Biomedical informatics: the future for drug development

Michael N. Liebman

The problems that exist in drug development are well documented: the limited number of new chemical entities, increased cost of drug development, problems in clinical trials (Phase III), product launches that result in withdrawal, and pressure to reduce the cost of pharmaceuticals from the government. It appears that the promise of genomics has not yet reached its full potential to impact the process. This review identifies the need to develop and implement the area of biomedical informatics for increased success in drug development and healthcare in general.

Michael N. Liebman
Computational Biology
Abramson Cancer Center of the University of Pennsylvania
Philadelphia
PA 19104 USA
tel: +1 215 746 5570
fax: +1 215 573 2486
e-mail: liebmanm@mail.med.upenn.edu

▼ Although the use of genomics in pharmaceutical research and healthcare continues to expand, it is clear that more than just sequence information is needed to successfully achieve its full potential. The ongoing evolution of bioinformatics attempts to translate genome-based information into knowledge, either for immediate diagnostic purposes or in the identification of new targets for therapeutic development. The resulting high-throughput access to potential pharmaceutical targets has proven more significant in highlighting our limited understanding of normal physiological processes and disease. Few 'one gene-one disease' examples exist that can lead to major pharmaceutical products, other than for diagnostic tests (e.g. Her2/neu [1]) or prognostic tests (e.g. Huntington's disease [2]). Thus, it appears that truly monogenetic events are rare and are most pertinent to orphan drug development – an area that is not a primary business opportunity within the major pharmaceutical companies. Major pharmaceutical products (as determined by revenue production and research commitment [3]) as well as consumer healthcare (as determined by consumer costs, provider costs and major causes of death [4]) are focused on what is termed complex diseases (e.g. depression and diabetes). Even apparent single-gene disorders, such as sickle cell trait, which impacts approximately 10% of black Americans of western African origins, are likely to be observed as complex disorders, involving the perturbation of additional genes and systems that produce the difference in symptoms observed in individual patients [5]. Thus, there exists a gap between sequence-based information and the need to understand the basis of complex disease. The term 'biomedical informatics' is used to describe a perspective that complements bioinformatics and that incorporates data and observations from the clinic (medical informatics).

Research progress in the human genome project

Research progress in the Human Genome Project (HGP) has led to the development of multiple '-omics', (e.g. proteomics, genomics and metabolomics) and has subsequently impacted both technology development and financial investment [6]. Weinstein noted that the use of this derivative is correct, as each -omic represents a collection of information about its root area of study [7]. The integrated study of these -omics presents significant computational challenges to enable the development of new knowledge [8]. This has resulted from the knowledge base being built 'bottom-up', with each -omic focusing only on a component of the bigger overarching process – human physiology. The strategic goal of the HGP depends on the tactical collection of crucial biological, biochemical and genomic information, but also requires its integration and conversion into knowledge within the perspective of genetic risk and disease. The goal of biomedical informatics is to construct and analyze synergistic interactions at the levels of molecules, pathways, cells, tissues, organs and the physiological system that produces the clinical observations of behavior and response of the organism. This is highly

1359-6446/02/$ – see front matter PII: S1359-6446(02)02469-8

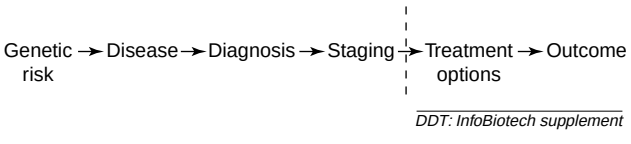

Figure 1. A generalized model for disease progression, from genetic risk to active disease and treatment. The dotted line separates the difference between clinical data (to the left of the line) and clinical trial data (to the right of the line) where the first is key to understanding the onset, progression and stratification of the disease and individual response, and the second focuses on treatment and response, without necessarily incorporating a detailed mechanism of disease progression.

compatible with systems biology [9] but focuses on refining our understanding of observations made when studying the behavior of the system, rather than attempting to predict behavior from the component parts of the system.

