46% of warfarin recipients who are difficult to dose because they require abnormal quantities of warfarin, the ordinary linear regression model with two biomarkers performed best and was better than models with clinical data alone. For these patients, model predictions could be life saving.

The work drew on PharmGKB (Pharmacogenomics Knowledge Base), which required knowledge engineering to structure its database schema and content [60]. Patient-specific amounts of known biomarkers were used as input variables for the analysis.

Extensions to this work could include text and data mining efforts [10,19] to identify additional markers for analysis (e.g., genetic loci correlating with phenotype) and interactions between warfarin and food additives or other drugs. In turn, this would require the redesign of regression models to include additional variables and likely second-order interactions between them. Analysis would then require a larger sample of patient records. To obtain such a sample, still larger multi-institutional collaborations would be required, the design of which would, in turn, benefit from dynamic models of research collaboration.

Case study 2. Fragile X syndrome, the *Fmr1* gene, and fenobam

Fragile X syndrome [61] is an X-linked dominant genetic disorder [62] most frequently caused by transcriptional inactivation of the *Fmr1* gene, either due to expansion of a CGG repeat found in the 5′-untranslated region or deletion of the gene [63]. Fragile X syndrome is associated with intellectual and emotional disabilities, ranging from mild cognition impairment to mental retardation and autism.

Experiments with transgenic *Fmr1* knockout mice [64] demonstrated that, in the absence of functional FMR1 protein, neuronal synapses are altered and dendritic protrusions are structurally malformed [65]. It was shown that two different antagonists of the mGluR5 receptor can rescue fragile X-related neuronal protrusion morphology [65].

An unexpected discovery was that researchers could compensate for the absence of FMR1 protein by tuning the state of molecules in its interacting neighborhood by inhibiting the mGluR5 receptor. Fenobam is a relatively old anti-anxiety drug [66] and a potent antagonist of mGluR5 receptor [67]. It was therefore natural to suggest that fenobam might rescue Fragile X-related abnormalities. Preliminary results have been encouraging and a clinical trial of an oral drug therapy option for adult patients with fragile X syndrome is currently under way [68].

