Computational Disease Models
(Curator Thomas Lengauer)

1. Definition of research

The technological revolution in molecular biology that was initiated with the sequencing of the human genome also greatly impacts medicine. For the first time in history we have the chance to collect data in sufficient quality and quantity to begin understanding the molecular basis of diseases at a patient-specific level. Such understanding can place the prevention, diagnosis, prognosis and therapy of diseases on a completely new foundation. The perspectives are three-fold: (1) diagnosis can be based on a detailed molecular “fingerprint” obtained from the patient by a diagnostic assay, whose discriminatory power greatly exceeds what is possible with classical methods of diagnosis. (2) Increased understanding of the molecular basis of a disease can afford effective development of new drugs. (3) Prevention, prognosis and therapy selection are improved by providing molecular fingerprint data – which in some contexts are called biomarkers – to computational models of the progression of the disease in dependence of patient-specific factors and proposed therapies. Such models can have widely differing granularity. On the one hand, statistical models can estimate the chance of effectiveness of the therapy. On the other hand, and certainly more visionary, quantitative dynamical models of disease progression can simulate the development of the disease and the impact of countermeasures in dependence of the parameters related to patient, drug, and environment.

One of the major challenges in this area is computational. Progress in lab technology is dramatic; the volume of data that can be generated is steeply rising; the costs at which one can generate such data are collapsing. However, the interpretation of the data – reaping biological insight from measurement – remains a central challenge. Here, new mathematical, statistical and informatics methods are urgently required.

2. Status of the field

2.1 Diseases, in general

Disease-targeted biology research has entered a new era with molecular screening technology available at various levels (genome, transcriptome, proteome, metabolome, interactome, etc.). Such technology is available today in a research setting, whereby in carefully crafted studies the generated data are used to uncover aspects of the molecular basis of disease. At the same time, due to the substantial drop in cost, such data can increasingly be generated in clinical practice. For example, the perspective of being able to sequence individual patient genomes at very moderate cost affords completely new approaches towards elucidating the molecular basis of disease. Genome-wide association studies (GWAS)\(^1\) and gene-expression profiling\(^2,3\) can provide genomic markers for assessing the risk of acquiring a disease or for diagnosing disease. However, the statistical analysis of data is difficult. In the face of the fact that in complex diseases there are no single causes, a statistically stringent integration of multiple lines of evidence is a major challenge. New method development is highly critical.
Beyond the genome, the challenge of understanding what is happening in cells in healthy and diseased tissues and organs is the object of intense studies. Experimental methods of manipulating cells, e.g. via RNAi-based gene-knockdown\textsuperscript{4}, and producing novel readout with innovative labeling and imaging technology afford completely new approaches towards generating data on intracellular processes. These approaches belong to the greater field of medical systems biology, which uses data from cell-wide screening experiments together with computational analysis and modeling methods to derive hypotheses on molecular networks, here in the context of diseases\textsuperscript{5, 6}. Again, the resulting data need to be interpreted and crafted into computational models that simulate the disease process or predict effects of the disease and of therapies.

Beyond the static structural analysis of molecular interaction networks, quantitative dynamical models, e.g., of the viral life cycle inside the host cell, afford insight into the factors that influence virus replication\textsuperscript{7}. Such models can be used as building blocks for dynamical modeling of disease processes or can be applied to optimize virus yield for vaccine production\textsuperscript{8}.

Simultaneously, in the clinical setting data collection is becoming more systematic with the formation of high-quality databases on patient histories involving personal and clinical parameters\textsuperscript{9, 10}. The accompanying storage of biological tissues is often critical\textsuperscript{11}. There are also efforts underway to create databases with quite comprehensive goals covering many diseases and including storage of biological samples\textsuperscript{12}. These developments facilitate the availability of substantial data volumes that introduce the personal aspect of the patient into the analysis. One major challenge is to bring both lines of research together and thus empower the new technology not only in a research lab setting but in a clinical setting, i.e., bring it from the bench to the bedside\textsuperscript{13, 14}. Major information infrastructure is built up to facilitate this process\textsuperscript{15}.

As a point in case, for AIDS, computational support of the effective selection of combination drug therapies exists to date and is in clinical use\textsuperscript{16, 17}. Such an approach is also a perspective for other infectious diseases and beyond (e.g., cancer\textsuperscript{18}). Improving disease phenotype classification and corresponding semantic ontologies will also be required for better patient-specific models of disease processes.

