CHAPTER 8.1
Viral bioinformatics

B. Adams[1], A. Carolyn McHardy[1], C. Lundegaard[2] and T. Lengauer[1]

[1]Max-Planck-Institut für Informatik, Saarbrücken, Germany
[2]Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Kongens Lynby, Denmark

1 Introduction

Pathogens have presented a major challenge to individuals and populations of living organisms, probably as long as there has been life on earth. They are a prime object of study for at least three reasons: (1) Understanding the way of pathogens affords the basis for preventing and treating the diseases they cause. (2) The interactions of pathogens with their hosts afford valuable insights into the working of the hosts' cells, in general, and of the host's immune system, in particular. (3) The co-evolution of pathogens and their hosts allows for transferring knowledge across the two interacting species and affords valuable insights into how evolution works, in general. In the past decade computational biology has started to contribute to the understanding of host-pathogen interaction in at least three ways which are summarized in the subsequent sections of this chapter.

Taking influenza as an example the computational analysis of viral evolution within the human population is discussed in Sect. 2. This evolutionary process takes place in the time frame of years to decades as the virus is continuously changing to evade the human immune system. Understanding the mechanisms of this evolutionary process is key to predicting the risk of emergence of new highly pathogenic viral variants and can aid the design of effective vaccines for variants currently in circulation.

Section 3 addresses the molecular basis of how such vaccines can be developed. Vaccines present the human immune system molecular with determinants of viral strains that elicit an immune response against the virus and activate the buildup of molecular immune memory without being pathogenic. That section also gives a succinct introduction to the workings of the human immune system.

Section 4 addresses the issue of highly dynamic viral evolution inside a single patient. Some viruses have the capability of this kind of evolution in order to evade the

Corresponding author: Thomas Lengauer, Max-Planck-Institut für Informatik, Campus E1 4, 66123 Saarbrücken, Germany (e-mail: lengauer@mpi-sb.mpg.de)

immune response of the host or the effects of a drug therapy. HIV is the example discussed here. Drug therapies against HIV become ineffective due to the virus evolving to a variant that evades the therapy. If this happens the therapy has to be replaced with another therapy that effectively targets the viral variant now present inside the patient.

2 Viral evolution in the human population

Influenza is a classic example of a pathogen that evades immunity at the population level. Due to a strong immune response in the host, which clears the virus within a few days, the virus can only survive by moving on quickly. Following an infection, hosts retain strong immunity to a particular antigenic type. As immunity accumulates in the population, there is increasing selection for pathogens with altered antigenic types that are less effectively recognized and thus have a higher probability of finding a susceptible host. By rapid evolution influenza is able to persist at relatively high prevalence in the human population. Consequently, vaccines must be frequently updated to ensure a good match with the circulating strain. However, even with current vaccination programs, endemic influenza remains a significant burden and is associated with an estimated 37,000 deaths in the U.S. alone.

In addition to the endemic activity, influenza pandemics occasionally occur when avian forms of the virus adapt to humans or provide genetic material that is incorporated into existing human forms. The antigenic novelty of these variants allows them to sweep though the global population, often causing severe disease. There were three such pandemics in the twentieth century. The most severe of them, the 'Spanish Flu' of 1918, resulted in 30 to 50 million deaths.

Thus, two key goals of influenza research are predicting viral evolution in the human population to determine optimum vaccine configurations and the early recognition of potential pandemic strains circulating in, or emerging from, the avian population. Large-scale genome sequencing and high-throughput experimental studies of influenza isolates from various sources have a central role in both of these endeavors.

2.1 Biology and genetics

Influenza viruses are single-stranded, negative sense RNA viruses of the family *Orthomyxoviridae* (Webster et al. 1992). Three phylogenetically and antigenically distinct types currently circulate, referred to as influenza A, B and C. All types infect humans and some other mammals. Influenza A also infects birds. This section will focus on influenza A, because of its high prevalence and increased virulence in humans, compared to types B and C.

The influenza A genome is composed of eight RNA segments totaling approximately 14 kb of sequence. The segments encode eleven proteins that are required for the

replication and infection cycle of the virus. The two major determinants recognized by the human immune system are the surface glycoproteins hemagglutinin (HA) and neuraminidase (NA). Hemagglutinin is responsible for binding to sugar structures on the ephithelial cells lining the respiratory tract and entry into the cell during the first stage of infection. Neuraminidase plays a part in releasing assembled viral particles from an infected cell by cleaving terminal sugar structures from neighboring glycoproteins and glycolipids on the cell surface. Several subtypes of influenza A are distinguished on the basis of the antigenic properties of the HA and NA proteins. There are 16 known subtypes for HA and 9 for NA, all of which occur in birds. In humans, subtypes H2N2 and H3N8 have circulated in the past but currently only H3N2 and H1N1 are endemic. Of these H3N2 is more virulent and evolves more rapidly.

There are two distinct mechanisms by which the influenza genome evolves. One is the acquisition of mutations, deletions or insertions during the replication process. This occurs at a higher rate than for DNA-based viruses, as RNA polymerases do not possess a proof-reading mechanism. Some of these changes subsequently become fixed in the viral sequence, either through the random fixation process of genetic drift or because they confer a selective advantage. This gradual change and its impact on the phenotype level is referred to as antigenic drift. The second mechanism of evolution is reassortment (see Fig. 1). If two different strains simultaneously infect the same host, a novel strain may arise with a combination of segments from the two. The phenotypic change associated with the emergence of such a viral variant is referred to as antigenic shift.

2.2 Vaccine strain selection for endemic influenza

The human immune system primarily targets the hemagglutinin surface protein of the influenza virus. Whether primed by infection or vaccination, antibodies provide long lasting immunity to that particular HA configuration. However, due to antigenic drift

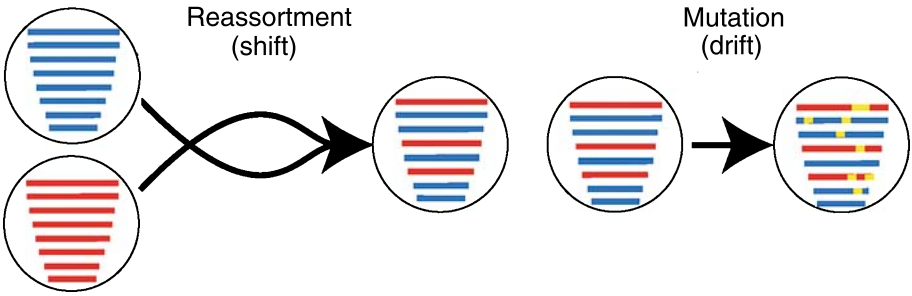

Fig. 1 Schematic representation of influenza evolution by reassortment (left) and mutation (right). Each viral genome is composed of 8 RNA segments. Reassortment of the 8 segments from two distinct viruses can result a new viable form of the virus. Drift occurs when errors during viral replication produce novel variants with small changes, i.e. insertions, deletions or mutations in the sequence segments

within just a few years those antibodies do not efficiently recognize the circulating HA. Influenza vaccines must thus be regularly updated and re-administered. The WHO makes vaccine recommendations based on the prevalence of recently circulating strains. If a new genotype, based on the HA segment, appears to be increasing in prevalence, then hemagglutination-inhibition (HI) assays using post infection ferret sera are carried out to determine whether this is associated with phenotypic change in terms of the antigenicity. If there is significant phenotypic change, the current vaccine is unlikely to be effective against the proposed emergent strain and must be updated. The genotype-phenotype map for influenza virus is unclear and genotyping is only used to choose candidate strains for HI assays. However, recent advances have indicated several ways in which genome-based methods may improve vaccine selection.