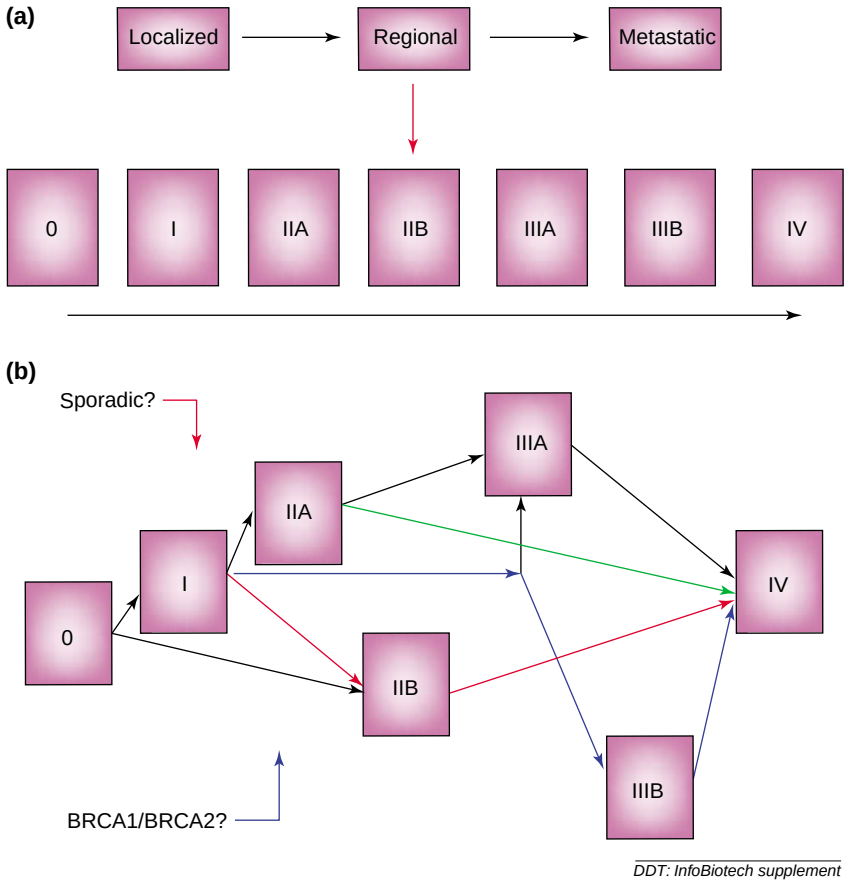

Figure 2. (a) Linear progression of tumor invasiveness and tumor staging. The assumed progression of a tumor is represented in terms of its invasiveness and the analogous linear progression of its staging based on its pathological classification. (b) Possible non-linearity of tumor progression. Because patients are treated upon diagnosis of breast cancer, the actual natural disease progression is not followed in an untreated individual. This might obscure differences in natural progression that could occur differently in BRCA1/BRCA2 breast cancer versus sporadic breast cancer.

The current standard of care for treatment of disease will probably undergo a paradigm shift as genomics becomes a component of diagnosis and treatment. This will require genome data to be refined to a higher quality, and bioinformatics to continue integrating computational solutions from other disciplines. This new paradigm (Fig. 1) will significantly impact the practice of medicine and the associated development of diagnostic and therapeutic intervention because it focuses on a more rigorous stratification of the disease process and recognition of environmental interaction with the patient's genotype.

Fundamental to this change will be the recognition that disease is a process, extending from risk (genotype) through to the stages that reflect not only the individual's cumulative exposure over time but also interaction with environment and lifestyle. Disease will be viewed as a process over time, rather than a state at a point in time, and almost all disease classifications that are currently used will be further stratified using molecular pathology to reflect the complex interaction between genotype and environment. The need to expand the definition of disease can be readily seen in the current diagnosis of breast cancer (Fig. 2a). It is assumed that breast cancer might progress from a local to a regional to metastatic disease. This progression aligns with the more highly defined tumor staging shown in Fig. 2b. Ethics and common sense in medical practice do not permit a patient diagnosed with breast cancer to be denied treatment. Thus, the actual disease progression cannot be followed, which raises the issue noted in Fig. 2b, namely, is there only one unique path through tumor stages?

