Chapter 1
Informatics for Infectious Disease Research and Control

Vitali Sintchenko

1.1 Introduction

Infectious disease informatics has been defined as a new field that studies knowledge creation, sharing, modeling and management in the domain of infectious diseases (Zeng et al. 2005). Its emergence has been fueled by rapid increases in the amount of biomedical and clinical data, and demands for data analyses. The resulting combinations of experimental and informatics evidence have reshaped the ways of conducting infectious disease research, raising the expectation of better control of infectious diseases. The authors of this book argue that informatics has not only changed the scale on which the infectious disease research is being done but has also conceptually opened up different ways of managing patients and making discoveries in the field of infectious diseases.

The goals of infectious disease informatics are lofty and include the optimization of the development of antimicrobials, the improved design of more effective vaccines, the identification of biomarkers for transmissibility and clinical outcomes of infectious diseases, and a better understanding of host-pathogen interactions. In the last two decades, the emergence of new informatics methods and integrated databases has facilitated the realization of these goals. This chapter outlines the major challenges and opportunities that infectious disease informatics faces in the twenty-first century.

V. Sintchenko
Centre for Infectious Diseases and Microbiology, Syndey Medical School,
The University of Sydney, Sydney, NSW 2006, Australia

V. Sintchenko (ed.), *Infectious Disease Informatics*,
DOI 10.1007/978-1-4419-1327-2_1, © Springer Science+Business Media, LLC 2010

1.2 Handling New Data Types

1.2.1 Microbial Genome Assembly and Annotation

"New Age" infectious disease informatics rests on advances in microbial genomics, the sequencing and comparative study of the genomes of pathogens, and proteomics or the identification and characterization of their protein related properties and reconstruction of metabolic and regulatory pathways (Bansal 2005). The speed of microbial genome sequencing has been steadily accelerating since the introduction of modern DNA sequencing methods more than thirty years ago (Sanger et al. 1977). The accumulation of sequenced genomes of bacteria shows a good fit to exponential functions with a doubling time of approximately 20 months (Koonin and Wolf 2008). Despite the historical bias towards the "working horses" of bacterial genomics, such as commensals *E. coli* and *B. subtilis* (Collado-Vides et al. 2008), the depth and breadth of the coverage of sequences belonging to different species of viral, bacterial, fungal and protozoan pathogens has been rapidly expanding.

Microbial genomes are thousands or millions of base pairs in length, requiring both a global view of the genome and the ability to zoom in on details for the purpose of analysis and annotation. Annotation is the extraction of biological knowledge from raw nucleotide sequences (Médigue and Moszer 2007). Such decoding of the genomes allows the prediction of protein-coding genes and therefore, the proteins the organism is able to produce. Desktop computer sequence editors such as Chromas Lite (http://chromas-lite.software.informer.com/), Trace Edit (http://www.ridom.de/traceedit/) or commercial products like LaserGene (http://www.dnastar.com/products/lasergene.php) or Sequencher (http://www.sequencher.com/) are helpful in the initial sequence assessment. The task of assembling of sequences from re-sequencing experiments, when a reference sequence is available, can be supported by tools like TraceEditpro (http://www3.ridom.de/traceeditpro/) or SeqScape.

Different software pipelines have been developed to automate microbial genome annotation and assembly (Table 1.1). The Integrated Microbial Genome (IMG) system, hosted by the Joint Genome Institute (JGI), and the RAST (Rapid Annotation using Subsystem Technology) server are examples of open resources. Major sequencing centers offer genome viewers and browsers through their websites (McNeil et al. 2007). For example, Manatee (J. Craig Venter Institute (JCVI)) has been developed to view and to alter initial automatic annotations of prokaryotic genomes. The Sanger Institute's Pathogen Sequencing Unit has been maintaining freeware for sequence analysis, viewing and annotation, such as Artemis and the Artemis Comparison Tool (ACT) (Carver et al. 2008). The alignment of genomes of three strains of *Staphylococcus aureus* using ACT is shown in Fig. 1.1. Alternatively, multiple genome alignments in the presence of large-scale evolutionary events, such as rearrangement and inversion, can be efficiently constructed and visualized using the Mauve program (http://gel.ahabs.wisc.edu/mauve/download.php) (Darling et al. 2004). These tools assist in the rapid identification of protein-coding

Table 1.1 Bioinformatics analysis tools

Analysis tasks	Tools	URL
ORF or gene identification	ORF Finder	http://www.ncbi.nlm.nih.gov/gorf.html
	GeneMark	http://opal.biology.gatech.edu/GeneMark/genemarks.cgi
	GLIMMER	http://www.cbcb.umd.edu/software/glimmer/
Sequence alignment	ClustalW	http://www.ebi.ac.uk/clustalw/
	Tcoffee	http://www.tcoffee.org/Projects_home_page/
	MUSCLE	http://www.drive5.com/muscle/
Genome annotation	RAST	http://rast.nmpdr.org/
	Artemis and ACT	http://www.sanger.ac.uk/Software/
	IMG	http://rast.nmpdr.org/
	MAUVE	http://genome-alignment.org/mauve/
Phylogenetic analysis	Phylogeny programs	http://evolution.genetics.washington.edu/phylis/software.html
	SplitsTree	http://www.splitstree.org
	MEGA	http://www.megasoftware.net
Microarray analysis	Gene Expression Omnibus	http://www.ncbi.nih.gov/geo/
	Microarray informatics EBI	http://www.ebi.ac.uk/microarray
Metabolic pathway analysis	KEGG	http://www.genome.ad.jp/kegg/kegg2.html
	UniPathway	http://www.grenoble.prabi.fr/obiwarehouse/unipathway
Whole genome visualization	BacMap	http://wishart.biology.ualberta.ca/BacMap/index_2.html
	GenomeAtlas	http://www.cbs.dtu.dk/services/GenomeAtlas/

genes, as well as other features like non-coding RNA genes, repetitive sequences or recently acquired DNA.

Web servers like Integrated Microbial Genomes (Joint Genome Institute; http://img.jgi.doe.gov) or the Bacterial Annotation System (BASys, http://wishart.biology.ualberta.ca/basys/cgi/submit.pl) also support comparative analysis and the automated annotation of bacterial genomic (chromosomal and plasmid) sequences (Van Domselaar et al. 2005). They accept raw sequence data and gene identification information, and provide textual annotation and hyperlinked image output.

Strings of nucleotides are assembled into draft sequences that can be characterized by the following: (1) > 90% of genome in contigs, (2) average contig length > 5 kb, (3) >90% of a set of conserved genes present, (4) contig N90 length > 5 kb, (5) >90% of bases > 5× read coverage, (6) scaffold N90 length > 20 kb. The information used to annotate genomes comes from three types of analysis: (1) ab initio gene finding programs, which are run on the DNA sequence to predict protein coding genes; (2)