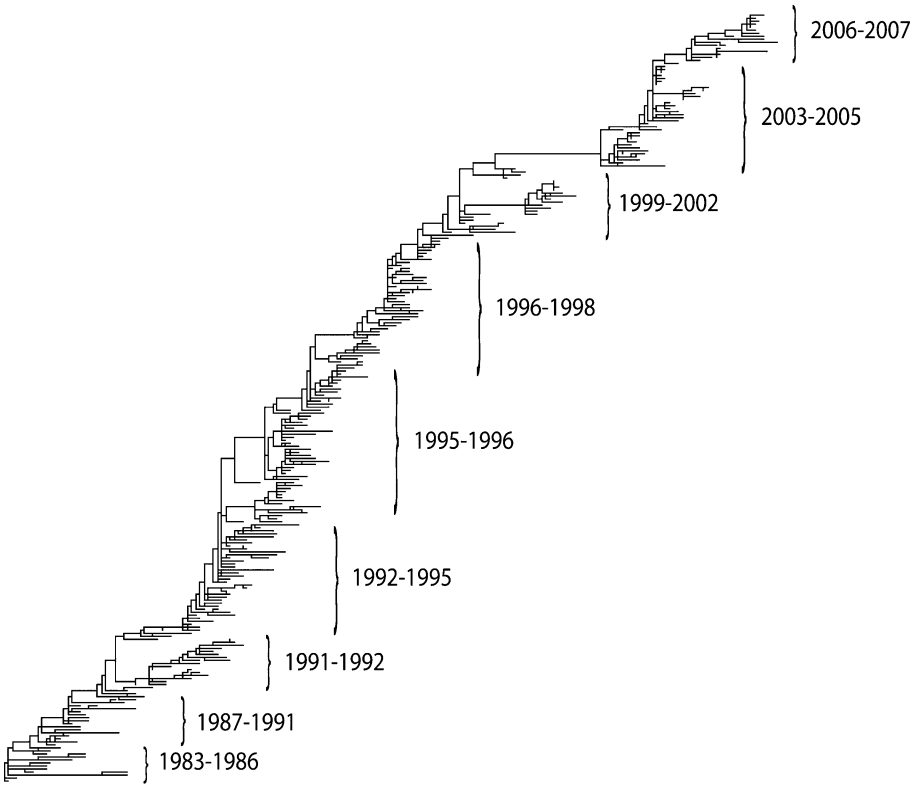

Fig. 2 Phylogenetic tree for the influenza HA coding sequences constructed by maximum parsimony using the software PAUP (http://paup.csit.fsu.edu/) from the sequences of 507 viruses isolated between 1983 and 2007. Dates to the right of the tree indicate the year that the majority of sequences contributing to that section were isolated. The tree has a distinctive cactus like shape characterized by constant turnover and limited diversity at any point in time

Bioinformatic analyses of the hemagglutinin encoding sequences have revealed characteristics of the evolutionary process and also determined relevant properties with respect to viral fitness. Phylogenetic trees of these sequences have a cactus-like topology (see Fig. 2). A diverse strain repertoire is periodically replaced by just a single strain, which constitutes the progenitor for all future lineages (Fitch et al. 1997). Population genetic theory states that such trees can be derived by random genetic drift if population size becomes very small or by selection if fitter variants emerge and periodically replace all others.

Further analyses of such trees led to the identification of a set of rapidly evolving codons in the antibody-binding and receptor-binding sites of the protein (Bush et al. 1999). These codons show a significantly higher ratio of synonymous to nonsynonymous substitutions than expected by chance, indicating that the driving force in the evolution of the HA gene is selection for variants that are fitter in terms of the evasion of host immunity acquired from previous infections. These positively selected for codons also possess predictive value with respect to the future fitness of a set of viral strains.

The relationship between the influenza genotype and phenotype has been elucidated by the application of multidimensional scaling to create a low dimensional representation of antigen-antibody distances measured with hemagglutinin inhibition assays (Smith et al. 2004). This showed that genotypes isolated over the same 2–5 year period cluster in phenotype space. Significant differences between clusters mostly localize to antibody-binding sites, the receptor-binding site and positively selected codons of the HA sequence. As more data become available, the combined analysis of genotypes and their relationship to the antigenic phenotype will enhance our capability to predict dominant circulating strains and estimate the efficacy of proposed vaccines.

2.3 Pandemic influenza

Antigenic drift allows partial immune evasion, but the host population, on average, always has some degree of immunity. Occasionally however, novel strains with no antigenic history cause global pandemics. In the twentieth century this happened in 1918, 1957 and 1968. Further pandemics are considered inevitable unless their origin can be rapidly detected or, better still, predicted (Taubenberger et al. 2007). Whole genome analysis has shown that the 1968 and 1957 pandemic strains were reassortants that introduced avian HA, PB1 and, in 1957 NA, segments into viruses already circulating in, and adapted to, the human population. The antigenic novelty of the 1918 pandemic strain also stems from its introduction from an avian source. Whether it crossed to humans directly from birds, circulated in swine first, or was a reassortment of existing avian and human strains remains a matter of debate.

Since 1997, the avian H5N1 subtype has been considered a serious candidate for a novel pandemic, due to a small but increasing number of human cases. This requires the avian HA protein to undergo adaptation to bind to human receptors. Analysis of the

viral genotypes responsible for the human H5N1 cases has identified several common amino acids changes in and around the binding region. It has also shown that the virus is repeatedly crossing directly from birds, without reassortment or sustained human to human transmission (2005). So far, an H5N1 strain with pandemic potential has not emerged, but continual surveillance is vital. Early detection of the accumulation of mutations that may facilitate a host switch, the mixing of genetic material from human and avian forms or evidence of human to human transmission will be critical for containment strategies

The efficiency of such surveillance measures may also be improved by targeting particular geographic regions. Based on a phylogenetic tree of avian H5N1 sequences, a phylogeography of significant migratory trajectories has been constructed for Eurasia by minimizing the number of migration events necessary to keep the phylogeny geographically consistent (Wallace et al. 2007). These data indicated that Indochina is a largely isolated subsystem in terms of H5N1 evolution and Guangdong in China is the main source of diversity and diffusion throughout Eurasia. It may therefore be practical to invest more of the surveillance effort into this region.

2.4 Conclusion

Even with modern medicine the burden of annual influenza is significant and the threat of a pandemic constantly hangs over the world. Vaccines, chemo prophylactics, detection and containment strategies are all in use. But the influenza virus, like malaria and HIV, is a constantly moving target and optimizing pharmaceutical design and public health policy is a complex problem requiring an integrated knowledge of, among other things, epidemiology, immunology and molecular biology. Bioinformatics has provided, and will continue to provide, vital insights in all of these areas.

3 Interaction between the virus and the human immune system

3.1 Introduction to the human immune system

The human immune system rests basically on two pillars. One pillar is solely genetically determined and remains unchanged throughout the life of an individual. This so-called innate immune system basically provides physical protection barriers and registers if generally recognizable foreign substances are entering the organism. If such substances are detected a fast and general protection mechanism sets in whose nature is determined by the type of substance registered. The innate protection mechanisms also include an activation of the other pillar of the immune system, the adaptive immune system. This part evolves during the life, and its present state is highly

dependent of the infection history of the individual. The adaptive immune system is itself basically split up in two parts. First the humoral immunity, which happens outside cells within the body liquids and is antibody-driven. Special immunoglobulin molecules (antibodies) mediate the humoral response. Antibodies are produced by B lymphocytes that bind to antigens by their immunoglobulin receptors, which is a membrane bound form of the antibodies. When the B lymphocytes become activated, they start to secrete the soluble form of the receptor in large amounts. Antibodies are Y-shaped, and each of the two branches functions independently and can be recombinantly produced and is then known as fragments of antibodies (Fab). The antibody can coat the surface of an antigen such as a virus and generally this will inactivate whatever undesired function the respective object may have, and facilitate the uptake of the antibody-bound object *via* phagocytosis by macrophages, which will then digest the object. Macrophages, B cells, and dendritic cells are all so-called professional antigen presenting cells (APC). They carry a special receptor named the major histocompatibility complex (MHC) class II. This receptor is able to present peptides derived from degraded phagocytosed proteins. Other cells (T cells) carries a receptor, the T cell receptor (TCR), which, if the T cell also carries a so called CD4+ receptor, is able to bind to MHC class II molecules presenting a foreign peptide, e.g. one not originating from the human proteome. Such an interaction will stimulate B cells to divide and further progress to produce more antibodies as well as survive for a long time as memory B cells. The presence of memory B cells enables the immune system to react faster in a subsequent infection by the same pathogen. The CD4+ T cells actually also belong to the second part of the adaptive immune system, which is the cellular immune system. Another important feature of cellular immunity regards T cells with the CD8 coreceptor (CD8+ T cells). The TCR of CD8+ T cells can recognize foreign peptides in complex with membrane bound MHC class I molecules on the outer side of nucleated cells. Such an interaction will activate the T cells to signal and induce cell death of the cell presenting the foreign peptide.