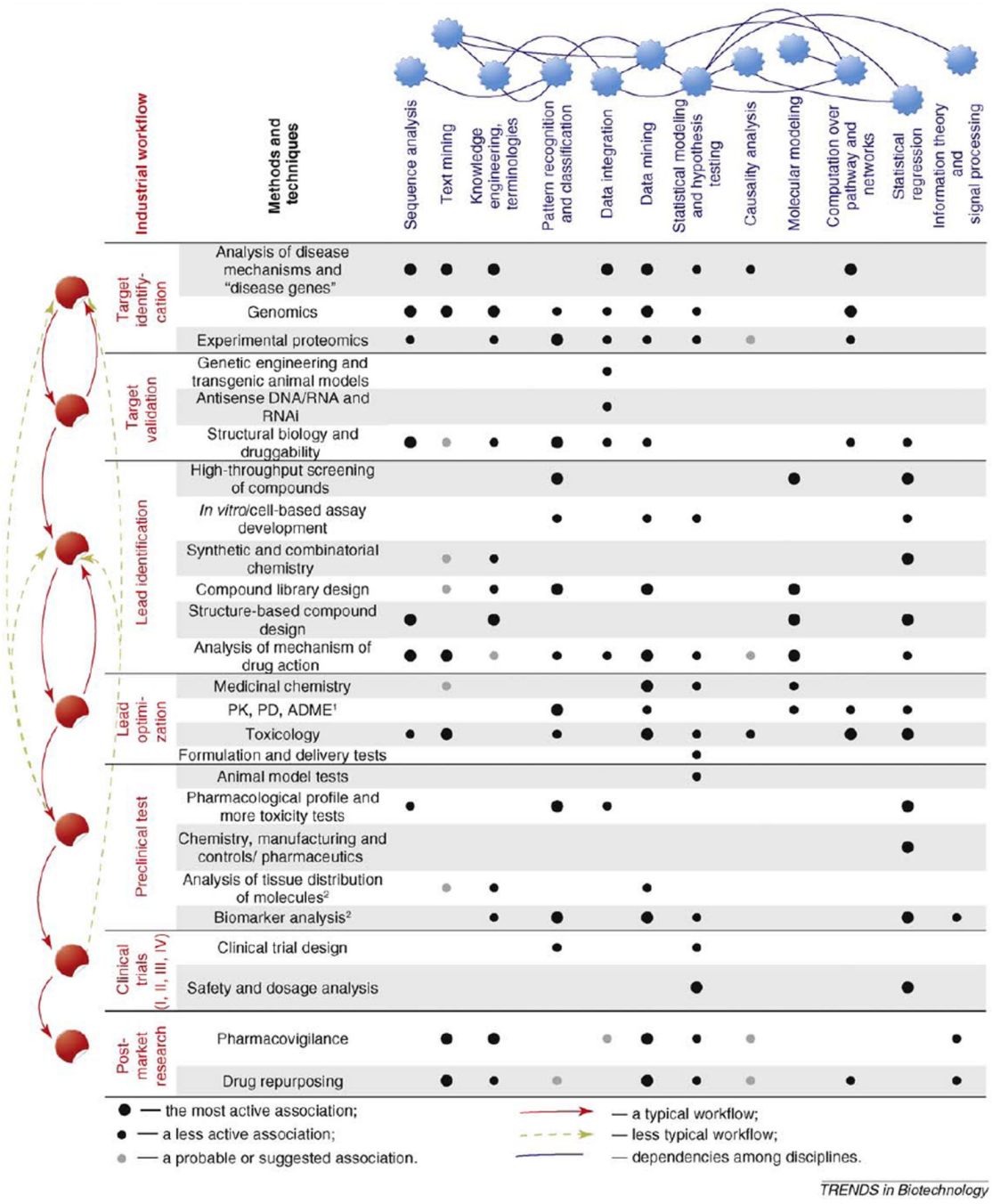

Figure 1.

Role of computational technologies in the drug discovery process. This figure summarizes how computational biology can impact drug discovery. The various stages of the drug discovery process (See Box 1 for detailed background on each step) are listed in the left column. We note that the traditional linear process is shifting to become more parallel, simultaneous and cyclical. Red arrows indicate the traditional process and yellow dashed arrows suggest novel workflows that are increasingly adopted by pharmaceutical and biotechnological companies to increase productivity. Biomarkers and analysis of the tissue distribution of target molecules are the most recently introduced checkpoints and are not required by the FDA.

Computational biology methods discussed in the main text* are listed along the top row. Blue lines illustrate how each method is related to others. For example, sequence analysis relies on pattern recognition and classification; text mining, terminologies and knowledge engineering are entwined, as are pattern recognition and classification. The impact of each computational technique on each stage of drug discovery is classified into three categories: actively or heavily used (large black dot), less actively used (small black dot) and our suggestion (small gray dot).

*We do not emphasize chemical informatics in the main text because it relates to issues from chemistry and not biology. Chemical informatics comprises a wide range of approaches from computational and combinatorial chemistry that model lead properties and their interaction with targets. These include chemical structure and property prediction; structure–activity relationships; molecular similarity and diversity analysis; compound classification and selection; chemical data collection, analysis and management; virtual drug screening; and prediction of *in vivo* compound characteristics.

PK, pharmacokinetics; PD, pharmcodynamics; ADME, absorption, distribution, metabolism and excretion.