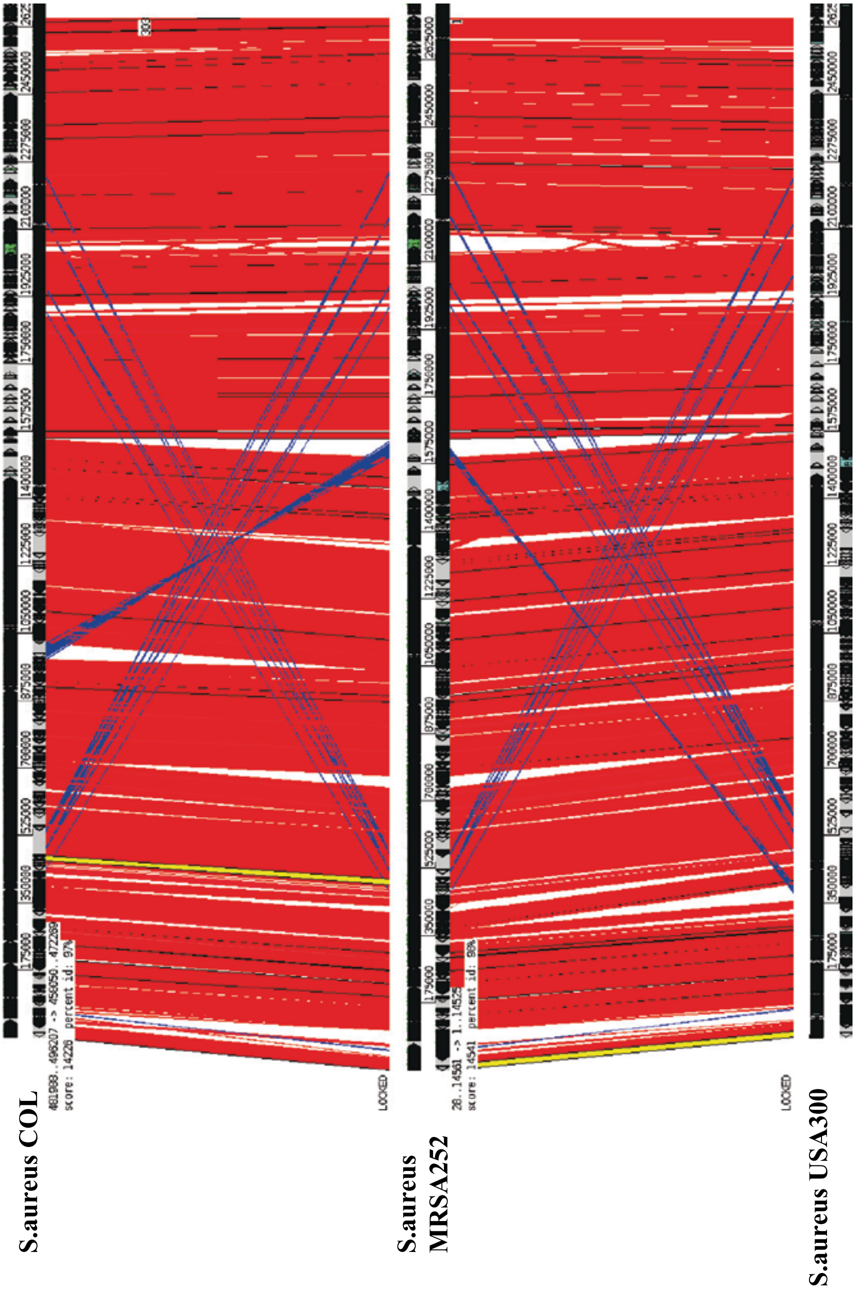


Fig. 1.1 Alignment of genomes of three strains of *Staphylococcus aureus*. DNA sequences that find a perfect match are connected with red lines or blocks. *Blue areas* are inversions or transitions and *white areas* represent indels. The figure was produced using Artemis software (The Wellcome Trust Sanger Institute, UK)

evidence-based gene calling or translating alignments of the DNA sequence to known proteins; and (3) aligning cDNAs from the same or related species. Gene finding has progressed far beyond the simple identification of open reading frames. The programs aligning cDNA and protein sequences to genomic DNA can locate the protein coding regions by searching the publicly available databases or by applying machine learning algorithms such as Hidden Markov Models (HMM). There is a long list of such programs including GeneMark, mORFind, PRODIGAL (Prokaryotic Dynamic programming Genefinding Algorithm), Argon and GLIMMER (Gene Locator and Interpolated Markov Modeller) (Delcher et al. 1999; Suzek et al. 2001; Majoros 2007). They differ in the time required for automated annotation as well as the quality of gene calling (Guigo et al. 2006). Problems with the accuracy of current gene finders reflect not only the performance of their algorithms but also the quality of the primary resources and the abundance of non-coding DNA regions in microbial genomes. Genome assembly annotation methods and tools including new applications for RNA genes, were reviewed in detail elsewhere (Stothard and Wishart 2006; Médigue and Moszer 2007; Brent 2008; Pop and Salzberg 2008).

Recent breakthroughs in high-throughput sequencing technologies have posed new challenges for genome assembly, annotation and analysis. These technologies make it feasible to sequence not only static genomes but also entire transcriptomes expressed under different conditions (Shendure and Ji 2008). However, they can produce read lengths as short as 35–40 nucleotides, which cannot be analyzed with software developed for Sanger data as they are often non-unique, lack neighborhood context and have a different distribution of errors. The task of linking such short-reads may be accomplished using a comparative assembly algorithm, in which new sequences are put together by mapping them onto close relatives or the “reference genomes.” Not surprisingly, the comparative assembly strategy works best when the two species are more than 90% identical. Alternatively, when no “reference genome” is available, the new cohort of assembly algorithms based on de Bruijn graphs – a way to transform sequence data into a network structure – has risen to the task (Chaisson and Pevzner 2008; MacLean et al. 2009). Strategies and systems that address these new challenges have recently been reviewed elsewhere (Pop and Salzberg 2008; MacLean et al. 2009; Ussery et al. 2009). Tables 1.1 and 1.2 provide examples of informatics tools for pathogen annotation and analysis.

1.2.2 Meta-Omics: Metagenomics and Metaproteomics

The metagenomics or the sequencing of genomes of complex mixed communities has emerged at the interface of genomics, microbiology and information technology. This field examines the interplay of hundreds of microbial species present at specific sites of potential infections in space and time (Hutchinson 2007; Smarr et al. 2009). Significantly, metagenomics has extended its focus from environmental microorganisms to microbial communities or “community whole genome sequences” of the human host (Field et al. 2006; Verberkmoes et al. 2009).

Table 1.2 Examples of bioinformatics resources for pathogens with epidemic potential

Analysis	Tools	URL
Sequence databases and tools	GenBank	http://www.ncbi.nlm.nih.gov/sites/entrez
	Protein Data Bank	http://www.rcsb.org/pdb/
	Microbial Genome Database	http://mbgd.genome.ad.jp/
Workbenches	Virology on the WWW	http://www.virology.net
	Viral Bioinformatics	http://www.biovirus.org
	Research	http://www.microbase.gr
	Microbase	http://xbase.bham.ac.uk/
	xBASE	
	SEED	http://www.theseed.org
	Influenza Virus Resources	http://www.ncbi.nih.gov/genomes/FLU/FLU.html
		http://www.biohealthbase.org
		http://www.flu.lanl.gov/
Pathogen specific datasets	European Hepatitis C	http://euhcvdb.ibcp.fr/euHCVdb/
	database	http://hcv.lanl.gov/content/hcv-db/index
	Hepatitis C database	http://www.hiv.lanl.gov/content/index
	HIV databases	
	Poxvirus Resource	http://www.poxvirus.org
	SARS Bioinformatics Suite	http://athena.bioc.uvic.ca/database.php?db = cooronaviridae
	DengueInfo	
		http://www.dengueinfo.org
	Neisseria.org	http://neisseria.org/
	TB Database	http://www.tbdb.org/
	Plasmodium Genome Resource	http://plasmodb.org/plasmo/
Antimicrobial resistance	ARDB	http://ardb.cbcb.umd.edu
	ARGO	http://www.argodb.org/
	Compendium of TEM genes	http://www.lahey.org/studies

Most of the 10–100 trillion microorganisms in the human gastrointestinal tract live in the colon (Turnbaigh et al. 2007). The genomes of these microbial symbionts have been collectively defined as the microbiome or ecosystem in which the number of microbial genes is estimated to be many folds higher than those present in the human genome. The Human Gut Microbiome Initiative, a logical conceptual extension of the Human Genome Project, aims to discover genomes of at least 100 new intestinal species. This approach has targeted the totality of genes involved in the gut biofilms, the mechanisms of horizontal gene transfer, and the role of the microbial pan-genome (Field et al. 2006). The Microbiome project aims to address some of the most inspiring and fundamental scientific questions today in order to identify new ways to determine health and predisposition to diseases and define parameters needed to design, implement and monitor strategies for intentionally manipulating the human microflora (Turnbaigh et al. 2007).

1.2.3 Global Genome Analysis

In addition to conventional strings of nucleotides, large-scale sequencing can provide new types of data reflecting global genome architecture and the properties of pathogens. These data include the size of a genome and its nucleotide composition, the locations of genes and intergenic regions, GC percentage and gene density. Microbial genomes are compared by the number of particular sets of genes, gene order (synteny) and the presence or absence of important genes. Other metrics include gene set properties (the number of two component system regulatory genes) and nucleotide sequence-based measures (distance between paired two-component system genes and consensus sequence) (Whitworth 2008; Ussery et al. 2009). These metrics represent a global view of genomes but often have limited biological meaning. Thus, "signature" sequences have been suggested as a means of identifying organisms or genes with sequence profiles correlating with the pathogen phenotype or disease outcomes. Examples of genome characteristics that are more directly related to biologically important behavior are bacterial IQ (a measure of the number of signal transduction proteins as a function of genome size) and extrovertedness (the proportion of signaling proteins predicted to sense external stimuli) (Galperin 2005).

Analyses of genomics data challenge the traditional taxonomy of microbial species. Recent projects have focused on producing simple analytical diagnostic tools based on strong taxonomic knowledge collated in the DNA reference libraries such as the DNA Barcode of Life Data System (BOLD; http://www.boldsystems.org). These types of data enable the acquisition, storage, analysis and publication of DNA barcode results, and provide clues about the global distribution of species. Their genetic diversity and structure is based on two postulates: first, that every species is represented by a unique DNA barcode (indeed there are 4^{650} possible ATGC combinations compared to an estimated 10 million species remaining to be discovered (Frézal and Leblois 2008)), and second, that the genetic variation between species exceeds the variation within species. DNA barcoding requires a minimum sequence length of 500 bp and more than three individual sequences per species. The initial Barcode of Life framework was based on the sequence of a single universal marker – the cytochrome c oxidase gene – but has evolved since then, giving rise to a flexible description of DNA barcoding, a larger range of applications and the broader use of the term "barcode" (Frézal and Leblois 2008). For example, the whole microbial genome's barcodes were defined as frequency distributions of periodic DNA sequences or *k*-mers across the whole genome (Zhou et al. 2008). It has been postulated that such barcode similarities are proportional to the genomes' phylogenetic closeness and could be utilized in metagenome analyses (Zhou et al. 2008).