Both antibodies and TCRs are composed of a light and a heavy chain. These chains are translated from genes resulting from a genetic recombination of two and three genes, respectively, during the B-cell development in the bone marrow. These genes exist in several nonidentical duplicates on the chromosome and can be combined into a large number of different rearrangements. However, the molecular processes linking the genes are imprecise and involve generation of P (palindromic) nucleotides, addition of N (non-templated) nucleotides by terminal deoxynucleotidyl transferase (TdT) and trimming of the gene ends and therefore also play a major role in the generation of the huge diversity needed to be able to respond to any given pathogen. The T cells having a mature TCR are being validated in the thymus. The host will eliminate T cells having a TCR that is either unable to bind to an MHC:peptide complex or that will recognize an MHC with a peptide originating from the hosts proteome (self peptides). All the above is highly simplified text book immunology (Janeway 2005).

3.2 Epitopes

To be able to combat an infection the immune system must first recognize the intruder as foreign. The specific parts of the pathogen that is recognized and induces an immune response are called epitopes. Epitopes are often parts of larger macromolecules, which most often happen to be polypeptides and proteins. B-cell epitopes are normally classified into two groups: continuous and discontinuous epitopes. A continuous epitope, (also called a sequential or linear epitope) is a short peptide fragment in an antigen that is recognized by antibodies specific for the given antigen. A discontinuous epitope is composed of residues that are not adjacent in the amino acid sequence, but are brought into proximity by the folding of the polypeptide.

The cellular arm of the immune system consists as desribed of two parts; the CD8+ cytotoxic T lymphocytes (CTLs), and the CD4+ helper T lymphocytes (HTLs). CTLs destroy cells that present non-self peptides (epitopes). HTLs are needed for B cells activation and proliferation to produce antibodies against a given antigen. CTLs on the other hand perform surveillance of the host cells, and recognize and kill infected cells. Both CTLs and HTLs are raised against peptides that are presented to the immune cells by major histocompatibility complex (MHC) molecules, which are encoded in the most polymorphic mammalian genes. The human versions of MHCs are referred to as the human leucocyte antigens (HLA). The cells of an individual are constantly screened for presented peptides by the cellular arm of the immune system. In the MHC class I pathway, class I MHCs presents endogenous peptides to T cells carrying the CD8 receptor (CD8+ T cells). To be presented, a precursor peptide is normally first generated by cutting endogenous produced proteins inside the proteasome, a cytosolic protease complex. Generally, resulting peptides should bind to the TAP complex for translocation into the endoplasmic reticulum (ER). During or after the transport into the ER the peptide must be able to bind to the MHC class I molecule to invoke folding of the MHC before the complex can be transported to the cell surface. When the peptide:MHC complex is presented on the surface of the cell, it might bind to a CD8+ T cell with a fitting TCR. If such a TCR clone exists a CTL response will be induced and the peptide is considered an epitope. The most selective step in this pathway is binding of a peptide to the MHC class I molecule. As mentioned above, the MHC is the most polymorphic gene system known. The huge variety of protein variants brought forth by this polymorphism is a big challenge for T-cell epitope discoveries, enhancing the need for bioinformatical analysis and resources. It also highly complicates immunological bioinformatics, as predictive methods for peptide MHC binding have to deal with the diverse genetic background of different populations and individuals. On a population basis, hundreds of alleles (gene variants) have been found for most of the HLA encoding loci (1839 in release 2.17.0 of the IMGT/HLA Database, http://www.ebi.ac.uk/imgt/hla/). In a given individual either one or two different alleles are expressed per locus depending on whether the same (in homozygous individuals) or two different (in heterozygous individuals) alleles are present on the two

different chromosomes. Each MHC allele binds a very restricted set of peptides and the polymorphism affects the peptide binding specificity of the MHC; one MHC will recognize one part of the peptide space, whereas another MHC will recognize a different part of this space. The very large number of different MHC alleles makes reliable identification of potential epitope candidates an immense task if all alleles are to be included in the search. Many MHC alleles, however, share a large fraction of their peptide-binding repertoire and it is often possible to find promiscuous peptides, which bind to a number of different HLA alleles. The problem can thus be largely reduced by grouping all the different alleles into supertypes in a manner were all the alleles within a given supertype have roughly the same peptide specificity. This grouping generally requires some knowledge regarding the binding repertoire of either the specific allele or an allele with a very similar amino acid sequence.

The peptides recognized by the CD4+ T cells are called helper epitopes. These are presented by the MHC class II molecule, and peptide presentation on this MHC follow a different path than the MHC class I presentation pathway: MHC class II molecules associate with a nonpolymorphic polypeptide referred to as the invariant chain (Ii) in the ER. The Ii chain is a type II membrane protein, and unlike MHC molecules the C-terminal part of the molecule extends into the lumen of the ER. The MHC:Ii complex accumulates in endosomal compartments and here, Ii is degraded, while another MHC-like molecule, called HLA–DM in humans, loads the MHC class II molecules with the best available ligands originating from endocytosed antigens. The peptide: MHC class II complexes are subsequently transported to the cell surface for presentation to the CD4+ T helper cells. The helper T cells will bind the complex and be activated if they have an appropriate TCR.

3.3 Prediction of epitopes

A major task in vaccine design is to select and design proteins containing epitopes able to induce an efficient immune response. The selection can be aided by epitope prediction in whole genomes, relevant proteins, or regions of proteins. In addition, prediction of epitopes may help to identify the individual epitopes in proteins that have been analyzed and proven to be antigens using experimental techniques based on, e.g., Western blotting, immunohistochemistry, radioimmunoassay (RIA), or enzyme-linked immunosorbent assays (ELISA).

Today, the state-of-the-art class I T-cell epitope prediction methods are of a quality that makes these highly useful as an initial filtering technique in epitope discovery. Studies have demonstrated that it is possible to rapidly identify and verify MHC binders from upcoming possible threats with high reliability, and take such predictions a step further and validate the immunogenicity of peptides with limited efforts, as has been shown with the influenza A virus (see next subsection). It is also possible to identify the vast majority of the relevant epitopes in rather complex organisms using class I MHC

binding predictions and only have to test a very minor fraction of the possible peptides in the virus proteome itself. MHC class II predictions can be made fairly reliably for certain alleles. B-cell epitopes are still the most complicated task. However, some consistency between predicted and verified epitopes is starting to emerge using the newest prediction methods (Lundegaard et al. 2007).

