AGGREGATION OF BIOLOGICAL KNOWLEDGE FOR IMMUNOLOGICAL AND VIROLOGICAL APPLICATIONS

OLIVO MIOTTO
B.Sc. (Hons), ARCS, Imperial College London

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2008

For Francesca and Alessandro

Haec ornamenta mea

Acknowledgments

My deepest thanks go to my supervisors, Assoc. Prof. Tan Tin Wee of the Department of Biochemistry and Dr. Vladimir Brusic of Dana-Farber Cancer Institute, Harvard Medical School, who both supported me every step of the way through this life-changing adventure, educating me as a researcher while respecting me as a peer. In particular, this work would have never started without Vladimir's vision and determination, and I will forever treasure his generosity and friendship. It has also been a privilege to collaborate with Prof. J. Thomas August of Johns Hopkins Medical School, whose enthusiasm and experience will continue to inspire my work for a long time to come. Very heartfelt thanks go to Asif M. Khan, and A. T. Heiny, fellow research students at the Department of Biochemistry, for working with me on many projects, and for helping me improve my techniques and tools. I am also grateful to Lim Swee Cheang, Director of the Institute of Systems Science, for recognizing and valuing my research efforts, and to Howard Russon, the best manager I could have ever hoped for, for his unflagging support throughout my candidature.

For making my childhood dream come true, I thank my father Mario, who taught me to love knowledge, and my mother Toni, who listened to a 12-year old's explanation of Darwin's theories as only a mother can. I thank Francesca and Alessandro, who have constantly energized me with their affection, even when they have had to compete with my studies for attention. And most of all I thank Tracy, whose love has made everything possible. Thank you for turning every day of my life into a splendid adventure.

Facendomi nascere, essi mi hanno fatto discendere dal Cielo alla Terra.
Lei con il suo amore, lei mi ha riportato dalla Terra al Cielo.

Roberto Benigni

Table of Contents

Summary

Advances in biotechnology have produced an unprecedented growth in the volume and diversity of biological data. To answer complex research questions, bioinformatics analysis needs to aggregate increasing quantities of information from expanding number of diverse sources, combining multiple tasks into analysis pipelines. Even as bioinformatics becomes integrated with the daily work of biomedical researchers, the lack of advanced computing skills restricts their access to complex computational analysis. In this thesis, we identified major issues for the scalability of bioinformatics: system and information heterogeneities must be overcome when aggregating knowledge from diverse sources; intuitive user interfaces are needed for life scientists to control analysis processes; and knowledge representation standards are needed to support knowledge flow between analysis tools. To model complex bioinformatics processes, we have developed a model of biological knowledge mining, which facilitates integration of new data with existing knowledge. We present a novel knowledge aggregation approach based on user-defined structural rules, which provides researchers with an intuitive user interface mechanism for overcoming information heterogeneities.

The knowledge aggregation method was implemented in the ABK software tool, and applied to multiple knowledge aggregation tasks. The AVANA software tool was developed to support information-theoretical diversity analysis of multiple sequence alignments, supporting peptide entropy and mutual information methods. A large-scale influenza A protein sequence dataset complete with descriptive metadata (including host, strain, geographic and temporal information), was constructed from over 90,000 public database records using the ABK platform. Using this dataset, we conducted a large-scale comparison of human-transmissible influenza strains against avian strains, using a novel method based on mutual information. The resulting catalogue of 70 adaptive amino acid mutations, distributed over eight influenza proteins, is the most comprehensive to date and reveals complex patterns of adaptations to humans. Genomic adaptation signatures, derived from this catalogue, were

used to assess the pandemic potential of H5N1 and other avian influenza strains. The ABK and AVANA tools were applied, in a collaborative research, to a systematic whole-genome analysis of vaccine targets. Conserved peptides with high HLA-binding potential were identified from large datasets of viral sequences. The conservation analysis method introduced peptide entropy, a novel measure of antigenic variability, followed by the use of HLA binding prediction algorithms to select candidate peptides. This method has been applied to multiple pathogens. In the final application, our knowledge mining approach was extended to the analysis of biomedical text for the curation of an allergen database. We devised a text mining approach based on active learning, which can be user-controlled via a simple annotation interface. Since no domain knowledge needs to be built into this text mining tool, it can be reused on a variety of curation tasks. The feasibility and utility of this approach were demonstrated by extending the ABK platform with text analysis tools. The diverse applications presented in this thesis demonstrate that our new knowledge aggregation approach is both practical and versatile, and represents an important contribution to bioinformatics and to the fields of biomedical research in which it is applied.