Microbial species diversity can be also estimated by the average nucleotide identity (ANI) using the list of orthologs and deriving the overall divergence of the core genome by averaging the percentages of identity at the nucleotide level (Konstantinidis and Tiedje 2005). Another approach to measure distances between genomes is based on estimating the proportion of common genes by calculating the ratio of orthologs to the total number of genes of the reference genome. More recently, similar methods such as DNA content, BLAST distance phylogeny and the MUM (maximal unique and exact matches) index have been suggested as more sensitive measures for intra-species comparisons (Deloger et al. 2009).

1.3 Changing the Way Discoveries Are Made

1.3.1 Knowledge Discovery from Comparative Genomics

The true power of large-scale comparative genomic studies lies in their ability to identify and characterize biological trends and rules that explain particular phenomena (Field et al. 2006). Computational methods have become essential steps in formulating hypotheses about gene functions. The comparative approach has not only yielded fundamental insights into the function and evolution of microbial genomes, but has also led to practical results. Comparative genomics has allowed the accurate estimation of the structure of genomes and the speed of gene movements, including the role of natural selection versus genetic drift, the origin of the pandemic strains, and the ecology of a pathogen in its natural reservoir (Chen et al. 2005; Yang et al. 2008a). Computational studies identified unexpected relationships between genomic features and ecological niches, demonstrated diversity in the microbial world and helped to reconstruct evolutionary relationships among genomes (Binnewies et al. 2006; Field et al. 2006).

Comparisons made between different genomes can also generate new hypotheses for testing, usually relating to the unexpected presence or absence of particular genes with respect to other genomes (Whitworth 2008). The studies of three main forces shaping genome evolution – gene loss, gain and change – have been especially fruitful in this respect (Burrack et al. 2007; Whitworth 2008). Discoveries of gene duplication in many bacterial pathogens, resulting in increased numbers of key gene clusters or the expansion of important protein families have led to the development of new diagnostic methods. For example, the gene clusters encode a secreted protein called the early secretory antigenic target 6 or ESAT6, which was identified as one of the key virulence factors in *Mycobacterium tuberculosis* and was subsequently used in the interferon-gamma release assays for the diagnosis of tuberculosis (Pallen and Wren 2007; Behr 2008).

Comparative genomics has also revealed that pathogens undergo a process of genome decay or a reduction in the number of biosynthetic pathways, resulting in a dependence on the infected host for certain essential functions. The most surprising snapshots of genome decay have come from relatively recently emerged pathogens that have changed their lifestyles by adopting a simpler host-associated niche. For example, the genomes of *Yersinia pestis* (Parkhill et al. 2001b) and *Salmonella enterica* serovar Typhi (Parkhill et al. 2001a) contain hundreds of pseudogenes. These findings challenge the traditional view that bacterial genomes never contain "junk" DNA and that every gene in a bacterial genome must have a function. Instead, every genome should be viewed as a work in progress, burdened with some non-functional "baggage of history" (Pallen and Wren 2007).

As the smallest-scale variation in microbial genomes occurs at the level of single-nucleotide polymorphisms (SNPs), SNP detection has been applied extensively to many pathogens (Yao et al. 2008). While SNPs are generally considered rare, at one per several thousand base pairs, two genomes of *M.tuberculosis* of 4 Mb each may have some 1,0002008 SNPs between two isolates (Behr). Whole-genome sequencing has been proven as an even more powerful tool to detect SNPs. It enabled the differentiation of *Escherichia coli* strains that had diverged for as few as 200 generations (Shendure and Ji 2005) and revealed genomic changes in pathogens in the process of human infection (Chen et al. 2006; Forst 2006; Pallen and Wren 2007).

1.3.2 Automatic Recognition of Functional Regions

In the pre-informatics era, virulence factors were typically identified either by biochemical studies or through genetic screens. Informatics has enabled innovative strategies for the recognition of virulence gene recognition through the analysis of genetic signatures (Pallen and Wren 2007). Despite the variety of microbial life styles and associated genomic and metabolic complexity, pathogen genomes share common architectural principles. As a result, computational techniques assist in exploring similarities between virulence factors and other genes with known functions. This association can then be tested using targeted genetic methods such as the inactivation of the putative virulence gene followed by the comparison of phenotypes of the original and modified microorganisms (Chen et al. 2005; Raskin et al. 2006). A strategy that does not rely on sequence similarity for identifying potential genes is the detection of coding sequences, which is based the gene context "grammars" supplemented with machine learning models (Garrido et al. 2008). For example, functional gene recognition tools GeneMark and GLIMMER employ Hidden Markov models, in which the preceding nucleotide bases are used to predict the next base in a coding region, and the algorithm is trained on a trusted set of sequences. Gene coding regions are then identified using probability estimates of the correct coding "grammar" in a region (Dougherty et al. 2002). Different statistical and machine learning methods for gene prediction have been reviewed elsewhere (Majoros 2007).

Gene-gene interactions specifically associated with a phenotype or a particular disease can be explored with or without a prior biological knowledge. Several techniques utilizing Bayesian networks, pair-wise mutual information and graphical

Gaussian models have been proposed for this purpose. Coupled with biological knowledge, the identification of such phenotype-specific interactions can shed light on the responsible pathways. The complexity of data handling and visualization has led to efforts to develop dedicated comparative genomics resources such as GenDB (Meyer et al. 2003), CMR, ACT, (Table 1.1) xBASE and Microbes OnLine as well as data management systems such as SEED (Table 1.2) (Chaudhuri et al. 2008).

1.3.3 Enabling the Dynamic View of Infectious Diseases

Informatics has been instrumental in the change from static to a dynamic view of the microbial world. In contrast to the static view of genome annotations focused on the gene or protein prediction, the dynamic view places information obtained into a biological context to identify interactions between the genomic components and the reconstruction of regulatory networks (Médigue and Moszer 2007; Sakata and Winzeler 2007). Under the network vision of the microbial world, microbial chromosomes are not envisaged as strictly defined genotypes gradually changing in time but rather as islands of temporary, relative dynamic stability that form tightly connected (vertically and horizontally) areas of the network (Koonin and Wolf 2008). The infection cycle should be considered as a whole and the links between growth, virulence, immune evasion and transmission should be assessed (Restif 2009).

Biological interactions vary in their nature and are spatially and temporally heterogeneous. One can abstract the actions of proteins and metabolites by representing genes acting on other genes as a gene network or as genetic regulatory, transcription or expression networks. Such networks can be constructed using computationally assigned functional linkages inferred by Rosetta Stone, Operon or similar methods (Rachman and Kaufmann 2007; Harrington et al. 2008), and often point to highly connected and central proteins frequently referred to as "hubs" (Wu et al. 2008). Biological interaction and communication networks share several commonalities: they are scale free (only a few nodes are highly connected) and are small world networks (highly clustered with short distances between any two nodes) (Kann 2008). Increasingly, disease pathogenesis and the mechanisms of drug action are viewed from a biological systems perspective (Wu et al. 2008). From this perspective, a deeper understanding of infectious diseases may rely on an exhaustive characterization of all potential interactions occurring between proteins encoded by viruses and those expressed in infected cells. Thus, the integration of all protein-protein interactions into an infected cellular network, or "*infectome*," offers a powerful framework for the virtual modeling and analysis of infections (Navrati et al. 2009). The terms "*interactome*" and "*phenomics*" have been coined in this context (Lussier and Liu 2007).