B-cell epitope prediction is a highly challenging field due to the fact that the vast majority of antibodies raised against a specific protein interact with parts of the antigen that are discontinuous in the polypeptide sequence. The prediction of continuous, or linear, epitopes, however, is a somewhat simpler problem, and may be still useful for synthetic vaccines or as diagnostic tools. Moreover, the determination of continuous epitopes can be integrated into determination of discontinuous epitopes, as these often contain linear stretches. More successful methods combine scores from the Parker hydrophilicity scale and a position specific scoring matrix (PSSM) trained on linear epitopes. Different experimental techniques can be used to define conformational epitopes. Probably the most accurate and easily defined is using the solved structures of antibody–antigen complexes. Unfortunately, the amount of this kind of data is still scarce, compared to linear epitopes. Furthermore, for very few antigens all possible epitopes have been identified. The simplest way to predict the possible epitopes in a protein of known 3D structure is to use the knowledge of surface accessibility and newer methods using protein structure and surface exposure for prediction of B-cell epitopes have been developed. The CEP method calculates the relative accessible surface area (RSA) for each residue in the structure. The RSA is defined as the fraction of solvent exposed surface of a given amino acid in the native structure relative to the exposed surface the same amino acid placed centrally in a tri-peptide, usually flanked by glycines or alanines. It is then determined which areas of the protein are exposed enough to be antigenic determinants. Regions that are distant in the primary sequence, but close in three-dimensional space will be considered as a single epitope. DiscoTope (www.cbs.dtu.dk/services/DiscoTope) uses a combination of amino acid statistics, spatial information and surface exposure. The system is trained on a compiled dataset of discontinuous epitopes from 76 X-ray structures of antibody–antigen protein complexes. (Haste Andersen et al. 2006). B-cell epitope mapping can be performed experimentally by other methods than structure determination, e.g., by phage display. The low sequence similarity between the mimotope (i.e. a macromolecule, often a peptide, which mimics the structure of an epitope) identified through phage display and the antigen complicates the mapping back onto the native structure of the antigen, however, a number of methods have been developed that facilitate this.

A number of methods for predicting the binding of peptides to MHC molecules have been developed. The majority of peptides binding to MHC class I molecules have a length of 8–10 amino acids. Position 2 and the C-terminal position have turned out generally to be very important for the binding to most class I MHCs and these positions are referred to as anchor positions (Fig. 3). For some alleles, the binding motifs

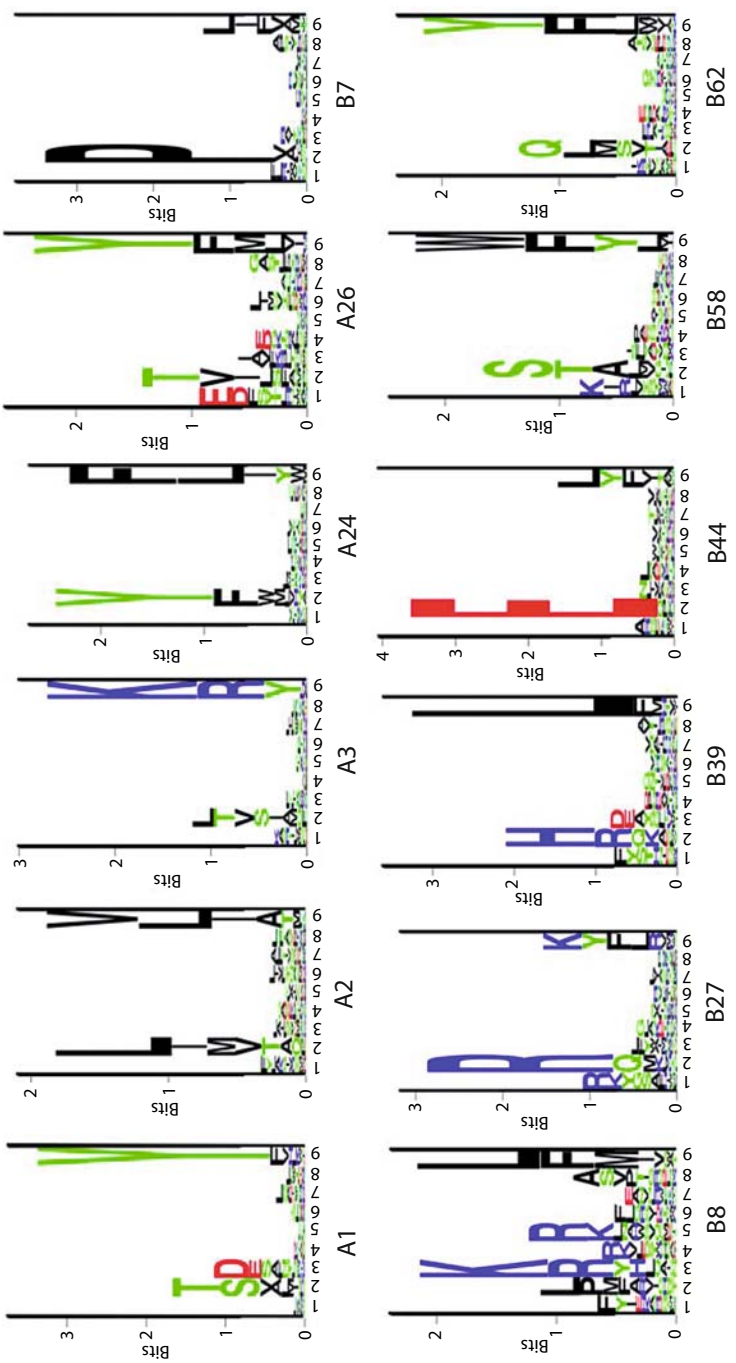

Fig. 3 Sequence Logos of raw count statistics of peptides measured to bind with a IC_{50} stronger than 500 nM. Peptides were mainly selected from the ImmuneEpitope database

have additional anchor positions. E.g., eptides binding to the human HLA-A*0101 allele have positions 2, 3 and 9 as anchors (Rammensee et al. 1999) (Fig. 1A). The discovery of such allele-specific motifs led to the development of the first reasonably accurate algorithms. In these prediction tools, it is assumed that the amino acids at each position along the peptide sequence contribute a given binding energy, which can be added up to yield the overall binding energy of the peptide. Several of these matrix methods are trained on exclusively positive examples like peptides eluted from MHCs on living cells, peptides that have been shown to induce significant interferon gamma responses in CTL assays, or peptides that bind the MHC more strongly than a certain binding affinity value (usually below 500 nM). Other matrix methods, like the SMM method, aim at predicting an actual affinity and thus use exclusively affinity data. However, matrix-based methods cannot take correlated effects into account (when the binding affinity of peptide with a given amino acid at one position depends on amino acids that are present at other positions in the peptide). Higher-order methods like ANNs and SVMs are ideally suited for taking such correlations into account and can be trained with data either in the format of binder/non-binder classification, or with real affinity data. Some of the recent methods combine the two types of data and prediction methods. The different types of predictors are reviewed in (Lundegaard et al. 2007) and an extensive benchmark of the performances of the different algorithms have been published by (Peters et al. 2006).

Representing a supertype by a well-studied allele risks the confinement to selecting epitopes that are restricted to this allele, exclusing other alleles within the supertype. Thus another, and potentially more rational approach, would be to select a limited set of peptides restricted to as many alleles as possible. This should be within reach with new methods that directly predict epitopes that can bind to different alleles (Brusic et al. 2002), or pan-specific approaches that can make predictions for all alleles, even those whose sequences are not yet known (Heckerman et al. 2007; Nielsen et al. 2007a). Finally, even though MHC binding is the most limiting step in the class I pathway the cleavage and transporting events are not insignificant. Several tools have been developed that integrate predictions of the different steps, and this has been shown to improve the predictions of actual CTL epitopes (Larsen et al. 2007).

Unlike the MHC class I molecules, the binding cleft of MHC class II molecules is open at both ends, which allows for the bound peptide to have significant overhangs in both ends. As a result MHC class II binding peptides have a broader length distribution even though the part of the binding peptide that interacts with the MHC molecule (the binding core) still includes only 9 amino acid residues. This complicate binding predictions as the identification of the correct alignment of the binding core is a crucial part of identifying the MHC class II binding motif. The MHC class II binding motifs have relatively weak and often degenerate sequence signals. While some alleles like HLA-DRB1*0405 show a strong preference for certain amino acids at the anchor positions, other alleles like HLA-DRB1*0401 allow basically all amino acids at all

positions. In addition, there are other issues affecting the predictive performance of most MHC class II binding prediction methods. The majority of these methods take as a fundamental assumption that the peptide:MHC binding affinity is determined solely by the nine amino acids in binding core motif. This is clearly a large oversimplification since it is known that peptide flanking residues (PFR) on both sides of the binding core may contribute to the binding affinity and stability. Some methods for MHC class II binding have attempted to include PFRs indirectly, in terms of the peptide length, in the prediction of binding affinities. It has been demonstrated that these PFRs indeed improve the prediction accuracy (Nielsen et al. 2007b).