List of Tables

List of Figures

1. INTRODUCTION

Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

T. S. Eliot

The primary goal of biomedical research is to relieve humanity of the burden of disease, to improve the quality and length of human life. Much of this research is focused on furthering our understanding of cellular and molecular mechanisms, essential to the design of targeted therapies and preventions. In recent years, advances in biotechnology have dramatically enhanced our ability to observe these mechanisms: high-throughput methods, such as genome-scale sequencing and microarray analysis, produce massive quantities of data about a multitude of genes and their products (Hall 2007; Bernstein and Kellis 2005). We are now able to observe and measure the behaviour and characteristics of multiple molecular components, and repeat these observations under the same or different conditions. Constantly decreasing operational costs allow genomic studies to process samples from populations of individuals, under a variety of environmental and medical conditions, to pinpoint multiple contributing disease factors (Wellcome Trust Case Control Consortium 2007). The shift from targeted studies of single molecules to multidimensional studies of thousands of interacting molecules provides opportunities to build more complete biological models, and perhaps preempting disease by mapping detailed identification of risk factors (Zerhouni 2005). However, as the volume and variety of biological data increases, the gap between the wealth of knowledge that can be extracted from this data and our ability to extract it is growing. Although current computational methods are able of discerning underlying patterns in large, multi-dimensional sets of data, most computational biology methods tend to support relatively small-scale analyses, often focused on a single task. Researchers are increasingly left to "swim a sea of data" (Roos 2001) without being able to extract the maximum value possible. New computational approaches are needed, to allow biomedical researchers to perform large-scale analyses and make full use of the large-scale datasets to answer ever more complex

biological questions.

This thesis addresses the question of scaling up the computational discovery process to make the best use of the growing quantities of biological data that are becoming available. We have developed a new conceptual approach in bioinformatics, *biological knowledge mining*, which enhances the analysis of *biological data* with contextual information that provides conceptual models that describe the data. Rather than relying on the analysis of raw data entities, knowledge-enabled analysis discovers significant patterns using biologically meaningful models, and therefore allows new types of analysis to be performed. The task of *knowledge aggregation* is the essential component of this approach, in which knowledge from diverse databases is extracted, combined and encoded so it can be used by analysis tools. Our approach enables biomedical researchers to aggregate large-scale datasets comprising thousands of records, through user-friendly interfaces that do not require specialized programming. The knowledge-enabled analysis tools implemented in this thesis are capable of analyzing these large datasets, using the accompanying knowledge to organize the analysis data. This approach and methods have been applied to three real-life biomedical research problems. The results of these applications demonstrate the utility of our approach, and its applicability to a variety of diverse analysis tasks.

The work presented in this thesis is a systematic and multi-disciplinary effort, containing theoretical and practical components from multiple fields: *bioinformatics*, the discipline that focuses on the computational analysis of biological data; *knowledge management*, a branch of computing that deals with the acquisition, representation and analysis of knowledge; and *software engineering*, the branch of computing concerned with the construction of software tools. The applications presented are relevant to the biomedical fields of *immunology* and *virology*. These applications were chosen as proofs of concept, and the applications of the approach presented here are not limited to these fields.

1.1 Background

Technological advances in the fields of *biotechnology* and *information technology* (IT) have

revolutionized life sciences research. Biotechnology advances have produced cheaper, more flexible and faster methods for obtaining a variety of molecular and cellular-level observations, such as genomic or proteomic sequencing, gene and protein expression levels, cell sorting, functional profiling, and others. The Human Genome Project took 13 years to sequence the fist human genome, at a cost of US$400 million by its completion in 2003. Only four years later, James Watson's genome was sequenced in less than two months, for under US$1M (Patrick 2007). Inexpensive large-scale DNA sequencing has opened up new research opportunities for the study of large-scale genomic datasets (Mardis 2008), which are already producing a significant impact to the study of pathogens (Medini *et al*. 2008).