Numerous resources have been developed to explore host-pathogen interactions (PHI) (Table 1.3). Specifically, PHI-base (Winnenburg et al. 2006), PHIDIAS (Xiang et al. 2007), BioHealthBase (Squires et al. 2008), PIG (Driscoll et al. 2009) VirusMINT (Chatr-aryamontri et al. 2009) and VirHostNet (Navrati et al. 2009) have been

Table 1.3 Knowledge discovery tasks from the host-pathogen interactions

Levels	Microbial genomes	Microbial proteins	Microbial metabolome	References
Human proteins	Gene-protein interactions, networks, defining protein functions	Protein-protein interactions, protein structure prediction, epitope mapping		An and Faeder 2009 Chatr-aryamontri et al. 2009 Driscoll et al. 2009 Garrido et al. 2008 Kann 2008 Xiang et al. 2007
Human metabolites	Pathway mapping and reconstruction	Protein function prediction	Pathway comparison	Lisacek et al. 2006 Winnenburg et al. 2006
Human phenome	Genotype-patient outcome mapping	Disorder prediction, virulence prediction	Biomarker discovery	Burrack and Higgins 2007 Forst 2006
			Virulence prediction	Lengauer et al. 2007
			Drug resistance prediction	Navrati et al. 2009 Raman et al. 2008
	Effect of diseases on gene expression	Drug target identification		Reddy et al. 2009 Squires et al. 2008
	Disease reclassification	Drug resistance prediction		Stavrinides et al. 2008
	Disorder prediction, virulence prediction			
	Drug resistance prediction			

suggested to study and visualize pathogen-related pathways. For example, the VirHostNet is a knowledge base for the management and analysis of proteome-wide virus-host interaction networks and a resource of manually curated interactions defined for a wide range of viral species (Navrati et al. 2009). Genomic and proteomic data is often informationally synergistic, allowing for the reconstruction of known pathways from the first principles. The combination of these forms of data have been used to identify libraries of recurring motifs, where the mixed semantics of the pattern promises to be more informative than any single data source taken in isolation in building biological networks (Michael et al. 2008; Stavrinides et al. 2008).

Systems biology has arisen from various attempts to move away from the reductionist approach, which is hindered by the difficulty of breaking a system into separable and meaningful parts. It encompasses several high-throughput analytic technologies, including genomics, transcriptomics to measure gene expression and its regulation at the level of messenger RNA and microRNA production, proteomics to measure changes in protein production, and computational biology, which depends on analytic software packages for analyzing, organizing, and interpreting those data (Sakata and Winzeler 2007). Such an approach treats pathogens and their environments as a series of hierarchical levels or networks from gene products to whole organisms and integrates the time dimension in order to structure knowledge and to determine rules that would allow navigation between levels (Lisacek et al. 2006). This approach demands new tools for data management, the integration of which offers the opportunity to correlate multiple lines of evidence and to reduce uncorrelated noise.

1.3.4 Cross-Validating the Knowledge Sources

The major difference between the pre- and post-genomics eras is that one can now potentially account for and keep track of all components at once. However, the gathering of a large collection of data does not guarantee that we can make sense of it or that new knowledge will emerge (Collado-Vides et al. 2009). The chance for enriching biomedical knowledge can be increased by mixing various streams of data and gaining robustness from the "cross-validation" of the knowledge sources (Guyet et al. 2007). Public websites like Galaxy (http://galaxy.psu.edu) and InterPro (http://www.ebi.ac.uk/interpro/) offer integration toolsets for genomics and proteomics analyses.

As generating data remains a costly undertaking, computational models have a pivotal role to play in the integrative science. They help researchers to illuminate the underlying processes and identify the key questions that need to be addressed experimentally (Restif 2009). Compared to conventional, small-scale experimental approaches, they give a wider, often more relevant view of host responses to infections or other health insults. These computational models have the capacity to guide and direct wet lab experimental efforts complimenting traditional in vivo, in situ, and in vitro testing with the emerging in silico approach (Lengauer et al. 2007;

Raman et al. 2008). Some impressive starts have been made on bacterial models in the form of simulation tools. For example, the reconstruction of metabolic networks gave birth to the first examples of in silico strains that can be utilized to explore alternative ways of identifying new drug targets (Jamshidi and Palsson 2007). The end result of these simulations may be the genomic bioengineering of microorganisms based on knowledge of interacting systems and networks of genes and gene products.

Text mining tools are being created to query the PubMed literature database and to integrate the available genomic and proteomic information to map the genes and their interrelationship with particular networks of a disease (Korbel et al. 2005; Jelier et al. 2008; Rzhetsky et al. 2008; Zaremba et al. 2009). An unsupervised, systematic approach for associating genes and phenotypic characteristics (G2P) that combines literature mining with comparative genome analysis has been successfully applied and has uncovered clusters of unsuspected G2P associations (Korbel et al. 2005).

1.4 Enabling Knowledge Communities: eScience

The phase of history in which biomedical science could be significantly advanced by individual researchers without data sharing has come to a close. The global, collaborative analyses of data and the exchange of the results across social, political and technological boundaries have created the demand for new cyber-infrastructures for research. There has been a major effort, in the form of e-Science, to develop technologies to fulfill these demands (Craddock et al. 2008).

1.4.1 Novel Infrastructures Support Knowledge Communities

The chance of making a discovery or replicating the finding is greatly increased if there are effective mechanisms for different groups to share data and thereby enlarge the number of samples that are studied. This paradigm has been successful in both human genomics and infectious disease research (e.g., including the rapid discovery and identification of emerged pathogens such as the Nipah virus and the novel coronavirus that caused the SARS epidemic). Post-genomic era solutions such as federated databases and other technologies that enhance connectivity and data retrieval have created a new knowledge environment (Birkholtz et al. 2006; Thorisson et al. 2009). The level of technical competence required of the users is being reduced by the provision of “off-the-shelf” solutions. For example, the GEN2PHEN project offers “database-in-a-box” installation packages, which include an open-source complete genetic association database system with the option for federation (Thorisson et al. 2009).

Alternative infrastructures for e-Science with significant advantages over conventional Internet technologies are offered by grid and cloud computing and the Semantic Web (Numann and Prusak 2007; Craddock et al. 2008). First, grids provide unique access to high performance computing power, distributed applications and sources (see Chap. 14 for examples). Second, grids increase data storage spaces, and allow data and tools to be shared by geographically dispersed users. However, developing and maintaining grid or cloud architectures remains a complex task and requires further advances in security and privacy models before they can be embraced by diagnostic laboratories (Lisacek et al. 2006).

1.4.2 Data Aggregation

Tasks that require an e-Science approach or global science that is performed in silico are typically computationally intensive and use heterogeneous resources that must be integrated across distributed networks (Craddock et al. 2008). Increasingly, the genomic, proteomic and metabolomic data have to be integrated with traditional literature in a machine-readable way. Typical sets of experimental data yield component lists with quantitative content data and a catalog of interactions and networks. This requires the establishment of a middleware to convert experimental data into a format suitable for manipulation and viewing by end-users. For example, the Generic Model Organism Database project (GMOD; http://gmod.org) aims to link experimental data with corresponding contextual meta-data about experimental conditions and protocols in a multi-user, multi-center environment. It offers a collection of open source tools for creating and managing genome-scale biological databases ranging from a small database of genome annotations to a large web-accessible community database. Another approach is to trade off the width of integration for more depth with regard to a particular analysis task, and to employ workflow systems such as InforSense (http://www.inforsense.com) or Taverna (http://taverna.sf.net). These act as glue layers between various data sources and analysis packages and are also often referred to as pipelines, in silico protocols or *e*-experiments (Turnbaigh et al. 2007). “Pipeline” is mostly used to describe executable workflows, while the other terms are dedicated to abstract workflows (Lisacek et al. 2006).

Many innovative solutions for the multi-dimensional integration of data produced by experimental laboratories have been introduced by Bioinformatics Resource Centers for Biodefense and Emerging/Re-Emerging Infectious Diseases through regional Biodefense Centers of Excellence (McNeal et al 2007; Greene et al. 2007). Sets of task- and domain-specific online query and display tools are being developed to allow the end-user to view data in a number of different formats and to run informative comparisons of data with existing libraries (Louie et al. 2007; Glassner et al. 2008). The most striking change in data collection and representation is expressed by the move from flat databases to atlases or collections of interconnected maps (Lisacek et al. 2006).

The uneven content and quality of data and the constant evolution of biomedical knowledge remain the main obstacles to data integration (Lisacek et al. 2006). The quality of data is affected by a number of factors including the accuracy of the mapping algorithms and reference datasets, the standardization of data formats and the level of detail of the experiment description (Stead et al. 2008). In addition, an increasing number of genomes are being released in "draft" form, before the finishing stage of a sequencing project, with high sequencing error rates (De Keersmaesher et al. 2006; Médigue and Moszer 2007). Recent developments in databases and browsers for genomics have been summarized by Schattner (2008).

There is an urgent need for data structures suitable for infectious disease space that can be applied to emerging "omics" data sets. The *Pathogen Information Markup Language* (PIML) has also recently been introduced to enhance the interoperability of microbiology datasets for pathogens with epidemic potential (He et al. 2005) by capturing the data elements that describe determinants of pathogen profiles. However, the jury is still out on the question of which data integration architectures are best suited to assembling large scale and highly diverse genomic data.

Integrating high-throughput techniques with other analytic tools brings a new understanding of infectious processes and introduces an era of personalized strategies for managing infectious diseases. In this way, informatics becomes an irreplaceable platform for the constant cross-fertilization and interplay between focused and genome-wide studies.