### 2.2 Infectious diseases

For the purpose of focus, the subsequent discussion is restricted to infectious diseases. Such diseases combine several attractive properties for research: (1) They remain highly relevant and affect large patient populations. (2) They involve clear molecular targets pertaining to the life cycle of the pathogen and to host-pathogen interactions. (3) They are ever changing, due to rapid evolution of the pathogen and consequent adaptation of the immune system in the patient population. This poses special challenges to modeling approaches.
The status of the field of modeling infectious diseases can be summarized as follows:

Host-pathogen interactions

There are major systems biology efforts underway to identify host factors for several pathogens and the first respective databases are available (e.g. AIDS\textsuperscript{18-21}, hepatitis C\textsuperscript{22}, malaria\textsuperscript{23}, influenza\textsuperscript{24, 25}, tuberculosis\textsuperscript{26, 27}). There are also databases comprising information on host-pathogen interactions for several pathogens\textsuperscript{28}. For bacterial pathogens, modeling based on metabolic network models, partly including models for enzyme kinetics affords predictions of phenotypes such as growth rate under varying conditions\textsuperscript{29}. For viruses the same can be achieved based on models of the viral life cycle which are based on data achieved by specially tailored assays and screens\textsuperscript{30}. In general modeling biological processes based on host-pathogen interaction networks is still in early stages.

Vaccine design

Vaccination remains the most cost-efficient intervention measure in medicine\textsuperscript{31}. Vaccines against numerous diseases have proven their high value for public health. Global incidences of polio have been reduced to a few thousand cases annually, of measles to some hundred thousand cases annually and small pox has been eradicated globally. In all of these cases, vaccination was instrumental to success. The basis of successful vaccination thus far has been the induction of antibodies which attack microbial invaders soon after infection. However, for several major diseases, including HIV/AIDS, hepatitis C, malaria and tuberculosis, we have not been able to generate efficacious vaccines. Principally, efficacy of a vaccine depends on its antigenicity and its immunogenicity. Antigens employed for vaccination and can be selected based on the combined analysis of antigenic and genetic information based on phylogenetic methods and antigenic maps\textsuperscript{32}. Regarding immunogenicity, which describes the strength of the immune response elicited, strong responses can be achieved by adjuvants which elicit potent immune reactions even if the antigen does not do so by itself. Combination vaccination comprising a prime, e.g., with a viable vaccine, and a boost with a subunit vaccine will be most efficacious. This strategy proved successful for tuberculosis. Similarly, the first (minor) success of vaccination against HIV/AIDS followed a prime–boost strategy. Clinical trials of vaccines need to be monitored by a set of biomarkers which predict clinical end-point, i.e., disease outbreak. Biomarkers, which serve as surrogates of protection (i.e., provide educated information whether a vaccine candidate will induce protection or not) can markedly shorten trial duration. This is also the best way to save cost. Reliable surrogates of protection could have predicted failures of previous clinical trials and thus saved large financial investment. In general, modeling of vaccination and disease processes can lead to improved vaccination strategies and facilitate definition of biomarkers. This is best done in an iterative process which includes data from wet-lab as well as ongoing clinical trials\textsuperscript{26, 33}.

Analysis of pathogen evolution

This topic has two facets. One is the evolution of the pathogen in the human population, mostly due to evasion of the host immune system. A prominent example is influenza, which rapid antigenic and genetic evolution due to selection for antigenic change\textsuperscript{34, 35} The evolutionary dynamics (phyldynamics) of influenza can be analyzed with phylogenetic methods and, in combination with phenotypic
information on the individual strains, the resulting models provide the basis for the selection of viral strains for vaccine design\textsuperscript{36}.

The second facet is the evolution of the pathogen to resistance. This phenomenon happens inside the individual patient and has been studied especially intensively with HIV. Databases have been developed that collect clinical data on viral resistance\textsuperscript{9,10}. On the basis of these data statistical models have been derived that have predictive power with respect to the resistance of individual viral strains obtained from patients\textsuperscript{37}. Moreover statistical models have been developed that predict viral escape from individual combinational drug therapies and estimate therapy effectiveness based on this prediction\textsuperscript{38}.

\textit{Epidemiology}

One characteristic of infectious diseases is that epidemiology is a key factor in understanding the risks involved. Epidemiological models go beyond biology and have to incorporate demographic data and social aspects of human behavior\textsuperscript{39}. The resulting simulation models elucidate the dependence of the spread of the disease on a broad range of factors and afford prediction of risk of infection depending on such factors. Furthermore, such studies are important for the adequate selection of pathogen strains for vaccine design\textsuperscript{32}.