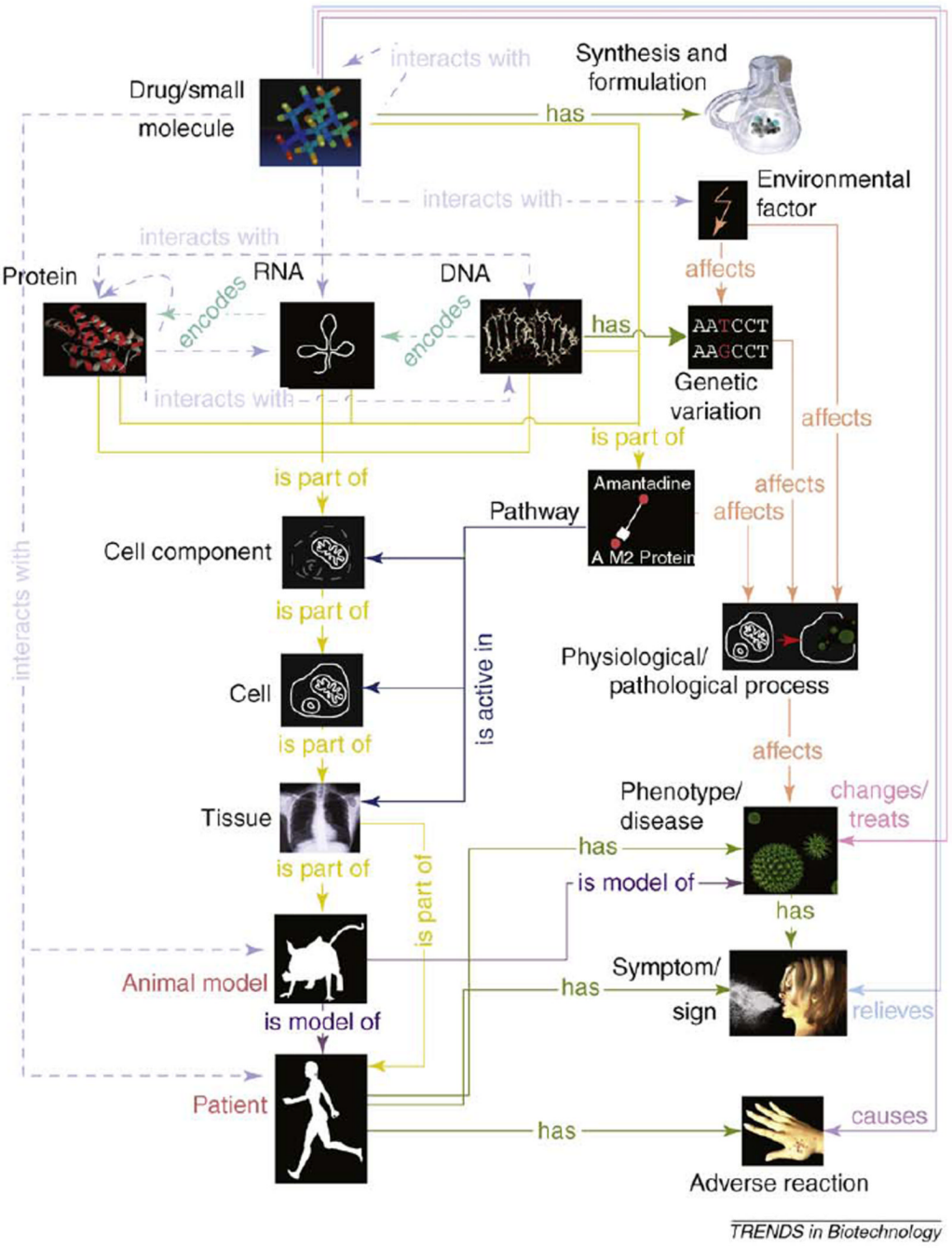

Figure 2.

Provisional ontology for drug discovery. The ontology features a concept for drugs or small chemical molecules and another for drug synthesis and formulation. Furthermore, a drug is known to target a specific molecule: protein, RNA, or DNA. A drug target is biologically active in the context of a pathway, cellular component, cell, and tissue. The drug is typically tested using an animal model for a specific human phenotype, usually a disease. A patient or human with the disease phenotype has a set of symptoms/signs, some of which can be unrelated to the disease. Drugs treat patients, but can cause adverse reactions. Finally, we link the patient's drug response to genetic variation in her genome and environmental factors encountered. These concepts are linked to directed or undirected relations, such as encodes

(DNA encodes RNA), interacts with (drug–drug interactions), is part of (a cell is part of a tissue), affects (environmental factor affects genetic variation) and several others. We illustrate our ontology with the drug amantadine, which acts as a prophylactic agent against several RNA virus-induced influenzas that afflict the respiratory system and result in coughing and sneezing. Amantadine can cause a skin rash as a side effect.