The significance of the potential difference in disease pathways that pass through similar stages relates directly to the accuracy of the diagnosis and designation of treatment for the underlying disease. Thus, a patient exhibiting Stage II breast cancer that arises through association with genetic mutations in BRCA1 might not be on the same disease path as a patient with sporadic breast cancer. The correct treatment and patient outcome is dependent on identifying and evaluating the differences in these pathways. This stratification will provide a more realistic level of specificity for diagnosis and treatment than the hypothetical 'personalized medicine'. As a result,

medical practice will probably evolve to focus more on the determination of risk and intervention in the pre-disease process rather than amelioration of symptoms (Fig. 1) as the case at present. This requires a fundamental understanding of disease etiology, and will probably achieve success in a disease-specific manner before the complex interactions produced by the underlying pathophysiology can be fully recognized. Biomedical informatics will focus on developing the tools necessary to extend the present boundaries of medical informatics and bioinformatics to accomplish these goals.

Medical informatics

Although the practice of medicine, in some fashion, probably dates back to pre-historic times, the development of medical informatics, or the process of information collection and exchange in medicine, only began in the late 1940s, with its real growth occurring in France during the 1960s [10]. This obviously refers to the organized collection and sharing of data concerning clinical practice, and so on. The recording of medical data, in general, began in places such as the Mayo Clinic at its inception in the early 1900s (http://www.mayo.edu). The significant gap between the recording of medical data and the development of processes to enable collection and sharing still exists, as most hospitals, including Mayo, struggle to develop a comprehensive electronic medical record. The primary drivers for an electronic medical record have been the need to support patient billing, scheduling of appointments, and filing of insurance claims. The development of standards for data incorporation from laboratory systems (HL-7; http://www.hl7.org/) has supported general data collection, but no standards exist as to what data (fields) need to be collected. In addition, because much of the reported diagnostic coding is used for insurance reimbursement, claims-based medical records within the US frequently do not contain diagnoses that are sufficiently accurate to support clinical research.

In general, medicine is practiced in a tactical manner: treating the symptoms of immediate concern without fully exploring the aspects of cause and effect (as depicted on the right of Fig. 1). Diagnosis requires disease classification and staging within standard definitions (ICD-10; http://www.who.int/whosis/icd10/) that correspond with diagnostics and treatment options (i.e. standard of care). These crucial tasks are performed using laboratory- and physical-observations within a relatively small timeframe on disease processes that could have a much longer time course and be further complicated by other underlying diseases or physiological differences that are not readily observable. The standard nomenclature used in the literature and for general classification, UMLS (http://www.nlm.nih.gov/research/umls/umlsmain.html) and SNOMED (http://www.snomed.com/), reflects the use of classifications made at the specific time of diagnosis, focusing on the state model of disease rather than the evolving model of disease.

Biomedical informatics should build on the extensive base of existing knowledge but change the perspective from which it is collected and analyzed. This data collection will require a concerted effort in the development of a longitudinal electronic medical record that the patient owns, and in establishing uniform data specifications for potential data integration and use in research. The data collected in clinical trials necessarily focus on the patient with a diagnosed illness and their response to treatment for support of FDA submissions. Standard minimal datasets should be established for use in all clinical trials so that cross-trial comparisons can also be carried out more effectively for the benefit of the industry, as well as for the benefit of the patient. Standard structures for family history should also be established to further enable the observation of crucial patterns of disease inheritance and to further integration with genomic studies. These activities will not be without technological challenges; perhaps even greater will be the cultural and ethical challenges that society is not as well prepared to confront. Issues concerning patient privacy need to be maintained, and balanced with the need to improve diagnosis and staging of disease, and treatment decisions associated with reducing the cost of healthcare overall. Pharmaceutical research will also have to evolve to accommodate the use of more stratified disease classification in the drug development process. Rather than facing the extreme case of truly personalized medicine, this middle ground can focus drug development on more selective targets with higher efficacy and better enrollment criteria for clinical trials, thus shortening the overall development process, and improving return on investment by extending the time a drug has under patent protection.