3.4 Epitope prediction in viral pathogens in a vaccine perspective

As described in Sect. 2 some of the important B cell antigens vary significantly between different influenza A viral strains. Current influenza vaccines are based on inactivated influenza virus and thus mimic only the B cell response obtained by a fully infection competent strain. This has the drawback that only closely related strains will be covered by this response and new vaccines have to be produced annually as a result of the antigenic drift (see Sect. 2). Thus the ideal influenza vaccine will raise an immune response against parts of the pathogen that are conserved between as many strains as possible. To identify these parts the described prediction tools will be an invaluable help. Initial *in silico* scans of the viral genome for potential immunogenic parts will reduce the potential epitope space, and thus make experimental validations feasible. In a published example all genomic sequenced strains of H1N1 were scanned for CTL epitopes. Only 9-mer peptides in the influenza proteome that were at least 70% conserved in all strains were considered. The top 15 predicted epitopes for each of the 12 supertypes were subsequently selected to be synthesized for further validation. Because of the limited size of the influenza genome and the high variability of some of the proteins the conservation criteria resulted in relatively low prediction scores of some of the chosen peptides. 180 peptides were selected and 167 were synthesized and further validated for MHC binding and CTL response. The fraction of validated MHC binding peptides (with a binding affinity of below 500 nM) was relatively low (about 50%) compared to some other studies (60–75%) (Sundar et al. 2007; Sylvester-Hvid et al. 2004), but 13 of the 89 binding peptides, or 15%, gave a positive output in a CTL recall assay. Obviously, the conserved epitopes were found in the less variable proteins, but the large majority of the validated epitopes (85%) turned out to be 100% conserved not only in H1N1 strains but also in the H5N1 avian strains that in the last few years have infected humans resulting in severe symptoms and high mortality (Wang et al. 2007). Such epitopes can be highly valuable starting points for vaccine development. Even though cellular immunity does not protect against infections it might protect against a fatal outcome of an infection with a new aggressive strain.

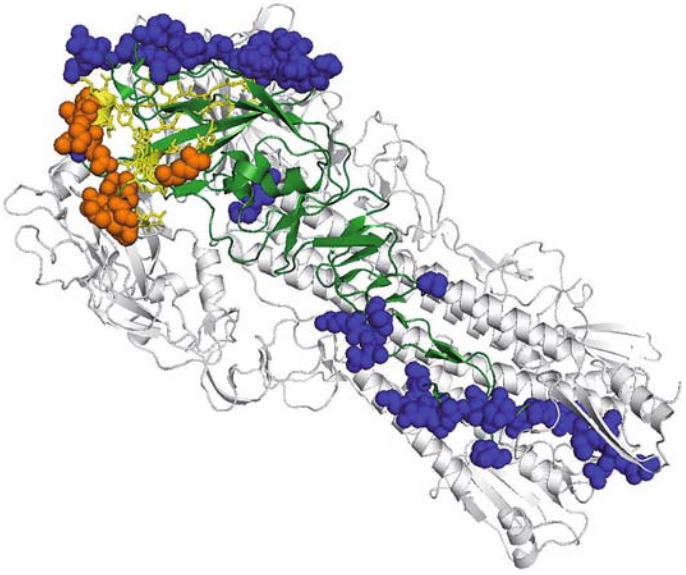

Fig. 4 3D structure of hemagglutinin with highlighted epitope predictions using chain A from the pdb entry 2IBX. White cartoon: Other chains in multimer not used for predictions. Green cartoon: Part of chain where no class II or B cell epitopes are predicted. Yellow sticks: Predicted helper epitopes (NetMHCII predictions) considering the DRB*0101 allele. Blue spheres: Predicted B cell epitopes (DiscoTope). Orange spheres: Residues predicted to be in both B cell and helper epitopes. The tools FeatureMap3D (www.cbs.dtu.dk/services/FeatureMap3D/) and PyMol (pymol.sourceforge.net/) were used to generate the drawing

To find conserved B cell and helper epitopes a similar approach could be used even though conserved conformational epitopes might be hard to find and even harder to direct a response to. Figure 2 displays a three-dimensional protein structure model of the variable surface protein hemagglutinin from a H5N1 strain. Predicted B cell epitopes are mapped on the structure, as well as helper epitopes restricted to the relatively common HLA-DRB*0101 allele.

4 Viral evolution in the human host

4.1 Introduction

The previous section has discussed the evolution which a pathogen population undergoes within the human population over a time span of years or longer. Some pathogens but not all, by any means, play a more dynamic evolutionary game inside the host by which they try to evade the host's immune system or the drug therapy that is applied to combat the disease. We observe this kind of process both with unicellular pathogens and

viruses. An example of the former is *Plasmodium falciparum* which causes Malaria. Here the pathogen evolves new suites of surface epitopes repeatedly to evade the adaptive immune response of the host, and the immune system of the host responds to the new populations of modified pathogens with recurrent fever bouts that manifest the periodic amplifications of the immune system activity.

This section will present a viral example, namely the case of Human Immunodeficiency Virus (HIV), which causes AIDS.

4.2 Replication cycle of HIV

HIV is a single-stranded RNA virus with two copies of the genome per virus particle. The replication cycle of the virus is schematically illustrated in Fig. 5.

The virus enters the human cell by attaching with its surface protein gp120 to the cellular receptor CD4. It needs one of the two cellular coreceptors CCR5 or CXCR4 to facilitate cell entry. After fusion with the cell membrane it releases its content and uses one of the viral enzymes, namely the *Reverse Transcriptase* (RT) to transcribe its genome back to DNA. Another viral enzyme, the *Integrase* (IN) splices the DNA version of the viral genome, the so-called *provirus*, into the genome of the infected host cell. This cell is often a T-helper cell of the host's immune system. Once this cell starts dividing, i.e., as part of the immune response to the HIV infection, the cell starts producing the building blocks of the virus. New virus particles assemble at the cell surface and segregate. During a final virus maturation phase, a third viral enzyme, the *Protease* (PR) cleaves the viral polyproteins into their active constituents. The dynamic evolution of HIV is manifested by the fact that RT lacks a proof-reading mechanism and introduces genomic variants during the copying process. The high turnover of over a billion virus particles per host and day during periods of high-activity immune response affords a sufficient genomic diversity for a selective evolutionary process that lends an advantage to forms of the virus that are resistant to the immune system and drug therapy with which they are confronted.

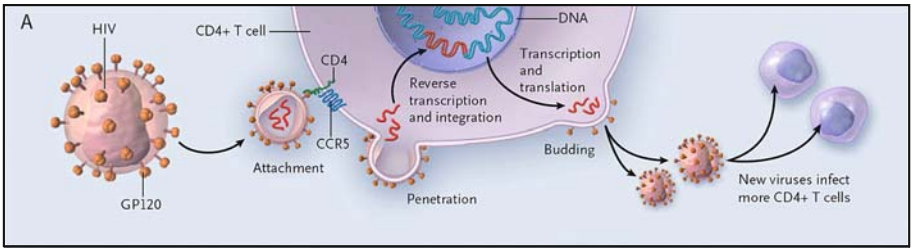

Fig. 5 Replication cycle of HIV (from (Markel 2005))

4.3 Targets for antiviral drug therapy

Antiviral drugs target one of several of the proteins involved in the viral replication cycle. The historically earliest drugs target RT and block it by providing "fake" nucleotides for the DNA assembly that act as terminators for the chain elongation process. These drugs are called *nucleoside analog RT inhibitors* (NRTIs). Another class of drugs targeting RT, the *non-nucleoside analogs* (NNRTIs), facilitate inhibition of the enzyme by binding to a specific part of its binding pocket. Since the mid 90s, inhibitors of PR (PIs) that substitute for the peptides to be cleaved by the enzyme have entered the market place. Inhibitors of integrase are just about to come to market. Finally, in recent years, several drugs have been developed that target the blockage of the process of viral cell entry, by blocking one of the involved proteins, either the viral surface protein gp41, or one of the cellular proteins, CD4, CCR5 or CXCR4. Within the older classes of drugs there are up to about a dozen different compounds in each class. The justification for so many compounds is that there are many different variants of HIV that have different resistance profiles. This is also the reason why, for over ten years, the so-called *highly antiretroviral therapy* (HAART) approach administers several drugs from several drug classes to the patient simultaneously, in order to present a high barrier for the virus on its evolutionary path to resistance. Still, after a time of several weeks up to about a year or two, the virus succeeds in evolving a variant that is resistant against the given therapy regimen. At this point, a new drug combination has to be selected to combat the new viral variant.