1.5 Translating "Omics" into Clinical Practice

1.5.1 Rapid Identification of Pathogens

Rapid and standardizable molecular identification systems have emerged during the last decade, with the development of sequence based species identification and sub-typing as the alternative to slow, labor-intensive and underpowered phenotypic techniques. Molecular identification usually relies on the detection of a single gene or multiple gene targets, or requires the comparison of whole microbial genomes. For example, in the pragmatic world of diagnostic bacteriology, conserved housekeeping genes such as the 16S rRNA gene, *rpoB* gene and others have been accepted as reliable targets. They are found in all microorganisms and show enough sequence conservation for accurate alignment as well as enough variation for phylogenetic analyses (Christen 2008). Furthermore, the 16S rRNA gene based phylogeny is sufficiently congruent with those based on whole genome approaches. Sequencing of six to eight genes or loci, as it typically done in multilocus sequence typing analysis, may constitute a reasonable compromise between single gene-based and whole genome-based methods for species diversity studies.

To streamline the process of the translation of sequencing-based identification into clinical practice, the concept of the pathogen profile has been introduced (Sintchenko et al. 2007). A pathogen profile is a single, multivariate observation or set of observations, comprised of classes of specific attributes (e.g., genome, transcriptome, proteome or metabolome data), which are designed to allow the interrogation of existing or future databases, and the integration of genomics and post-genomics data with clinical observations and patient outcomes. The profile may indicate the probability that a specific marker is associated with a clinically relevant phenotype such as in vivo antimicrobial resistance or high transmissibility. This information allows the classification of strains into "risk groups" for treatment failure or a propensity to cause outbreaks of infections. It is often important to capture the quantitative information about a pathogen, in vivo, i.e. viral or bacterial loads and their units of measurement. In contrast to traditional subtyping, which is based on phenotypic characteristics such as serotype, biotype, phage type or antimicrobial susceptibility, genetic profiling describes the phenotypic potential in the nucleic acid sequence. A pathogen profile is a synthesis of various markers and clinical end-points, which can be extracted from medical charts that characterize an individual patient's clinical and public health outcomes. The profile may be heuristic, when only a single genetic marker is associated with a specific patient outcome, while more insights can be achieved when attributes from different levels of the biological hierarchy (i.e. gene detection, gene expression, metabolite profiles etc) corroborate and complement each other. Machine learning algorithms, such as E-Predict (Urisman et al. 2005), are being developed to identify viruses and bacteria present in clinical samples. These profiles are based on the microarray hybridization patterns or DNA sequences of pathogens.

1.5.2 Guiding Antibiotic Prescribing Decisions

Many computerized evidence-based guidelines and decision support systems (DSS) have been designed to improve the effectiveness and efficiency of antibiotic prescribing (Samore et al. 2005; Buising et al. 2008). The most frequently utilized are electronic guidelines and protocols, especially for the empirical selection of antibiotics. The majority of DSS result in improvement in clinical performance and, in at least half of the published trials, in improved patient outcomes (Finch and Low 2002; Sintchenko et al. 2007; Sintchenko et al. 2008a). The revival of interest in prescribing-decision-support reflects the recent change in emphasis from support for diagnostic decisions towards support for patient management, and the changing focus from systems targeting a broad range of clinical diagnoses to task- and condition-specific decision aids. Despite reported successes of individual applications, the safety of electronic prescribing systems in routine practice has recently been identified as an issue of potential concern.

Bioinformatics assisted prescribing has become a new frontier in reducing the complexities of prescribing combinations of antimicrobials in the era of multidrug resistance. The great diversity of mutational patterns contributing to antimicrobial resistance complicates the choice of optimal therapies. A range of bioinformatics tools to predict drug resistance or response to therapy from a genotype, have been developed to support clinical decision-making (Beerenwinkel et al. 2003; Lengauer and Singh 2006). These tools use either a statistical approach, in which the inferred model and prediction are treated as regression problems, or machine learning algorithms, in which the model is addressed as a classification problem (Sintchenko et al. 2008a). A statistical learning approach to the ranking of therapeutic choices often relies on a direct correlation between the baseline microbial profiles, the therapeutic decision and the patient's response to treatment (e.g., expected reduction in viral load resulting from anti-HIV combination therapy). For example, several susceptibility scores have been used for combination antiretroviral therapy. These take into account specific resistance mutations and add up the activities of individual drugs in the regimen (Lengauer and Singh 2006). Computer-assisted therapy depends on the availability of widely shared databases that can correlate quality-controlled data from genotypic resistance assays and treatment regimens with short- and long-term clinical outcomes. Databases such ARDB (Liu and Pop 2009) capture differences in antimicrobial sensitivities and reflect variation in the amino acid composition of resistant microbes, but simply counting mutations may not be enough to predict functional differences, which affect treatment outcomes.

1.5.3 Linking Genomics to Clinical Outcomes

The molecular profiling of pathogens is based on the concept that various pathogens can be associated with different clinical outcomes. It brings together the pathogen and host factors as the pathogenesis and natural history of infection are determined by both the pathogen and human genetic susceptibility. The effectiveness of combining host and pathogen genetics in a single system or "genetics-squared" has been proven in studies of viral infections (Persson and Vance 2007). Investigations of the impact of host genetics on the susceptibility to HIV infection and the rate of disease progression have mainly used a candidate gene approach to reveal associations with a number of different genes. The genome-wide association studies look at the genetic variation across the human genome in order to uncover factors not previously suspected of influencing infection outcomes. For example, this strategy identified variants of the HIV virus associated with differences in the control of viral load at set points and in disease progression. However, unraveling the interaction between the host and microbial genetic factors requires large clinical trials, reinforcing the role of collaborative networks and data repositories.

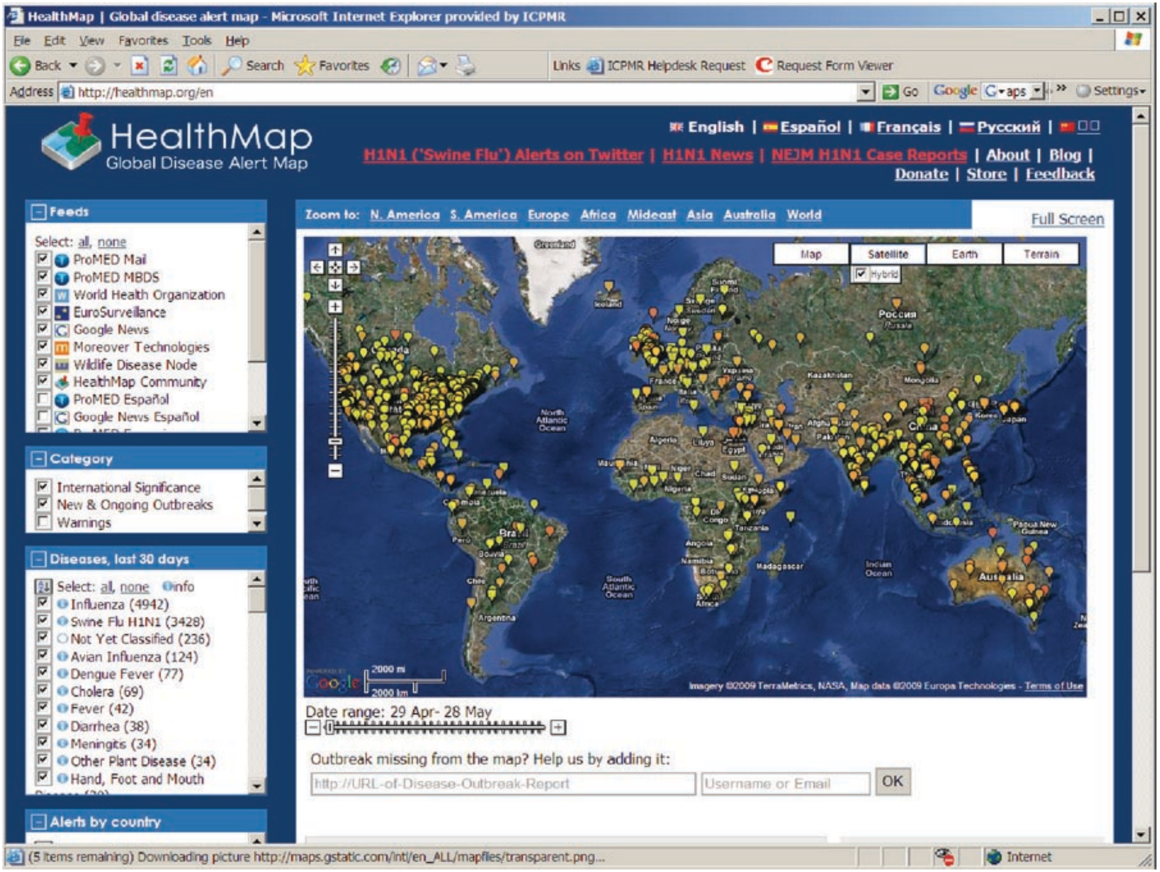
Informatics methods have become critical for data mining to decipher links between genetic variation and disease pathogenesis in order to define markers of disease progression, to guide the optimum use of therapeutics and to refine the drug and vaccine development (Mansmann 2005). A better understanding of the function of genes and other parts of the genome has enabled the reverse engineering approach, which may lead to the characterization and discovery of potential drug targets, vaccine candidates and diagnostic or prognostic markers (Davies and Flower 2007; Yang et al. 2008b). Proteins with essential biological functions present in multiple pathogens could be the best drug targets. Once the target genes essential for pathogen survival are identified, their susceptibility to specific compounds derived from large chemical libraries is examined in silico and in vitro (Muzzi et al. 2007; Biswas et al. 2008).