Bioinformatics applies computational techniques to the investigation of biological hypotheses, discerning patterns in the most relevant biological data (Brazma *et al.* 1998). Computational analyses (often referred to as *in silico assays*) provide data analysis and statistical support for hypotheses, and are widely used for computational prescreening, to narrow the scope of subsequent *in vitro* or *in vivo* experimentation. *In silico* results lead to rational and focused experimental design, which is becoming essential to the ability to investigate complex biomedical problems (Yu *et al.* 2004). Early successful applications of bioinformatics methods were limited to the study of molecular entities, such as genes, proteins and genomes, to answer biological questions at the molecular level. However, as our understanding of cellular mechanisms increases, it becomes clearer that the complexity and diversity of life has a combinatorial nature, deriving from interactions of many thousands of components (Zuckerkandl 2006). Phenotypic traits are rarely governed by single genes; rather, genes work in teams, regulating, enhancing and disrupting each other's function. Proteins produced by genes must therefore be viewed as components of complex and diverse systems, where they interact with other molecules, physically as part of assemblies, chemically as participants in cellular processes, and functionally as effectors or regulators of biological processes. To apply bioinformatics from a more systemic perspective, there is an increasing need to combine the analysis of data from multiple experiments with knowledge accumulated from the other kinds of analysis (Kanehisa and Bork 2003).

Although the high volume of diverse data generated by high-throughput biotechnology appears an ideal starting point for system bioinformatics, there is mounting concern that currently available bioinformatics tools and approaches are not sufficient to effectively tackle large-scale systemic research questions. In part, this problem is caused by the fact that most tools are limited to the analysis of entities, thus extracting knowledge solely from raw data. The analysis of systems and populations, on the other hand, requires the integration of *models* that describe the data. For example, current sequence phylogenetic tools can discover significant clades (clusters) of evolutionarily related organisms by analyzing sequence data alone. However, we also need tools that can use the clade model produced by phylogenetic analysis to identify systematic molecular differences between clades, co-evolving mutations, patterns of evolution, *etc*. In the absence of tools that are capable of "understanding" models, researchers have to organize datasets and analysis results manually, which is unsuitable for tasks that span thousands of records.

An additional obstacle to large-scale analysis is the difficulty of constructing large datasets. The 2008 Molecular Biology Database Collection lists over 1,000 publicly-accessible biological databases (Galperin 2008), which differ widely in size, purpose, level of detail and data structures, and use a diversity of standards for encoding and accessing their data. For biological researchers, the all-important task of extracting and combining information from these "information silos" is a daunting prospect, both technically and logistically (Philippi and Köhler 2006), as will be discussed in detail in Chapter 2 of this thesis. This problem is compounded by a relatively low level of computational skills among biomedical researchers, who often resort to manual methods to aggregate their datasets. As datasets grow to thousands, tens of thousands, or even greater numbers of sequences, manual curation becomes prohibitively expensive, highly error-prone and difficult. At this stage it represents a grater obstacle to large-scale analysis than the availability of suitable computational resources. Paradoxically, just as bioinformatics analysis tasks become increasingly important for the daily work of biologists, the level of technological skill required to perform these tasks increasingly exceeds the scope of a biologist's training.

1.2 Aims of this thesis

The key motivation for this thesis is to lay the foundation of what we term "second-generation" bioinformatics – analysis of biological data that is (a) *knowledge-enabled*, in that it makes use of *descriptive models* as well as raw data, and integrates analysis results with existing knowledge; (b) arbitrarily *scalable*, in that it can process increasing numbers of records, exceeding hundreds or thousands, aggregated from multiple and disparate sources; and (c) *biologist-friendly*, in that it directly empowers life science researchers to perform complex analyses with tools that do not require complex programming or significant IT infrastructure.

To realize this vision, we developed a novel conceptual framework, which models bioinformatics studies as complex pipelines of analysis tasks, with knowledge flowing from one task to the next. In this *biological knowledge mining* framework, *knowledge* comprises raw data, descriptive information about the data (metadata), and the results of analysis tasks. Knowledge thus encompasses data and models that describe the data; these models are augmented as more knowledge is aggregated, for example, as a result of the execution of an analysis task (Michalski 2003).