4.4 Manual selection of antiretroviral combination drug therapies

Even before the use of computers, doctors have selected drug therapies based on the genome of the viral variant prevalent inside the patient which, in developed countries, is routinely determined from virus in the patient's blood serum via sequencing methods. The basis for the selection is a set of *mutation tables*. There is one such table for each molecular target. The table lists, for each drug, the observed and acknowledged set of mutations (on the protein level) that have been observed to confer resistance against that drug. The offered tables are updated regularly by international societies such as the International AIDS Society (Johnson et al. 2007).

There are two problems with the mutation tables. (1) They regard different mutations as independent from each other. Any one of the mutations listed in the table is considered to confer resistance on its own. However, in some cases, mutations at different positions have been observed to interact in complex ways. For instance, a mutation can resensitize a virus to a drug to which an earlier mutation has rendered it resistant. (2) Mutations are selected to enter the mutation table by a consensus process among experts that cannot claim to be objective and

reproducible. Problem 1 has been countered by the introduction of rule-based expert systems that can implement complex resistance rules involving several mutations (Schmidt et al. 2002). Problem 2 has been approached by introducing bioinformatics methods for predicting resistance from the viral genotype. Such methods derive statistical models directly from clinical data that comprise experience on viral resistance development. We now survey the methods by which such statistical models are derived and applied.

4.5 Data sets for learning viral resistance

First we need data sets for deriving the statistical models. The availability of data in sufficient volume and quality is a major hurdle for bioinformatical approaches to resistance analysis. Data have been collected in several parts of the world, e.g., in the USA (Stanford HIV Database (Rhee et al. 2003)), over Germany (Arevir Database (Roomp et al. 2006)) and, more recently, over Europe (Euresist Database[1]). These databases contain two types of data.

1. *Genotypic data* list viral variants sampled from patients together with clinical information about the patient, including their viral load (the amount of free virus) and counts of immune cells in the blood serum. This allows for correlating the viral genotype with the virologic and immunological status of the patient.
2. *Phenotypic data* report results from laboratory experiments, in which virus containing the resistance mutations observed in the patient is subjected to different concentrations of single antiretroviral drugs and the replication fitness of the virus is measured. This results in a quantitative measure of viral resistance, the so-called *resistance factor*. Briefly, a virus with a resistance factor of 10 against some drug requires ten times the concentration of that drug in comparison to the wild-type virus in order to reduce the replication fitness of both viruses to the same extent.

In developed countries, genotypic data are collected routinely in clinical practice. Thus they are available in high volume (tens of thousands of data points). The viral genotypes are usually restricted to the genes of the target molecules (here RT and PR). Phenotypic data require high-effort laboratory procedures and cannot be collected routinely. Thus they are available in lower quantities (thousands of data points). While phenotypic data represent viral resistance in an artificial environment, they provide a highly informative quantitative value for resistance. Thus can are of substantial value for learning statistical models with high predictive power.

[1] http://www.euresist.org

4.6 Computational procedures for predicting resistance

We will survey approaches to solving three problems in resistance prediction:

1. *Quantifying the information that a mutation carries with respect to the resistance of any viral variant with that mutation against a given drug.* Any method solving this problem can be used to generate mutation tables such as the one derived by hand through expert panels.
2. *Predicting the resistance of a given genotypic variant against a given drug.* Any method solving this problem can also take complex interactions between different mutations into account and thus competes with the rule-based expert systems mentioned before.
3. *Assessing the effectiveness of a combination drug regimen against a given genotypic variant.* Methods solving this problem can take the future viral evolution into account. Thus, in effect, they can attempt to answer the question how effective the virus will be in evading the present combination drug therapy. Thus they go further methodically than any competing method.

We will now summarize the methods that are used to solve the above problems.

Several methods are available for solving Problem 1. Computing the mutual information content of a viral mutation with respect to the wild type is one alternative (Beerenwinkel et al. 2001). Another is to generate a support-vector machine model for predicting resistance against the drug and deriving the desired information from it (Sing et al. 2005). The resulting methods yield suggestions for new resistance mutations that are highly desired by the medical community.

Problem 2 can be solved with classical supervised learning techniques such as decision trees or support vector machines (Beerenwinkel et al. 2002). These methods provide classification of viral variants into *resistant* or *susceptible*, or regression of the measured resistance factor or the viral load observed in a clinical setting. The models incur error rates of about 10–15% against measured phenotypic data and the resulting web-based prediction servers[2] are very popular with practicing physicians and laboratories evaluating patient data. Figure 6 shows an excerpt of a respective patient report that presents an intuitive display of the level of the virus against each drug.

The solution of Problem 3 is somewhat more complicated. We need several ingredients for a respective method. First, we need a notion of success and failure, respectively, of a combination drug therapy that incurs more than a moment's observation of the patient. One way is to assess the effectiveness of a therapy after

[2] E.g. http://www.geno2pheno.org

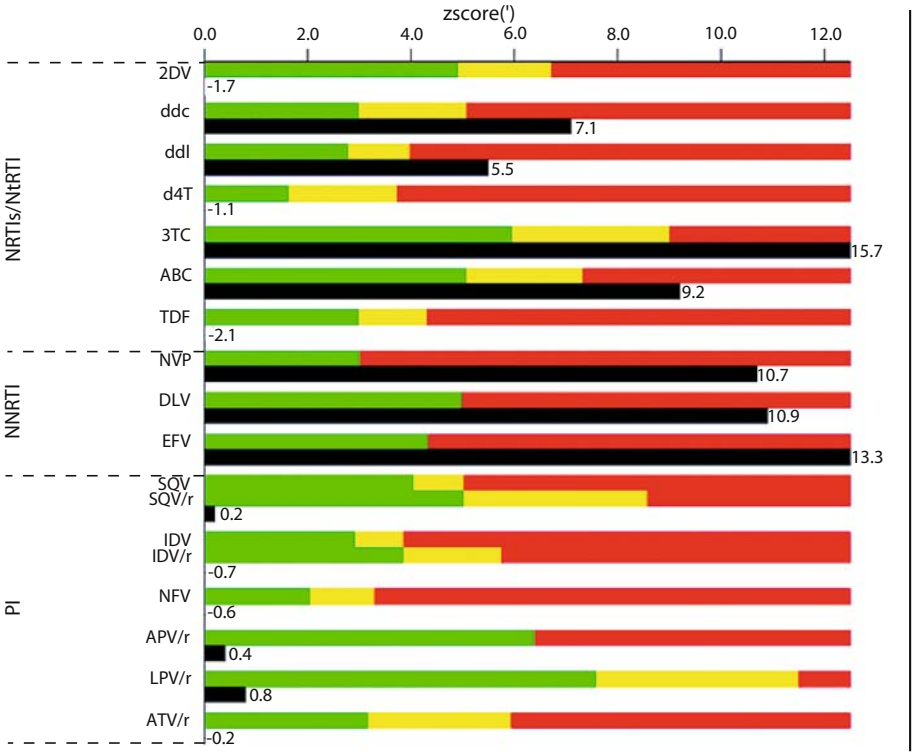

Fig. 6 Patient report by the geno2pheno resistance prediction server. There is one line for each drug. The level of resistance of the virus against the drug is represented by the length of the black bar. The colored bar above indicates the region of resistance (green – susceptible, yellow – intermediate, red- resistant)