Bioinformatics

Bioinformatics began in the mid- to late-1980s, with many early bioinformatisists starting out as X-ray crystallographers studying the relationship between structure and function in proteins [11]. Early research support came from the Department of Energy through studies focused on the potential biological effects of radiation exposure (http://www.ornl.gov/hgmis/project/about.html). The enabling technology that fueled this area was the development of molecular biology and nucleic acid sequencing technology. The sequencing of the human genome evolved from an 'intellectual challenge' in 1990 to a fully-fledged competitive research community within a few short years. Although technological advances have provided increasingly more rapid means to help refine its accuracy (by determining the sequence multiple times), much of the genome remains unsequenced because of its 'lack of genes'. An interesting challenge that has developed is determining how many genes are actually encoded in the human genome,

with initial estimates being in the 100,000–120,000 range and more recent estimates in the 30,000–40,000 range [12]. The number of gene products or proteins that exist appears much higher, so the mechanism by which proteins are formed from the genetic sequence, through mixing and matching different segments, reveals our limited understanding of the complex relationship of structure and function complexity in the genome. While the actual number of genes remains an academic question, the underlying mechanisms for gene to protein production are probably of more value to drug development because their potential for use as drug targets. Mutations in the amino acid sequences of the population have been observed for some time; characterization of hemoglobin mutations and their observed biochemical and/or physiological functional perturbations have been well documented [13]. More recently, observations of single nucleotide polymorphisms (SNPs) in the sequence are being studied that might establish the lack of homogeneity among individuals in the overall population, with some having structural or functional impact. The observation of these SNPs is an active area of research with an emphasis on their potential medical relevance.

Significant efforts to define and organize the increasingly complex and detailed levels of information resulting from genomics research have been focused in developing a 'gene ontology' (http://www.geneontology.org/) [14] that serves to categorize the terminology and establish the inter-relationships and context of component organization.

Development of additional high-throughput methodologies has focused on determining the levels of expression of mRNA and proteins, using a variety of array technologies (http://www.ncbi.nih.gov/geo/). Most experimental studies using these methods have been more correlative than mechanistic; the measurements can be viewed as somewhat equivalent to the biomarkers that are used in clinical diagnosis. As with all science, the most accurate observations are determined by evaluating the difference between two or more states, and this is also true for array technologies. The limitation in using such approaches is that it is a surrogate for the desired measurement of the difference in actual biological activity, and the observed differences in expression reflect only one step in the multi-step process that produces the biological agent and its action from either the mRNA or protein expression level (Eqn 1). Expression of a particular gene, as measured by transcript profiling and microarrays, produces, at most, a level of information about the gene that requires additional biological and biochemical processing before its biological activity can be established. Therefore, many additional observations are needed to completely determine the change in biological activity beyond just evaluation of the change in expression, as noted in Eqn 1.

$$[\Delta \text{ Expression}] \times [\text{Biological activity}] = [\Delta \text{ Expression}] \times [\text{Impact of transcription, translation, post-translation modification, transport and compartmentalization}] \quad \text{[EQN 1]}$$

Parallels between bioinformatics and medical informatics

Time and process

It is interesting to note the parallels between the current application of bioinformatics and medical informatics (Fig. 1 and Eqn 1, respectively). The important concept in both equations involves that of time and process. In Fig. 1, the process is the development of disease as a continuum, from genetic risk through active disease and into treatment and response; in Eqn 1, the process involves the conversion of mRNA expression into the gene product, and then further processing to generate its intrinsic biological activity. Biological and chemical processes occur over time and it is essential to incorporate this into the interpretation of experimental observations at both clinical and genomic levels.

Use of biomarkers

Biomarkers are used to indicate the presence of or level of activity of a given process, ideally, through the measurement of a reactant that is consumed or produced by that process. In this way, the biomarker is directly related to the mechanism-of-action. Unfortunately, most biomarkers are not mechanistically derived but rather arise from an observed correlation with the existence of the process. An example of a clinical biomarker is the finding that prostate specific antigen (PSA) is elevated in prostate disease, although recent studies have shown that diagnosis based on this parameter might be misleading [15]. Others examples include mutations in BRCA1 and BRCA2 [16] and levels of Her2/neu [17], which are associated with breast cancer risk and severity of disease, respectively. At the genomic level, changes in the level of mRNA expression or protein expression, as noted in Eqn 1, are markers that are used to signify changes in biological activity, although they only comprise a single component of the overall process involved in determining biological activity.