In most biological knowledge mining processes, the core task is the initial *knowledge aggregation* task, which deals with the construction of large-scale analysis datasets from diverse data sources. This task currently presents insurmountable problems for biologists, because data is fragmented across multiple databases, and presents *heterogeneities*, which are discussed in Chapter 2. In this thesis, we have developed a knowledge aggregation approach that overcomes common heterogeneities, empowering biologists (who have limited IT skills) to perform complex analyses. We implemented this approach as a desktop tool, the Aggregator of Biological Knowledge (ABK), which is able to process datasets consisting of tens of thousands of records on a current standard-configuration desktop computer system. ABK allows users to interact with data to be aggregated, specifying the data to be extracted by example through a simple point-and-click interface, and controlling the format of the

extracted data through user-defined vocabularies. From these user selections, ABK learns *structural rules* that are automatically applied to large sets of data records, seamlessly extracting the desired knowledge in the form required by the user. The level of automation provided by ABK, combined with the flexibility in specifying the knowledge to be extracted, dramatically reduces the effort needed for curation, allowing biologists to construct in a relatively short time large-scale datasets that would have been prohibitively laborious to assemble manually.

In practical implementations, *biological knowledge mining* requires *knowledge-enabled analysis tools*, which are able to utilize raw data as well as its accompanying knowledge for analysis. In addition, a *knowledge representation* standard must be defined, capable of describing accurately and expressively varied types of knowledge, but sufficiently generic to allow diverse tools to transfer knowledge along the analysis pipeline. In this thesis, we have developed the Antigenic Variability Analyzer (AVANA), a knowledge-enabled tool that performs information-theoretical analysis of variability in viral protein sequence alignments. AVANA integrates sequence data with descriptive metadata (such as sequence protein name, subtype, year and place of isolation, *etc*.), allowing the alignments to be partitioned into biologically meaningful subsets, subsequently used in comparative studies and meta-analyses. Since AVANA accepts arbitrary metadata fields, the user can model the population represented by the sequence alignment in the way that is most appropriate to the analysis task. As a result, biologists can rapidly test hypotheses and models by performing simple metadata queries, without having to reconstruct datasets, which is often necessary with current available tools. We identified *semantic technologies* as a suitable candidate platform for knowledge transfer in bioinformatics pipelines, and demonstrated that these technologies can use *reasoning* to perform useful data aggregation tasks, which will further reduce the need for programming. In this thesis, however, we do not define a full knowledge representation standard, as this field is still relatively immature.

To demonstrate that "second-generation" bioinformatics tasks can generate important results from large-scale analysis tasks, we applied the approaches and methods developed in

this thesis to three different applications, all of which addressed real biomedical research questions. The applications are relevant to the fields of immunology and virology, and were chosen to demonstrate the generality and versatility of our approaches and methods. However, other fields of biomedical research can be addressed.

In our first application, we constructed a dataset of influenza A proteins that comprised all sequences available in public databases, annotated with several metadata fields. This knowledge aggregation task, performed by the ABK system, reduced to two weeks the time taken to fully aggregate, curate and verify a dataset from more than 90,000 database records. Such rapid construction of a large-scale annotated dataset, which would have been practically impossible to build manually, demonstrates the applicability our knowledge aggregation method on real analysis tasks. The availability of descriptive metadata allowed the AVANA tool to partition the dataset, separating currently circulating human-infecting lineages and avian lineages, and perform comparative analyses to identify mutations that are characteristic of human lineages. Knowledge-enabled analysis, in combination with a novel mutual information algorithm, produced a catalogue of 70 characteristic mutations involved in human transmissibility of influenza A. This catalogue is twice as large as similar catalogues produced by previous studies, showing that the combination of large-scale dataset, accurate metadata and sensitive statistical measures can dramatically extend the analytical power. From a virological perspective, this extensive catalogue of mutations has revealed new insights into the systemic nature of human host adaptation in influenza A.

In line with our knowledge mining approach, the mutations catalogue (the result of an analysis task) constitutes important new metadata that can be used in further analysis of the dataset. We extended the AVANA tool to use the catalogue of characteristic mutations to produce *adaptive signatures* of avian influenza A isolates, which show the extent of presence of human-adaptive mutations. This further knowledge-enhanced analysis task has provided a tool for assessing the human-infecting potential of avian influenza, showing that recent H5N1 strains capable of jumping the host barrier are unusually rich in adaptive mutations. These important results are a clear demonstration that knowledge transfer in the analysis pipeline

can enable entirely new analysis tasks to be carried out.