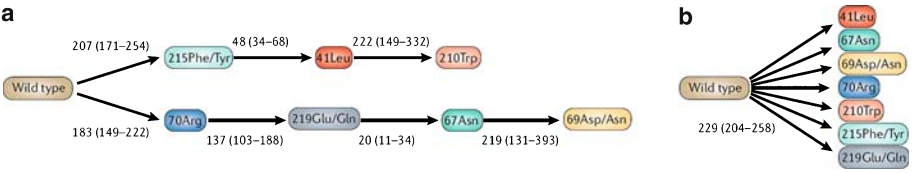

Fig. 7 Tree model of the evolutionary development of viral resistance against the NRTI zidovudine. The model consists of two trees. Tree (**a**) displays two clinically observed nontrivial paths to resistance, indicated by mutations that accumulate from the wild type from left to right. Tree (**b**) represents unstructured noise in the data. The method also return quantitative estimates for how much of the data is explained by what tree. In this case the left tree explains about 78% of the data

some time since onset, say eight weeks. The second is a model of viral evolution under drug therapy. We have developed a statistical model that represents the paths of the virus to resistance by a set of trees ((Beerenwinkel et al. 2005b), see Fig. 7)

Given such a tree model, we can derive a quantitative value for the probability of a virus to become resistant against a certain drug after a given amount of time, given a specific combination drug therapy. This value is called *the genetic barrier to drug resistance* (Beerenwinkel et al. 2005a). Finally we use multivariate statistical learning

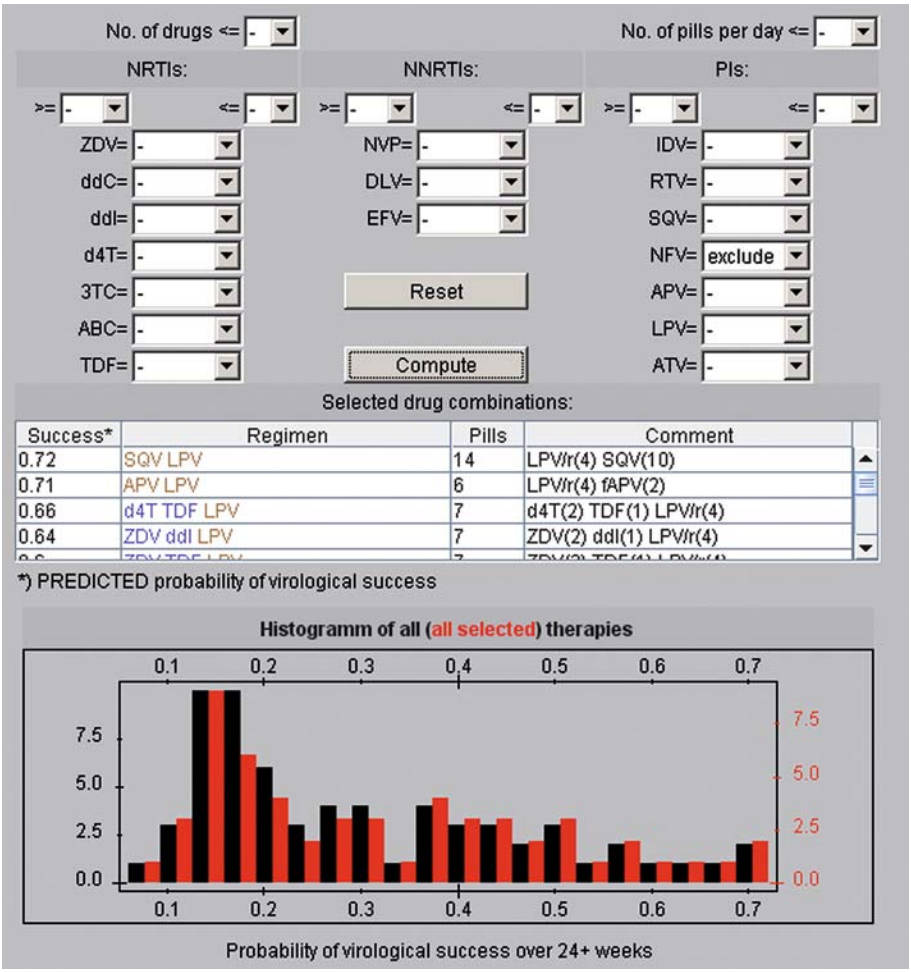

Fig. 8 Results of prediction of therapy effectiveness on the same sample as used for generating Fig. 2. At the top, the user can preselect, here, by excluding the use of the protease inhibitor NFV. In the middle a list of ranked therapies is given. The two top-ranking therapies involve two protease inhibitors, which is not surprising since, by inspection of Fig. 2 the viral variant displays few resistances against protease inhibitors. The distribution of therapy effectiveness with (red) and without (black) preselection is displayed at the bottom

techniques to generate models that classify the therapies into successes and failures based on input comprising the viral variant, the applied therapy, phenotypic resistance prediction (Problem 2), and the predicted genetic barrier to resistance. The results of the implementation of this method called THEO are displayed as illustrated in Fig. 8.

The resulting method reduces the error of therapy classification from about 24% (without any use of software) to under 15% (Altmann et al. 2007). While this is a substantial improvement in accuracy, doctors are still hesitant to use the method in clinical practice, for two reasons: (1) They would like more information on why the method arrives at its results, i.e., the ask for the results to be more interpretable. (2) They question the "objectiveness" of the data. In some sense the subjectivity of the expert decision is replaced with the arbitrariness of how the dataset is collected, from which the model are built.

Addressing both issues is possible but requires additional research which is currently under way.

4.7 Clinical impact of bioinformatical resistance testing

The methods described here are applied within clinical practice in the context of research projects and clinical studies. They improve the rate of selection of adequate drug combination therapies significantly. Besides the statistical evaluation by cross validation, there has been a retrospective study, in which previously applied therapies have been rechecked with the geno2pheno software (Problem 2 above) among other prediction systems, and the software has proven to pick successful therapies statistically significantly more often than therapies that turned out not to be successful. Among the single cases that can be reported is a patient who had been receiving HAART for 16 years within several therapy changes but without ever having virus cleared from his blood serum. After the mutation tables offered no more option for therapy, the bioinformatics software made a suggestion that was amended by the doctor. The resulting therapy was the first to clear the patient's blood of virus and held for at least 2.5 years. Thus, while the software does not make flawless suggestions it advances the state of therapy selection significantly.

Bioinformatics solutions to Problem 3 have yet to win acceptance with the practicing physicians.

The methods described here can be transferred to other diseases for which viral evolution to resistance inside the patient can be observed and for which the relevant genotypic and phenotypic data are available. Transferring the methods to Hepatitis B and C is in preparation.

A recent review on bioinformatical resistance testing is provided in (Lengauer and Sing 2006).

4.8 Bioinformatical support for applying coreceptor inhibitors

As the new coreceptor inhibitors are entering the marketplace and affording a completely new approach to AIDS therapy, there are also new problems that have to be dealt with and that can be supported with bioinformatics methods. We mentioned above that CCR5 and CXCR4 are the two coreceptors that are used alternatively by HIV to enter the infected cell. The clinical picture manifests that, almost exclusively, CCR5 is required for primary infection (R5 virus). As the disease progresses, the virus often switches to using CXCR4 (X4 virus). Some viral variants can use both coreceptors (R5X4 virus). The use of CXCR4 is often associated with enhanced disease symptoms and accelerated disease progression. Thus, preventing the virus from evolving to an X4 variant is a therapy goal. CCR5 seems to be inessential, as humans with an ineffective CCR5 gene shows no disease phenotype, but are highly resistant to developing AIDS. Thus CCR5 is an attractive target for inhibiting drugs. The first CCR5 blocker Maraviroc (Pfizer) has just entered the marketplace. Regulatory agencies, as they were admitting the drug for clinical use, prescribed accompanying tests of the virus for coreceptor usage, as it is ineffective to treat X4 viruses with CCR5 blockers.