Annotation

Annotation is used in genomics to assign a descriptor to one object because of its similarity to another, or to assign a new descriptor to a unique object for use in later comparisons. In this way, annotation of genomic sequences is by comparison of sequence analogy or homology, which ranges from <25% to >90%, and assignment of terms to indicate a close relationship between sequences (e.g. insulin-like growth factors). In medicine, the process of diagnosis is similar to annotation because it is also an attempt to classify observations within a known

pattern for use in making decisions about treatment and potential outcome. For example, in breast cancer, pathologists determine a TMN score (tumor size, metastatic state, and number of nodes involved), although ambiguities in this scoring exist (http://www.bioscience.org/atlases/tumpath/staging/llist.htm). In both instances, it should be noted that the use of annotation or classification can only reflect the accuracy of the classes or names that are being used to define the area, and frequently become dated as new information becomes available to support further 'evolution' of the classifications.

Ontology

Genomics has developed the use of ontologies to introduce some consistency into the list of parts and their hierarchical structure as observed in the genome and in fundamental biology. This provides a context in which these elements are defined and enables identification of potential variations in their specific functions based on environment or location. In medicine, one ontology represents the inherent hierarchical structure of an organism, which actually converges in a complementary manner with the ontology that is appropriately structured from the genome perspective. A second clinical ontology involves the definition of symptoms, clinical observations and definition of disease and staging, and is referred back to a standard classification of diagnostic codes (e.g. ICD-10). The definition of a specific diagnosis within these codes reflects a complex set of criteria, including insurance reimbursement and issues surrounding the expected standard of care, and can limit the value of claims-based clinical data for research evaluation.

Challenges and opportunities in biomedical informatics

There is a distinct relatedness between the first two comparisons (time and process) and the last two comparisons (annotation and ontology), described previously, as they provide insight into where biomedical informatics can make a significant contribution to achieving the paradigm shift depicted in Fig. 1. Therefore, the implementation of biomedical informatics (i.e. incorporating clinical data into the analysis of genetic information, and determining the mechanisms of disease progression), can address many of the issues facing the pharmaceutical industry as mentioned previously.

The similarity that time and process share between genomics and clinical observations, is that there is limited incorporation of the concept of 'process' rather than 'state'. Representing a vector by a point, or by a set of discrete observations that do not incorporate the time element, is very limited because of the loss of directionality and complexity when the vector exists in more than one dimension. Biomedical informatics can incorporate the perspective of process and time to improve the identification of biomarkers that are more mechanistically derived than correlatively derived in genomics, and can also stratify disease processes over time as a means of establishing disease subtypes for use in enhancing diagnosis and staging. This is of particular value in complex diseases [18] because of the inherent difficulty already observed in defining genomic traits through linkage analysis. The potential reduction of 'noise in the system' by grouping patients using more detailed information and according to phenotype should yield a better observation of genotypic relationships, both for disease risk and for response to their environment.

The similarity that both annotation and ontological approaches share between genomics and clinical observations stems from our limited ability to define accurately the crucial differences in specific descriptors, and an inherent human nature to develop classifications. The use of classification reflects a mechanism for simplifying complex information and enabling the complexity to be 'dealt with' in a manageable reference frame. Unfortunately, as science and technology continue to evolve, and vast amounts of data are generated for comparison, differences begin to appear that potentially disrupt the simple classifications. We are now in a transition period where we continue to make more observations than we can currently process effectively to evaluate their significance. Furthermore, their value might not simply be as independent observations, but rather as a relationship to other observables. These relationships will need to be evaluated in a way that is not only correlative, as noted previously, but also focuses on the potential biological mechanisms that produce this interaction of parameters. This is a fundamental component of biomedical informatics and perhaps the area where it will have its greatest impact on the integration of bioinformatics and medical informatics.

In breast cancer, for example, there is a significant difference in defining an ontology of: (1) cell components and structures that occur in the breast by location; and (2) the development process, from pre-natal through menarche to menopause of the breast structures, and where the deviation develops in the presentation of disease. The development of a process-based ontology that reflects the developmental phases of the breast cancer is being used as a data model for the integration of clinical and genomic information. A significant goal of this data mapping is to better enable the clinician and/or researcher to establish what is meant by 'normal', based on developmental phase, in the search for mechanistic changes associated with breast cancer progression.

Biomedical informatics: addressing crucial issues

Biomedical informatics will add complexity to the system rather than making things simpler, but hopefully, will only incorporate the complexity that really exists. Some of the issues that need to be addressed are described in the next section.