In our second application, we used the AVANA tool to perform conservation analysis of viral sequences, based on a novel *peptide entropy* measure, to identify universally conserved peptides that could be used as epitope-based vaccine targets. The large-scale annotated influenza dataset was partitioned based on the metadata, to model groups of influenza viruses that are important from an epidemiological viewpoint. AVANA performed conservation analysis within each group, and conducted meta-analysis to identified peptides conserved in all groups. This study, conducted by AT Heiny, produced a catalogue of 50 candidate vaccine targets, showing that our knowledge-enabled approached could easily be repurposed to extract new knowledge with a different perspective. The generality of our conservation analysis was further demonstrated by a similar study, performed by AM Khan on a dataset comprising over 12,000 dengue virus sequences, aggregated using the ABK tool. This study yielded a set of 34 candidate vaccine components, showing that the method can be successfully reused on multiple pathogens.

The third application extended our knowledge mining approach to a completely different problem: the identification of biomedical publications that contain information on allergen cross-reactivity, for the purpose of curation of an immunological database. The knowledge aggregation task collected large a large dataset of biomedical abstract, allowing the user to construct a metadata model by specifying relevant text features, and by annotating some examples. We demonstrated that standard machine learning software, given this metadata about the text abstracts, can select further relevant documents, reducing considerably the manual curation effort needed. Furthermore, user annotation of selected abstracts, effected by simple gestures, is also used to improve the accuracy of this text mining method. This application shows that the knowledge mining principles (large-scale dataset, knowledge-enabling and biologist-friendly interface) can be applied to a completely different class of problem.

Figure 1-2 shows a high-level overall model of the applications covered in this thesis. Each analysis “stage” should not be thought of as a single task, but rather as a number of

cascaded knowledge mining tasks, which will be described in detail in the relevant chapters. All three applications presented in this thesis scaled to handle tens of thousands of records on standard desktop computing hardware commonly available to life scientists, and there are no inherent limits to the scalability of the techniques used. Larger datasets can be handled by our knowledge aggregation methods and statistical measures, and the extent of automation in the knowledge aggregation tasks means that these methods will continue to be applicable as the volume of data increases.

In summary, our biological knowledge mining approach, through the application of knowledge aggregation and scalable knowledge-aware analysis tools, has enabled new classes of knowledge discovery tasks, which have revealed new, detailed knowledge about important immunological and virological problems. These approaches can be controlled by biomedical researchers, and do not require ad-hoc programming for task customization. The approach, techniques and tools presented herein constitute important contributions to biomedical research, and pave the way for second-generation bioinformatics analysis to become a reality.