1.5.4 Tracing Pathogens with Epidemic Potential

Increases in the use of electronic medical records and the availability of information technology tools have created opportunities for the automation of surveillance and facilitation of surveillance based on either syndromic or disease-specific signals (Amadoz and Gonzales-Candelas 2007; M'ikanatha et al. 2007). The automation of data collection improves the time and completeness of surveillance and allows infection control professionals to focus on interventions (Hota et al. 2008; Young and Stevenson 2008).

The comparison of chromosomal sequences allows the identification of the unique genomic signatures of pathogens for the purposes of infection control and "microbial forensics." Molecular typing methodologies, in contrast to classical phenotypic methods, allow the discrimination of variations among strains within a species, the elucidation of the route of contamination, the identification of the source of infection as well as the analysis of epidemics. The identification of the natural reservoir and any possible intermediate hosts of pathogens is critical for understanding the transmission modes, designing a long-term disease control strategy, and preventing future reintroduction (Sintchenko and Gallego 2009). Bioinformatics assisted biosurveillance addresses the inefficiencies of traditional surveillance, as well as the need for a more timely and comprehensive infectious disease monitoring and control. It leverages on recent breakthroughs in the rapid, high-throughput molecular profiling of microorganisms and text mining, as well as on the growing electronic body of knowledge about the molecular epidemiology of pathogens with epidemic potential. Such a framework combines the genetic and geographic data of a pathogen to reconstruct its history and to identify the migration routes through which the strains spread regionally and internationally (Cantón 2005; Sintchenko et al. 2008b). Computer-based geographic information systems (GIS) have offered an efficient way to visualize the dynamics of the transmission of infections, especially in the setting of a community outbreak (McKee et al. 2000; Schreiber et al. 2007).

Another way to track infectious diseases of public health concern is to monitor health-seeking behavior in the form of queries to online search engines used by the general public or health professionals. Epidemics of seasonal influenza in areas with a large population of Internet users have been successfully detected using Google search data and then correlated with visits to a doctor (Ginsberg et al. 2009; Brownstein et al. 2009). The advent of news aggregators has led to the development of new disease surveillance tools that can continuously mine, categorize, filter, and visualize multilingual online information about epidemics. The Global Public Health Intelligence Network (GPHIN), developed almost a decade ago by Health Canada in collaboration with WHO, HealthMap (http://www.healthmap.org/en) (Fig. 1.2) or Geosentinel (http://www.istm.org/geosentinel/main.html) among many others are examples of such early warning systems. Resources for infection prevention and control on the World Wide Web have been recently reviewed elsewhere (Brownstein et al. 2009; Johnson et al. 2009)


HealthMap | Global disease alert map - Microsoft Internet Explorer provided by ICPMR
File Edit View Favorites Tools Help
Back Search Favorites
Links ICPMR Helpdesk Request Request Form Viewer
Address http://healthmap.org/en
Go Google G•aps Settings
HealthMap
Global Disease Alert Map
English | Español | Français | Русский
H1N1 ('Swine Flu') Alerts on Twitter | H1N1 News | NEJM H1N1 Case Reports | About | Blog |
Donate | Store | Feedback
Zoom to: N. America S. America Europe Africa Mideast Asia Australia World
Full Screen
Map Satellite Earth Terrain
Hybrid
2000 mi
2000 km
Imagery ©2009 TerraMetrics, NASA, Map data ©2009 Europa Technologies - Terms of Use
Feeds
Category
Diseases, last 30 days
Alerts by country
OK
Internet

Select: all, none
☑ ProMED Mail
☑ ProMED MBDS
☑ World Health Organization
☑ EuroSurveillance
☑ Google News
☑ Moreover Technologies
☑ Wildlife Disease Node
☑ HealthMap Community
☐ ProMED Español
☐ Google News Español

☑ International Significance
☑ New & Ongoing Outbreaks
☐ Warnings

Select: all, none info
☑ Influenza (4942)
☑ Swine Flu H1N1 (3428)
☑ Not Yet Classified (236)
☑ Avian Influenza (124)
☑ Dengue Fever (77)
☑ Cholera (69)
☑ Fever (42)
☑ Diarrhea (38)
☑ Meningitis (34)
☑ Other Plant Disease (34)
☑ Hand, Foot and Mouth

Date range: 29 Apr- 28 May

Outbreak missing from the map? Help us by adding it:

http://URL-of-Disease-Outbreak-Report | Username or Email

(5 items remaining) Downloading picture http://maps.gstatic.com/intl/en_ALL/mapfiles/transparent.png...

Fig. 1.2 Screen shot of HealthMap (Global Disease Alert Map) displaying reports about recent outbreaks in English language sources (with permission, http://www.healthmap.org/en)

1.6 Conclusions

The reductionist approach to biomedical research focusing on the study of cells and molecules has peaked with the sequencing of the human genome. However, it is becoming increasingly clear that "taking apart" analyses have reached their limit, and the time has perhaps come for integrative science (An and Faeder 2009). Developments in informatics have been critical in supporting and engaging with both reductionist and integrative paradigms. On one hand, informatics has equipped comparative genomics with tools to scrutinize genes and explore genetic polymorphisms. On the other hand, informatics has enabled the generation of integrative and testable hypotheses through the discovery of knowledge in databases and through the study of gene-phenotype connections between a pathogen and its host environment. A variety of data sets can be integrated, including the patient's demographic and clinical presentation, the laboratory results, the pathogen's gene regulation and expression, and metabolic maps with different parameters reflecting the phenotypic behavior of a pathogen and host factors. In early years some skeptics saw informatics-assisted research as a distraction of effort and funding away from traditional hypothesis-driven inquiry. Since then, infectious disease informatics has verified its status as a platform for hypothesis generation and testing (Sintchenko et al. 2007).

New breakthroughs in infectious disease informatics (IDI) are the result of cross-pollination between different disciplines that use technologies to gather and disseminate knowledge (Fig. 1.3). Microbial genome sequence analysis and metagenomics have contributed intriguing new data types and data sources to IDI. Bioinformatics has brought to the IDI a range of analytic tools, databases and data standards. Conventional health informatics and computer science has provided high performance solutions for the data storage, sharing, analysis and visualization as well as clinical terminology libraries, data standards, decision support and technology evaluation frameworks. Importantly, the infectious disease informatics community has fed the lessons learnt from the implementation of clinical and public health systems back to the broader audience.

As the subsequent chapters of this volume testify, infectious disease informatics is set to lead to the more targeted and effective prevention, diagnosis and treatment of infections through a comprehensive review of the genetic repertoire and metabolic profiles of pathogens. The post-genomic era offers new opportunities for the efficient discovery of safe and efficacious subunit vaccines by shortcutting the enormous economic burden of the experimental process. Our analytical capacity has already become the rate-limiting step in biomedical research. At the same time, it provides an opportunity to apply the engineering paradigm to biomedical research, thereby mandating the development of tools that can dynamically represent a body of current knowledge. However, the simplistic application of brute force computational power to massive reams of biomedical data is unlikely to result in meaningful mechanistic insight. It cannot be overstressed that informatics initiatives should compliment "wet laboratory" practices. An iterative loop of discovery and validation between the two methodologies remains the best way forward.