3. \textbf{Research opportunities and needs}

Research challenges fall into two categories. First, in most subtopics of the field research is in an early phase. The individual research topics need to mature in several aspects.

i. Data collection: The coverage of data needs to be increased. For many of the relevant questions specific assays or screens must be developed. However, the basic underlying screening technologies are in a rapid stage of development and costs are decreasing dramatically. In the clinical setting the scope of systematic data collection can and needs to be increased. A special issue regarding data quality for several diseases is the identification of quantitative disease-related phentotypes\textsuperscript{40,41}.

ii. Data curation: Many data, especially those generated with less mature technologies are quite prone to experimental variation. Assuring data quality and comparability between data generated in different labs or at different times is not just a matter of disciplines but is a research topic in itself and is far from being solved. Statistical error models that separate technological from biological variation, are adequately validated and have diagnostic value are missing in many settings.

iii. Data availability: Whereas biological data enter the public domain in increasing volumes, clinical data are still largely unavailable to the public. The reasons are a mixture of privacy issues and reluctance of the data owners to share their data. Such attitude has caused a major blockage of the deposition of influenza data which could only recently be remedied by the world-wide GISAID initiative\textsuperscript{42}. In order to best support research, clinical data must also be complemented by the availability of the respective biological samples. In general progress on patient-specific methods continues to be substantially hindered by the fact that the relevant
data are not available to the broader scientific community. To resolve this problem, a multitude of legal, ethical and organizational issues have to be resolved.

iv. Method development: Here the two major issues are (1) the development of statistical methods that deal with sparse data sets in high dimensions and curb the risk of generating extraordinary large amount of false positives and (2) the development of methods for simulating dynamical stochastic systems. These challenges are not particular to the medical field but apply to the whole field of systems biology.

v. Visualization: Biological processes acting in complex biochemical networks and having stochastic behavior are inherently hard to illustrate and capture at an intuitive level. The tools for visualizing the input, structure and output of the respective models need to be further developed. Special emphasis must be taken that the resulting models are interpretable and that their output is presented to the medical professional, who is to apply them, in a fashion that makes the results transparent with respect to their content and their reliability.

vi. Validation: In a setting in which models are inherently partial and inaccurate convincing validation is of the essence. Validation methodology has to be brought forward in order to assess the reliability and accuracy of models and enhance the comparability of different studies.

vii. Legal and ethical issues: Privacy protection and sufficient coverage of the developing world in studies and with respect to distribution of the new technology and its products are important topics.

With respect to host-pathogen interactions the major challenge is to provide complete and accurate spatio-temporally resolved molecular interaction networks and to bring the analysis forward from the elucidation of the static structure of the networks to their dynamic behavior. This involves not only dynamic models of pathogen replication in individual cells but also a quantitative description of pathogen release and distribution in tissues and organs. Based on such models, the impact of vaccinations or the use of antiviral inhibitors can be evaluated on a more quantitative basis. More long-term goals are to bring essential parameters into the network such as the genetic variability of host and pathogen, aspects of the patient history, environmental factors and the effect of the applied drugs. With respect to pathogen evolution the technology developed for HIV can be extended to other pathogens such as HCV and HBV, for instance. Furthermore, we need more accurate models of viral escape that are not only based on larger bodies of clinical data comprising longitudinal samples more so than it is the case today.

The second and major challenge, though, is to bring the separate modeling efforts that focus on different aspects of the disease process together to cover a broader range of disease aspects in an integrated fashion. E.g., host-pathogen networks should also cover the genetic variability of host and pathogen, aspects of the patient history, environmental factors and the effect of the applied drugs. Models of pathogen evolution should be based at least in part on biomechanistic models of the life cycle of the pathogen. Epidemiological risk analysis should incorporate the analysis of the evolution of the pathogen. Also, drug application scenarios have an epidemiological component, which needs to be analyzed and integrated. Furthermore integration efforts have to span tremendous time scales. Chemical reaction happen in microseconds or below. Disease progression and healing requires days or
even years. So, there is a need of integration over several levels of abstraction wherever we look. This integration task is daring, but it is common to many other areas in today’s life sciences.

4. Perspectives for the next five years

Many of the perspectives outlined above are quite long-term. However, we expect that substantial progress will be made in the next five years. Host-pathogen interaction networks will grow in coverage and accuracy. So will the knowledge on host-drug interactions via pharmacogenetic studies. With a mixture of screening experiments and the computational interpretation of the resulting data, on the one hand, and targeted biological experimentation, on the other hand, we will be able to identify additional critical host factors and further the understanding of pathogen strategies and host defense. This knowledge will afford new approaches to drug design. Similarly, these approaches will help identification of conserved antigens and provide deeper insight into the mechanisms of vaccine-induced protective immunity. Software for suggesting combination drugs therapies will become more accurate and incorporate aspects not covered today, such as genetic host variation, patient history and a growing number of clinical correlates. We hope to progress in making clinical data available to the public although we realize that this is quite a challenge, non-scientific but essential. In summary, we expect that information based on computational disease models will come to the benefit of an increasing number of patients and substantially reduce the risk of inadequate or even damaging therapy.

5. At a glance

Screening technology in modern biology affords the comprehensive generation of patient-specific molecular data. Such data harbor invaluable information for the diagnosis, prognosis and therapy of diseases that can be harvested only with complex computational disease models.

6. References