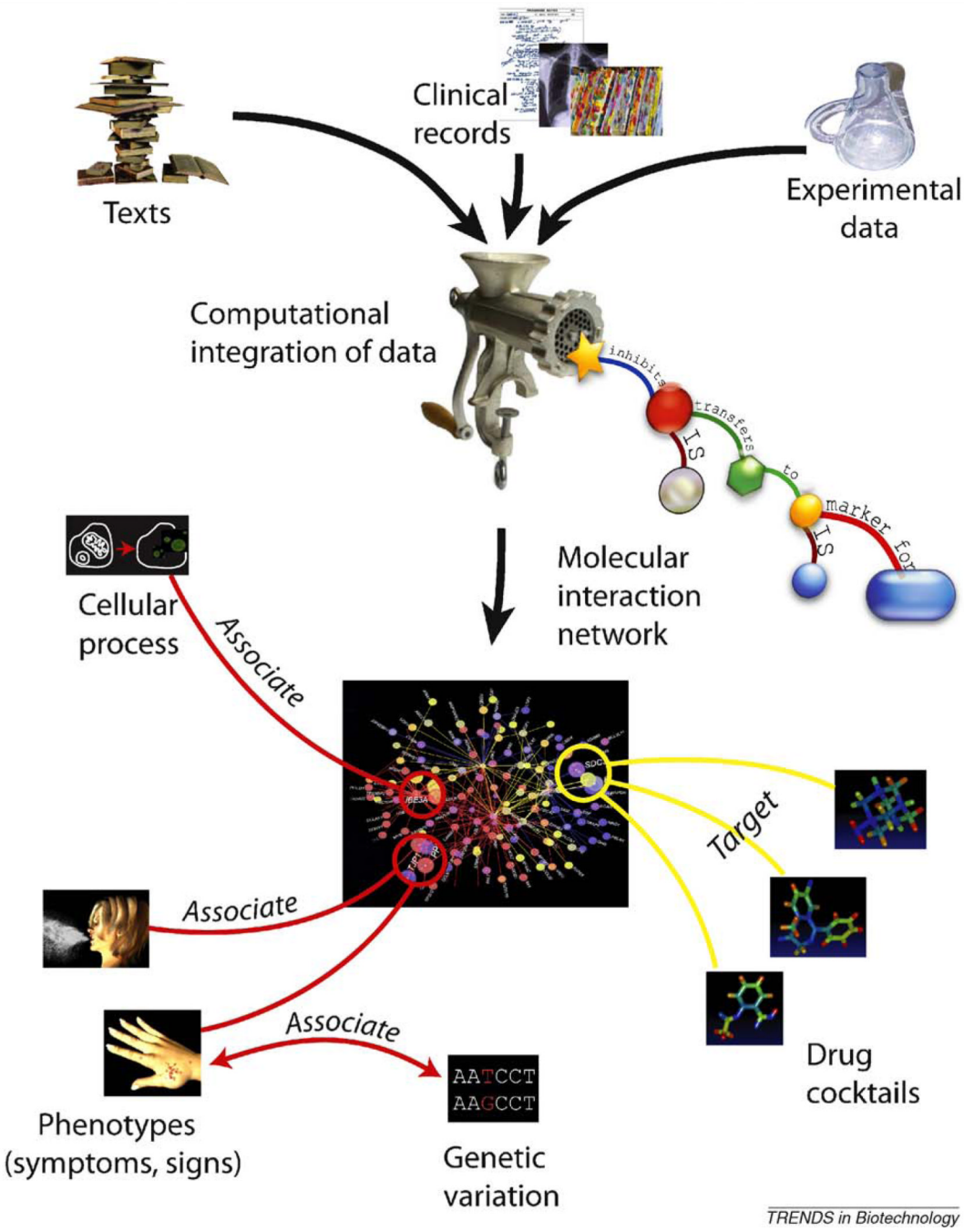

Figure 3.

Overview of promising computational opportunities in drug discovery. Text mining is used to extract information from publications and clinical records. Mathematical modeling helps to assess experimental data in the context of previously collected facts, whereas computational data integration distills multiple types of raw data into a collection of computable biological statements. The resulting network of semantic relations can serve as a scaffold for modeling biological processes, for design and optimization of therapeutic drug cocktails and for linking complex phenotypes to genotypes. The figure incorporates ontological concepts outlined in Figure 2: cellular process (such as tissue necrosis), symptoms (in this case, sneezing and allergic rash), genetic variation (depicted as a single

nucleotide polymorphism) and drugs (amantadine, valium and aspirin) listed here from top to bottom.