For testing of coreceptor usage we have a similar picture as for resistance testing. Coreceptor usage is determined based on the viral genotype. There are laboratory assays for measuring coreceptor usage. They are a little bit closer to clinical routine than phenotypic resistance tests, but they still suffer from limited accessibility, long times (weeks) to receive the results and high cost.

Using genotypic and phenotypic data, one can develop statistical models for viral coreceptor usage based on the viral genotype. Supervised learning models such as support vector machines or position-specific scoring matrices are used for this purpose. The methods are based mainly on the viral genotype (this time restricted to the hypervariable V3 loop of the viral gp120 gene that binds to the coreceptor). Prediction accuracy can be enhanced by including clinical parameters, such as patient immune status, in the model or by specifically offering 3D-structural information on the V3 loop in the form of a structural descriptor that is based on mapping the viral variant under investigation onto the x-ray model of a reference V3 loop. Reviews on bioinformatical prediction of coreceptor usage can be found in (Jensen and van 't Wout 2003; Lengauer et al. 2007).

5 Perspectives

In the last decade, computational biology has embarked on the analysis of host-pathogen interactions. However, the field is still in an early stage. The analysis of viral evolution inside the human population is currently targeting genetic drift but does not yet have a handle on analyzing and predicting genetic shift. The analysis of interactions between viral epitopes and molecules of the human immune system has brought forth effective methods for analyzing and predicting the strength of MHC-binding but has yet

to develop models that adequately represent the many stages of molecular interactions and molecular transport that lead to eliciting an immune response. And the support of the selection of new antiviral therapies in the face of emerging resistant strains inside a patient is still mainly based on statistical analysis of previously applied therapies (to many different patients) rather than on a mechanistic understanding of the molecular interaction networks manifesting the disease. In all fields we would greatly benefit from dynamic simulatable models of the molecular processes manifesting the disease and of the way in which molecular determinants of the virus, the immune system of the host and the applied drugs influence them. Basic research in the field of computational modeling of virus-host interactions will be directed towards generating this network-based understanding of the involved processes. Towards this end we need not only develop new computational models but also generate the relevant experimental data for calibrating the models and for identifying the molecular determinants involved.

References

The World Health Organization Global Influenza Program Surveillance Network (2005) Evolution of H5N1 avian influenza viruses in Asia. Emerg Infect Dis 11: 1515–1521

Altmann A, Beerenwinkel N, Sing T, Savenkov I, Däumer M, Kaiser R, Rhee SY, Fessel WJ, Shafer RW, Lengauer T (2007) Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. Antivir Ther 12: 169–178

Beerenwinkel N, Lengauer T, Selbig J, Schmidt B, Walter H, Korn K, Kaiser R, Hoffmann D (2001) Geno2pheno: interpreting genotypic HIV drug resistance tests. IEEE Intel Syst 16: 35–41

Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, Korn K, Selbig J (2002) Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proc Natl Acad Sci USA 99: 8271–8276

Beerenwinkel N, Däumer M, Sing T, Rahnenführer J, Lengauer T, Selbig J, Hoffmann D, Kaiser R (2005a) Estimating HIV Evolutionary Pathways and the Genetic Barrier to Drug Resistance. J Infect Dis 191: 1953–1960

Beerenwinkel N, Rahnenführer J, Däumer M, Hoffmann D, Kaiser R, Selbig J, Lengauer T (2005b) Learning multiple evolutionary pathways from cross-sectional data. J Comput Biol 12: 584–598

Brusic V, Petrovsky N, Zhang G, Bajic VB (2002) Prediction of promiscuous peptides that bind HLA class I molecules. Immunol Cell Biol 80: 280–285

Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. Science 286: 1921–1925

Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 94: 7712–7718

Haste Andersen P, Nielsen M, Lund O (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci 15: 2558–2567

Heckerman D, Kadie C, Listgarten J (2007) Leveraging information across HLA alleles/supertypes improves epitope prediction. J Comput Biol 14: 736–746

Janeway C (2005) Immunobiology: the immune system in health and disease. Garland Science, New York

Jensen MA, van 't Wout AB (2003) Predicting HIV-1 coreceptor usage with sequence analysis. AIDS Rev 5: 104–112

Johnson VA, Brun-Vezinet F, Clotet B, Gunthard HF, Kuritzkes DR, Pillay D, Schapiro JM, Richman DD (2007) Update of the Drug Resistance Mutations in HIV-1: 2007. Top HIV Med 15: 119–125

Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M (2007) Large-Scale Validation of Methods for Cytotoxic T-Lymphocyte Epitope Prediction. BMC Bioinformatics 8: 424

Lengauer T, Sander O, Sierra S, Thielen A, Kaiser R (2007) Bioinformatics prediction of HIV coreceptor usage. Nat Biotechnol 25: 1407–1410

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

Lundegaard C, Lund O, Kesmir C, Brunak S, Nielsen M (2007) Modeling the adaptive immune system: predictions and simulations. Bioinformatics 23: 3265–3275

Markel H (2005) The search for effective HIV vaccines. N Engl J Med 353: 753–757

Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Røder G, Peters B, Sette A, Lund O, Buus S (2007a) Quantitative, pan-specific predictions of peptide binding to HLA-A and -B locus molecules. PLoS-One 2: e796

Nielsen M, Lundegaard C, Lund O (2007b) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics 8: 238

Peters B, Bui HH, Frankild S, Nielson M, Lundegaard C, Kostem E, Basch D, Lamberth K, Harndahl M, Fleri W, Wilson SS, Sidney J, Lund O, Buus S, Sette A (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput Biol 2: e65

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50: 213–219

Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 31: 298–303

Roomp K, Beerenwinkel N, Sing T, Schülter E, Büch J, Sierra-Aragon S, Däumer M, Hoffmann D, Kaiser R, Lengauer T, Selbig J (2006) Arevir: A secure platfrom for designing personalized antiretroviral therapies against HIV. In: Leser U, Naumann F, Eckman B (eds) Third International Workshop on Data Integration in the Life Sciences (DILS 2006). Springer Verlag, Hinxton, U.K. 4075, pp 185–194

Schmidt B, Walter H, Zeitler N, Korn K (2002) Genotypic drug resistance interpretation systems – the cutting edge of antiretroviral therapy. AIDS Rev 4: 148–156

Sing T, Svicher V, Beerenwinkel N, Ceccherini-Silberstein F, Däumer M, Kaiser R, Walter H, Korn K, Hoffmann D, Oette M, Rockstroh J, Fätkenheuer G, Perno C-F, Lengauer T (2005) Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling and SVM-Based feature ranking. In: Alipio MJ, Torgo L, Bradzil PB, Camacho R, Gama J (eds) Knowledge discovery in databases: PKDD 2005. Lecture notes in computer science No. 3721, Springer Verlag, Berlin/ Heidelberg, pp 285–296

Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305: 371–376

Sundar K, Boesen A, Coico R (2007) Computational prediction and identification of HLA-A2.1-specific Ebola virus CTL epitopes. Virology 360: 257–263

Sylvester-Hvid C, Nielsen M, Lamberth K, Roder G, Justesen S, Lundegaard C, Worning P, Thomadsen H, Lund O, Brunak S, Buus S (2004) SARS CTL vaccine candidates; HLA supertype-, genome-wide scanning and biochemical validation. Tissue Antigens 63: 395–400

Taubenberger JK, Morens DM, Fauci AS (2007) The next influenza pandemic: can it be predicted? JAMA 297: 2025–2027

Wallace RG, Hodac H, Lathrop RH, Fitch WM (2007) A statistical phylogeography of influenza A H5N1. Proc Natl Acad Sci USA 104: 4473–4478

Wang M, Lamberth K, Harndahl M, Roder G, Stryhn A, Larsen MV, Nielsen M, Lundegaard C, Tang ST, Dziegiel MH, Rosenkvist J, Pedersen AE, Buus S, Claesson MH, Lund O (2007) CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. Vaccine 25: 2823–2831

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y (1992) Evolution and ecology of influenza A viruses. Microbiol Rev 56: 152–179