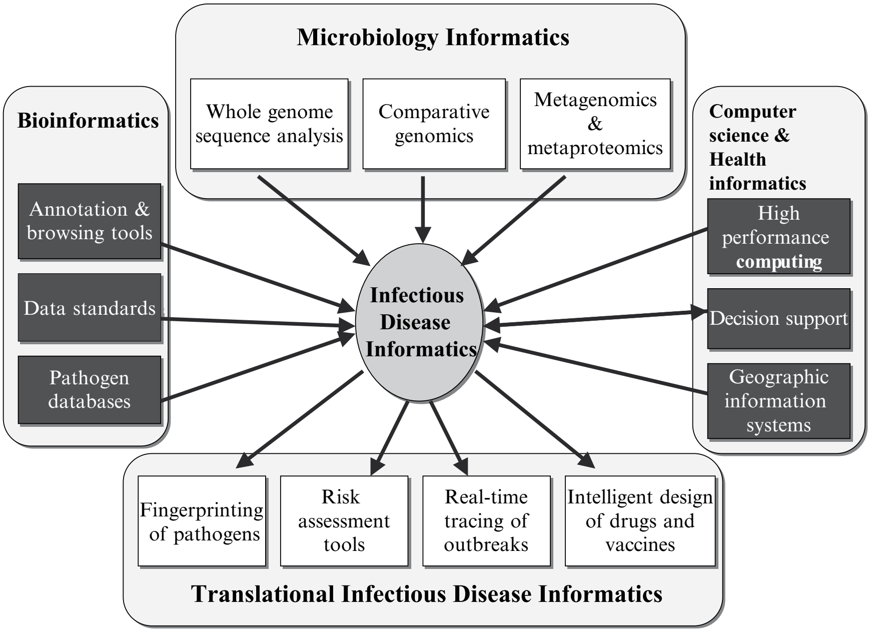

Microbiology Informatics
Whole genome sequence analysis
Comparative genomics
Metagenomics & metaproteomics
Bioinformatics
Annotation & browsing tools
Data standards
Pathogen databases
Computer science & Health informatics
High performance computing
Decision support
Geographic information systems
Infectious Disease Informatics
Fingerprinting of pathogens
Risk assessment tools
Real-time tracing of outbreaks
Intelligent design of drugs and vaccines
Translational Infectious Disease Informatics

Fig. 1.3 Inter-relations between branches of informatics and bioinformatics domains

References

Amadoz A, Gonzales-Candelas F (2007) epiPATH: an information system for the storage and management of molecular epidemiology data from infectious pathogens. BMC Infect Dis 7:32

An GC, Faeder JR (2009) Detailed qualitative dynamic knowledge representation using a BioNet Gen model of TLR-4 signaling and preconditioning. Math Biosc 217:53–63

Bansal AK (2005) Bioinformatics in microbial biotechnology – a mini review. BMC Microb Cell Factor 4:19

Beerenwinkel N et al (2003) Geno2Pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. Nucleic Acids Res 31:3850–3855

Behr MA (2008) Mycobacterium du jour: what's on tomorrow's menu? Microb Infect 10:968–972

Binnewies TT, Motro Y, Hallin PF et al (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct Integr Genomics 6:165–185

Birkholtz L-M et al (2006) Integration and mining of malaria molecular, functional and pharamacological data: how far are we from a chemogenomic knowledge space? Malaria J 5:110

Biswas S, Raoult D, Rolain J-M (2008) A bioinformatic approach to understanding antibiotic resistance in intracellular bacteria through whole genome analysis. Int J Antimicrob Agents 32:207–220

Brent MR (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet 9:62–73

Brownstein JS, Freifeld CC, Madoff LC (2009) Digital disease detection - harnessing the Web for public health surveillance. N Engl J Med 360:2153–2157

Buising KL, Thursky KA, Black JF (2008) Improving antibiotic prescribing for adults with community acquired pneumonia: does a computerised decision support system achieve more than academic detailing alone?-A time series analysis. BMC Med Inform Dec Mak 8:35

Burrack LS, Higgins DE (2007) Genomic approaches to understanding bacterial virulence. Curr Opin Microbiol 10:4–9

Cantón R (2005) Role of the microbiology laboratory in infectious disease surveillance, alert and response. Clin Microbiol Infect 11(Suppl 1):S3–S8

Carver T, Berriman M, Tivey A et al (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinform 24:2672–2676

Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18(2):324–330

Chatr-aryamontri A, Ceol A, Peluso D, Nardozza A, Panni S et al (2009) VirusMINT: a viral protein interaction database. Nucleic Acids Res 37:D669–D673

Chaudhuri RR et al (2008) xBASE2: a comprehensive resource for comparative bacterial genomics. Nucleic Acids Res 36:D543–D546

Chen L et al (2005) VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 33:D325.

Chen SL et al (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. Proc Natl Acad Sci USA 103:5977–5982

Christen R (2008) Identification of pathogens – a bioinformatic point of view. Curr Opin Bitech 19:266–273

Collado-Vides J, Salgado H, Morett E et al (2008) Bioinformatics resources for the study of gene regulation in bacteria. J Bacteriol 191:23–31

Craddock T, Harwood CR, Hallinan J, Wipat A (2008) e-Science: relieving bottlenecks in large-scale genome analyses. Nat Rev Microbiol 6:948–954

Darling ACE, Mau B, Blatter FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14(7):1394–1403

Davies MN, Flower DR (2007) Harnessing bioinformatics to discover new vaccines. Drug Discov Today 12:389–395

De Keersmaecker SCJ, Thijs IMV, Vanderleyden J, Marchal K (2006) Integration of omics data: how well does it work for bacteria? Mol Microbiol 62:1239–1250

Delcher AL, Harmon D, Kasif S et al (1999) Improved microbial gene identification with GLIMMER. Nucl Acids Res 27:4636–4641

Deloger M, El Karoui M, Petit M-A (2009) A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol 191:91–99

Dougherty TJ, Barrett JF, Pucci MJ (2002) Microbial genomics and novel antibiotic discovery: new technology to search for new drugs. Curr Pharmac Design 8:1119–1135

Driscoll T, Dyer MD, Murali TM, Sobral BW (2009) PIG - the pathogen interaction gateway. Nucleic Acids Res 37 (Database Issue):D647–D650

Field D, Wilson G, van der Gast C (2006) How do we compare hundreds of bacterial genomes? Curr Opin Microbiol 9:499–504

Finch RG, Low DE (2002) A critical assessment of published guidelines and other decision-support systems for the antibiotic treatment of community-acquired respiratory tract infections. Clin Microbiol Infect 8(Suppl 2):69–91

Forst CV (2006) Host-pathogen systems biology. Drug Discov Today 11:220–227

Frézal L, Leblois R (2008) Four years of DNA barcoding: current advances and prospects. Infect Genet Evol 8:727–736

Gallego B, Sintchenko V, Wang Q et al (2009) Biosurveillance of emerging biothreats using scalable genotype clustering. J Biomed Inform 42:66–73

Galperin MY (2005) A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. BMC Microbiol 5:35

Garrido C, Roulet V, Chueca N et al (2008) Evaluation of eight different bioinformatics tools to predict viral tropism in different human immunodeficiency virus type 1 subtypes. J Clin Microbiol 46:887–891

Ginsberg J, Mohebbi MH, Patel RS, Brammer L et al (2009) Detecting influenza epidemics using search engine query data. Nature 457:1012–1014

Glasner JD et al (2008) Enteropathogen Resource Integration Center (ERIC): bioinformatics support for research on biodefense-relevant enterobacteria. Nucleic Acids Res 36:D519–D523

Greene JM, Collins F, Lefkowitz et al (2007) National Institute of Allergy and Infectious Diseases Bioinformatics Resource Centers: new assets for pathogen informatics. Infect Immun 75:3212–3219

Guigó R, Flicek P, Abril JF et al (2007) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol 7(Suppl 1):S21–S31

Guyet T, Garbay C, Dojat M (2007) Knowledge construction from time series data using a collaborative exploration system. J Biomed Inform 40:672–687

Harrington ED, Jensen LJ, Bork P (2008) Predicting biological networks from genomic data. FEBS Lett 582:1251–1258

He Y, Vines RR, Wattam AR, Abramochkin GV et al (2005) PIML: the Pathogen Information Markup Language. Bioinform 21:116–121

Hota B, Jones RC, Schwartz DN (2008) Informatics and infectious diseases: what is the connection and efficacy of information technology tools for therapy and health care epidemiology. Am J Infect Control 36:S47–S56

Hutchinson CA (2007) DNA sequencing: bench to bedside and beyond. Nucleic Acids Res 35:6227–6237

Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the *in silico* strain *iNJ661* and proposing alternative drug targets. BMC Syst Biol 1:26

Jelier R, Schuemie MJ, Veldhoven A et al (2008) Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol 9(6):R96

Johnson LE, Reyes K, Zervos MJ (2009) Resources for infection prevention and control on the World Wide Web. Clin Infect Dis 48:1585–1595

Kahveijian A, Quackenbush J, Thompson JF (2008) What would you do if you could sequence everything? Nat Biotech 26:1125–1133

Kann MG (2008) Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform 8:333–346

Kommedal Ø, Karlsen B, Sæbø Ø (2008) Analysis of mixed sequencing chromatograms and its application in direct 16S rRNA gene sequencing of polymicrobial samples. J Clin Microbiol 46:3766–3771

Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA 102:2567–2572

Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 36:6688–6719

Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C et al (2005) Systematic association of genes to phenotypes by genome and literature mining. PloS Biology 3:e134

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299–306

Lengauer T, Sing T (2006) Bioinformatics-assisted anti-HIV therapy. Nat Rev Microbiol 4:790–797

Lengauer T, Sander O, Sierra S et al (2007) Bioinformatics prediction of HIV coreceptor usage. Nat Biotech 25:1407–1410

Lisacek F, Cohen-Boulakia S, Appel RD (2006) Proteome informatics II: bioinformatics for comparative proteomics. Proteom 6:5445–5466

Liu B, Pop M (2009) ARDB – Antibiotic Resistance Genes Database. Nucleic Acids Res 37:D443–447

Louie B et al (2007) Data integration and genomic medicine. J Biomed Inform 40:5–16

Lussier YA, Liu Y (2007) Computational approaches to phenotyping: high-througput phenomics. Proc Am Thorac Soc 4:18–25

M'ikanatha NM, Lynfield R, Van Beneden CA, de Valk H (2007) Infectious disease surveillance. Blackwell, Oxford

MacLean D, Jones JDG, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. Nat Microbiol Rev 2009 7:287–296

Majoros WH (2007) Methods for computational gene prediction. Cambridge University Press, Cambridge.