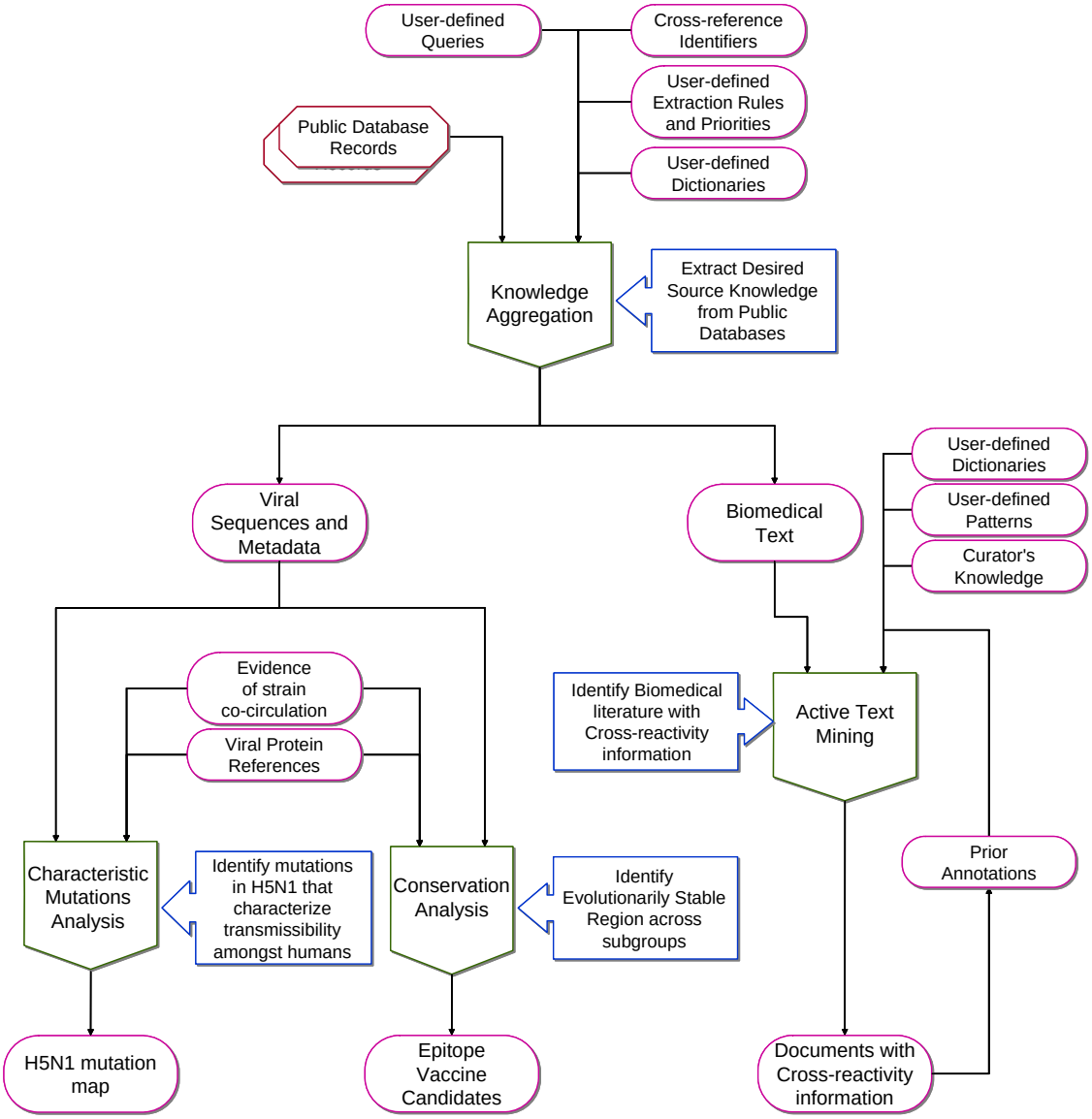


Figure 1-2: High-level representation of the Biological Knowledge Mining applications. Knowledge Aggregation, conducted using the ABK tool, constructs large-scale datasets by querying multiple databases with user-specified queries, and extracting the desired values using user-specified structural rules. This is the first stage for the three Biological Knowledge Mining tasks detailed in this thesis. The Conservation Analysis task performs a meta-analysis of conserved peptides in multiple co-circulating viral lineages, and selects conserved peptides predicted to bind to HLA molecules. The Characteristic Mutations Analysis task compares human-infecting influenza strains with avian strains, using mutual information to identify amino acid sites that present characteristic mutations. In the Active Text Mining application, free text abstracts are analyzed based on user-defined dictionaries and patterns, and on previously annotation abstracts. A simple user interface allows an expert to rapidly annotate top-scoring documents, and these annotations are fed back to improve classification accuracy.

1.3 Key Contributions

The original work presented in this thesis makes several important contributions by the author to the fields of bioinformatics, immunology and virology, which are summarized here:

- A new *biological knowledge mining* conceptual framework for modeling "second generation" biological discovery processes, in which knowledge flows through analysis pipelines consisting of multiple cascaded tasks.
- A novel *knowledge aggregation* method based on *structural rules* for extracting, aggregating and reconciling information from multiple heterogeneous biological databases, regardless of their native data structures. This method includes user-friendly interfaces for specifying rules, and mechanisms for overcoming information heterogeneities, including rule prioritization and text filters based on user dictionaries.
- The *Aggregator of Biological Knowledge (ABK)*, a desktop tool for performing large-scale knowledge aggregation tasks. ABK implements XML-based structural rules, and accesses diverse databases by means of an extensible mediator framework. Through an intuitive graphical user interface, ABK allows users to visualize and manage the extracted knowledge, which can be processed by plug-in analysis tools.
- An annotated dataset built from over 90,000 influenza A database records from GenBank and GenPept, complete with metadata describing sequence provenance, host organism, geographic origin, *etc*. The dataset was automatically curated by the ABK tools, then verified and completed by human curators. It was subsequently used for two major genome-wide analysis tasks. This work has been extended by collaborators to the analysis of other viruses, including dengue, rabies and hantavirus.
- An evaluation of the semantic heterogeneity of sequence records metadata in GenBank and GenPept, two key data sources for bioinformatics research.
- A proof-of-concept demonstration that aggregated biological knowledge, expressed using standard semantic technologies (RDF and OWL), can be processed by generic software. The author shows that simple semantic rules can be used to improve