Aging

Although aging is a process that operates independent of the disease, it still has an impact. In addition, it is not clear whether all physiological systems evolve at the same rate within an individual, or whether this is reflected in the genotype. For example, breast cancer that appears pre-menopause (before the age of 40) might have a strong potential for genetic linkage, while post-menopausal breast cancer (after the age of 50) is termed 'sporadic'. One model of the menopause describes the normal aging of the female reproductive system, and effectively enables the separation of disease analysis at varying stages of aging [19].

Environment

The impact of the environment has been acknowledged as a significant component of disease complexity, but one whose evaluation is limited because of the complexity in defining it. It is possible that the use of isolated populations, now studied for genomic linkage but having limited success in extending specific loci to the general population, could present an opportunity to study environmental impact that also reflects the isolation of that population. For example, the identification of genetic loci with an apparent link to depression in isolated populations (e.g. Finland and Costa Rica), does not necessarily give a general view because of potential interactions with the local environment, and the potential effects of additional modifier genes. In an attempt to identify such factors, an analysis of sickle cell disease and sickle cell trait is being initiated to compare the effects of environment and healthcare in the 30 million Nigerians that exhibit this trait, and in the African American population within the United States.

Disease association

Attempts to define genetic links in complex diseases have been limited by the non-homogeneous nature of the diseases themselves. However, patient populations can be stratified on the basis of the disease process on the assumption that genotypic and environmental contributions will be more similar in disease substrata as opposed to across disease substrata. For example, in sickle cell disease and sickle cell trait, and thalassemias, not all individuals with the same β-globin mutation profile exhibit the same clinical manifestations of disease. By analyzing the progression of disease using clinical observations and symptoms, patient populations can be clustered to reflect those exhibiting greater overall similarity; these clinically defined substrata can then be further evaluated for similarity in genetic or environmental factors. This is analogous to the general search for genetic linkage in patients exhibiting depression, which is significantly limited by the inability to develop a quantitative diagnosis of the disease that can associate patient groups with relevant disease subtypes; this, in turn, creates 'noise' in the analysis, and an inability to identify potentially weak signals.

Model organisms

Better recognition of the physiological systems of model organisms that distinguish them from humans might help to refine the choice of system to study. For example, breast development (both normal and abnormal) is studied in mice; however, mice do not undergo menopause, and this is a significant risk factor associated with sporadic breast cancer in humans (>90% of all breast cancer). Many of the existing animal models are currently selected for convenience, but as genome sequences for these organisms become available, it will be possible to identify, at the pathway level, not only topological analogy or homology with humans, but also the physiological responses for the whole pathway.

Opportunity and challenge

Currently, methodologies to add these levels of complexity do not exist solely within bioinformatics or medical informatics, and might exist either in other scientific or engineering disciplines, or not at all. The opportunity to use methods that have evolved to handle such complexities, as in engineering (e.g. stochastic modeling [20], control theory [21] and risk analysis [22]), and to evaluate their performance in these biologically significant areas, is just beginning to be explored [19]. Methods initially used might not succeed or even address the correct question; the available data might not be adequate in content or resolution to support the analysis. The opportunity and the challenge is to prioritize the complexities that need to be addressed, and to start the evaluation of how to incorporate them into the process of problem solving. This is potentially a more accurate definition of systems biology than simply the integration of genomics, proteomics and so on, because it reflects the physiological basis for their integration rather than the technological basis.

Concluding remarks

Medicine is at a significant crossroads in its development. Although many innovations in technology have made the 'practice' more efficient and effective, nothing has yet had the potential for revolutionizing the 'science' more than the results of the human genome project and its continuing generation of information and knowledge about disease processes and individual differences. The implementation of this new knowledge will present a new set of challenges and opportunities for the pharmaceutical industry that are only just beginning to be recognized.

Is the pharmaceutical industry prepared to enter this 'brave new world' and, if not, can it develop a proactive (strategic) rather than reactive (tactical) response to these changes? Such a question remains to be answered. Within the pharmaceutical industry today, clinical data are thought to refer to clinical trial data. Diagnosis by the physician or pathologist is relied upon to be more exact than it deserves. It could also be that the existing infrastructures within the industry, including organizational

and cultural characteristics, could be resistant to evolutionary change. If so, it will probably be the virtual pharmaceutical companies, created by the biotech industry, with a virtual vertical structure for drug development that will be more adaptable and more able to generate and test a new model, and one that incorporates biomedical informatics.