Mansmann U (2005) Genomic profiling: interplay between clinical epidemiology, bioinformatics and biostatistics. Methods Inf Med 44:454–460

McKee KT, Shields TM, Jenkins PR et al (2000) Application of a geographic information system to the tracking and control of an outbreak of shigellosis. Clin Infect Dis 31:728–733

McNeil LK et al (2007) The National Microbial Pathogen Database Resource (NMPDR): a genomic platform based on subsystem annotation. Nucleic Acids Res 35:D347–D353

Médigue C, Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. Res Microbiol 158:724–736

Meyer F et al (2003) GenDB – an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31:2187–2195

Michael H, Hogan J, Kel A et al (2008) Building a knowledge base for system pathology. Brief Bioinform 9:518–531

Muzzi A, Masignani V, Rappuoli R (2007) The pan-genome: towards a knowledge-based iscovery of novel targets for vaccines and antibacterials. Drug Discov Today 12:429–439

Navrati V, de Chassey B, Mayniel L et al (2009) VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. Nucleic Acids Res 37:D661–D668

Numann E, Prusak L (2007) Knowledge networks in the age of the Semantic Web. Brief Bioinform 8:141–149

Pallen MJ, Wren BW (2007) Bacterial pathogenomics. Nature 449:835–842

Parkhill J, Dougan G, James KD et al (2001a) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature 413:848–852

Parkhill J, Wren BW, Thomson NR et al (2001b) Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413:523–527

Persson J, Vance RE (2007) Genetics-squared: combining host and pathogen genetics in the analysis of innate immunity and bacterial virulence. Immunogenet 59:761–778

Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24:142–149

Rachman H, Kaufmann SHE (2007) Exploring functional genomics for the development of novel intervention strategies against tuberculosis. Intern J Med Microbiol 297:559–567

Raman K, Kalidas Y, Chandra N (2008) TargetTB: a target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. BMC Systems Biol 2:109

Raskin DM et al (2006) Bacterial genomics and pathogen evolution. Cell 124:703–714

Reddy TBK, Riley R, Wymore F et al (2009) TB Database: an integrated platform for tuberculosis research. Nucleic Acids Res 37:499–508

Restif O (2009) Evolutionary epidemiology 20 years on: challenges and prospects. Infect Genet Evol 9:108–123

Rzhetsky A, Seringhaus M, Gerstein M (2008) Seeking a new biology through text mining. Cell 134:9–13

Sakata T, Winzeler EA (2007) Genomics, system biology and drug development for infectious diseases. Mol BioSyst 3:841–848

Samore MH, Bateman K, Alder SC et al (2005) Clinical decision support and appropriateness of antimicrobial prescribing. J Am Med Assoc 294:2305–2314

Sanger F, Air GM, Barrell BG et al (1977) Nucleotide sequence of bacteriophage X174 DNA. Nature 265:687–695

Schattner P (2008) Genomes, browsers and databases. Cambridge University Press, Cambridge.

Schreiber MJ, Ong SH, Holland RCG et al (2007) DengueInfo: a web portal to dengue information resources. Infect Genet Evol 7:540–541

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature Biotech 26:1135–1145

Sintchenko V, Gallego B (2009) Laboratory-guided detection of disease outbreaks: three generations of surveillance systems. Arch Pathol Lab Med 133:916–925

Sintchenko V, Iredell JR, Gilbert GL (2007) Genomic profiling of pathogens for disease management and surveillance. Nat Microbiol Rev 5:464–470

Sintchenko V, Magrabi F, Tipper S (2007) Are we measuring the right thing? Variables that affect the impact of computerized decision support on patient outcomes: a systematic review. Med Inform Internet Med 32:225–240

Sintchenko V, Coiera E, Gilbert GL (2008a) Decision support systems for antibiotic prescribing. Curr Opin Infect Dis 21:573–579

Sintchenko V, Gallego B, Chung G, Coiera E (2008b) Towards bioinformatics assisted infectious disease control. BMC Bioinform 10:S10

Smarr L, Gilna P, Papadopoulos P et al (2009) Building an OptIPlante collaboratory to support microbial metagenomics. Future Gen Comp Systems 25:124–131

Squires B et al (2008) BioHealthBase: informatics support in the elucidation of influenza virus host-pathogen interactions and virulence. Nucleic Acids Res 36:D497–D503

Stavrinides J, McCann HC, Guttman DS (2008) Host-pathogen interplay and the evolution of bacterial effectors. Cell Microbiol 10:285–292

Stead DA et al (2008) Information quality in proteomics. Brief Bioinform 9:174–188

Stothard P, Wishart DS (2006) Automated bacterial genome analysis and annotation. Curr Opin Microbiol 9:505–510

Suzek BE, Ermolaeva MD, Schreiber M, Salzberg SL (2001) A probabilistic method for identifying start codons in bacterial genomes. Bioinform 17:1123–1130

Tettelin H, Masignani V, Cieslewicz MJ et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Nat Acad Sci USA 102:13950–13955

Thorisson GA, Muilu J, Brookes AJ (2009) Genotype-phenotype databases: challenges and solutions for the post-genomic era. Nat Rev Genet 10:9–18

Turnbaugh PJ et al (2007) The Human Microbiome Project. Nature 449:804–810

Urisman A, Fischer KF, Chiu CY, Kistler AL et al (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. Genome Biol 6:R78

Ussery DW, Wassenaar TM, Borini S (2009) Computing for comparative microbial genomics: bioinformatics for microbiologists. Springer-Verlag, London

Van Domselaar GH, Stothard P, Shrivastava S et al (2005) BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res 33:W455–W459

Verberkmoes NC, Russell AL, Shah M et al (2009) Shortgun metaproteomics of the human distal gut flora. ISME J 3:179–189

Whitworth DE (2008) Genomes and knowledge – a questionable relationship? Trends Microbiol 16:512–519

Winnenburg R et al (2006) PHI-base: a new database for pathogen host interactions. Nucleic Acids Res 36:D459–D464

Wu H-J, Wang A H-J, Jennings MP (2008) Discovery of virulence factors of pathogenic bacteria. Curr Opin Chem Biol 12:93–101

Xiang Z, Tian Y, He Y (2007) PHIDIAS: a pathogen-host interaction data integration and analysis system. Genome Biol 8:R150

Yang JY, Yang MQ, Arabnia HR, Deng Y (2008a) Genomics, molecular imaging, bioinformatics, and bio-nano-info integration are synergistic components of translational medicine and personalized healthcare research. BMC Genomics 9(Suppl 2):11

Yang X, Yang H, Zhou G, Zhao G-P (2008b) Infectious disease in the genomic era. Ann Rev Genom Hum Genet 9:21–48

Yan Q (2008) Bioinformatics databases and tools in virology research: an overview. In Silico Biol 8:71–85

Yao J, Lin H, Van Deynze A (2008) PrimerSNP: a web tool for whole-genome selection of allele-specific and common primers of phylogenetically-related bacterial genomic sequences. BMC Microbiol 8:185

Young J, Stevenson KB (2008) Real-time surveillance and decision support: Optimizing infection control and antimicrobial choices at the point of care. Am J Infect Control 36:S67–S74

Zaremba S, Ramos-Santacruz M, Hampton T, Shetty P et al (2009) Text-mining of PubMed abstracts by natural language processing to create a public knowledge base on molecular mechanisms of bacterial enteropathogens. BMC Bioinform 10:177

Zeng D, Chen H, Lynch C, Eidson M, Gotham I (2005) Infectious disease informatics and outbreak detection. In: Chen H, Fuller SS, Friedman C, Hersh W (eds) Medical informatics: knowledge management and data mining in biomedicine. Springer, New York

Zhou F, Olman V, Xu Y (2008) Barcodes for genomes and applications. BMC Bioinform 9:546.