Major pharmaceutical companies have several potential factors on which it can capitalize to facilitate this evolution. Understandably, internally funded basic research in a pharmaceutical company might not have any of these factors, but it can still have a significant impact on development through appropriate alliances with academic research institutions. The key to moving forward is to recognize that systems biology is more than just an integration of the genomic and proteomic perspective; it also uses the long-standing observations of clinicians. Pharmaceutical companies have worked through their marketing groups to establish communication networks that enable market penetration and saturation for new products, but have not yet capitalized on using these networks for competitive advantage at a research and discovery level. The true capture and analysis of clinical data beyond clinical trials alone might be outside the scope of most pharmaceutical companies. However, the opportunity to use its physician networks to describe, for example, why a conventional diagnosis is not adequate, or what patterns of response appear within populations, could significantly enable the implementation of the first phase of biomedical informatics into the discovery process. Ideally, this should be used as early in the discovery phase as possible, rather than in the selection or refinement of candidates and leads. These physician networks, managed by highly trained marketing professionals, remain an untapped resource that could facilitate the successful evolution and interface of biomedical informatics.

References

1 Xie, D. *et al.* (2000) Population-based, case-control study of HER2 genetic polymorphism and breast cancer risk. *J. Natl. Cancer Inst.* 92, 412–417

2 Schulman, J.D. *et al.* (1996) Preimplantation genetic testing for Huntington disease and certain other dominantly inherited disorders. *Clin. Genet.* 49, 57–58

3 Ph.R.M.A. Organization (2001–2002) Annual report, Ph.R.M.A., Washington, DC

4 National Vital Statistics Reports (2000) Deaths leading causes, CDC, Atlanta

5 Brugnara, C. (1995) Erythrocyte dehydration in pathophysiology and treatment of sickle cell disease. *Curr. Opin. Hematol.* 2, 132–138

6 Collins, F.S. (1998) New goals for the US Human Genome Project: 1998-2003. *Science* 282, 682–689

7 Weinstein, J.N. (1998) Fishing expeditions. *Science* 262, 627

8 Joubert, M. *et al.* (1998) UMLS-based concept queries in biomedical informatics databases: an overview with ARIANE. *J. Am. Med. Inform. Assoc.* 5, 52–61

9 Stikeman, A. (2002) Systems biology. *Technol. Rev.* March

10 Giuse, D.A. and Miller, R.A. (1995) Potholes in the road to professionalism in medical informatics. *Methods Inf. Med.* 34, 434–440

11 Altman, R.B. and Klein, T.E. (2002) Challenges for biomedical informatics and pharmacogenomics. *Annu. Rev. Pharmacol. Toxicol.* 42, 113–133

12 Lewis, R. and Palevitz, B.A. (2001) Human genes, how many? The Scientist 15, 21

13 Huisman, T.H.J. *et al.* (1996) *Syllabus of Human Hemoglobin Variants, The Sickle Cell Anemia Foundation*, Augusta, Georgia

14 Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* 11, 1425–1433

15 Ferriman, A. (2002) Advocates of PSA testing campaign to silence critics. *Br. Med. J.* 324, 255

16 Malone, K.E. *et al.* (2000) Frequency of BRCA1/BRCA2 mutations in a population-based sample of young breast carcinoma cases. *Cancer* 88, 1393–1402

17 Shak, S. (1999) Overview of the trastuzumab (Herceptin) anti-HER2 monoclonal antibody clinical program in HER2-overexpressing metastatic breast cancer. Herceptin multinational investigator study group. *Semin. Oncol.* 26, 71–77

18 Liebman, M.N. (2002) Opening Pandora's Box: clinical data and the study of complex diseases. *Science* (http://stke.sciencemag.org/cgi/content/full/sigtrans;2002/130/pe20?ijkey=C6Dswql3VBKyM&keytype=ref&siteid=sigtrans).

19 Tsavachidou, D. and Liebman M.N. (2002) Modeling and simulation of pathways in menopause. *J. Am. Med. Inform. Assoc.* 9, 461–471

20 Fall *et al.* (2002) *Computational Cell Biology*, Springer Verlag

21 Kampis, G. (1991) *Self-Modifying Systems in Biology and Cognitive Science*, Pergamon Press

22 Evans, J.R. *et al.* (2001) *Introduction to Simulation and Risk Analysis*, Prentice Hall