metadata structure and quality.

- A method for the rational identification of stable vaccine targets across whole viral genomes. This method identifies conserved peptides by *information entropy* analysis, and assesses the presence of immune epitopes by applying predictive algorithms. This method, jointly developed with biologists AM Khan and AT Heiny at Dept. of Biochemistry, was used by these researchers to identify whole-genome catalogues of conserved potential T-cell epitopes for both influenza A and dengue viruses.
- A novel method, based on *mutual information*, for identifying mutations which are characteristic of an aligned set of sequences, by comparison with other homologous alignments. This method supports the processing of large numbers of sequences, and processed our large-scale influenza A dataset to identify adaptive mutations implicated in host range determination.
- The most complete catalogue to date of amino acid mutations involved in the adaptation of influenza A viruses to transmissibility amongst human hosts. The catalogue comprises 70 amino acid sites in eight internal influenza proteins, indicating that host adaptation of this virus is complex and systemic, requiring the participation of entire protein constellations.
- A novel method of producing genomic *adaptation signatures* from the catalogue of characteristic sites, to visualize the presence of adaptive mutations in influenza isolates. Adaptation signatures are a powerful tool for assessing the potential for human infectivity and transmissibility of H5N1 and other avian influenza viruses.
- The Analyzer of Antigenic Variability (AVANA), a knowledge-enabled desktop tool for performing information-theoretical analysis of sequence alignments. AVANA analyzes peptide diversity and conservation using information entropy, and is capable of comparative analysis based on mutual information. AVANA allows comparison and meta-analysis of alignment subsets selected based on metadata values.
- A novel user-driven text mining method for document classification, which supports

database curation tasks by focusing the curator's efforts on relevant documents. The method is customizable by end users without the need for programming. Knowledge acquired through annotation is injected into the text mining process by means of an *active text mining* process, which gradually improves text mining performance.

.

1.4 Structure of this thesis

The first two chapters of this thesis introduce the field of Biological Knowledge Aggregation and Biological Knowledge Mining, analyzing the current problems inherent to the evolution of bioinformatics from small-scale entity-based discovery to large-scale systemic discovery.

- Chapter 1 provides an introduction, in which the background, aims and structure of this work are presented.
- Chapter 2 is a review of Biological Knowledge Mining, and Biological Knowledge Aggregation which is its major component. We review the knowledge analysis needs of the post-genome era, and discuss the need for scalability. We present the most significant obstacles to bioinformatics scalability in three dimensions: quantitative, integrative and hierarchical; some currently available solutions are evaluated. We introduce a framework for modeling "second-generation" bioinformatics tasks, and discuss the role of knowledge flow in this model.

In Chapters 3 to 5, we propose the Biological Knowledge Aggregation method and other Biological Knowledge Mining techniques, illustrating their effectiveness in overcoming many of the scalability obstacles discussed in Chapter 2.

- In Chapter 3, *structural rule-based knowledge aggregation* is discussed as a strategy for integrative scalability of bioinformatics. The ABK software for knowledge aggregation is presented, and we discuss its features and capabilities. We report the results of a knowledge aggregation task: the creation of a dataset of influenza A protein sequence, complete with descriptive metadata. These results have been used

to assess the extent of heterogeneities in public databases, and the effectiveness of structural rules in overcoming such heterogeneities.

- Chapter 4 discusses the use of *semantic technologies* for the representation, encoding, storage and interchange of biological knowledge, as required by large-scale bioinformatics analysis. We assess the quality improvement that can be achieved by applying *semantic rules* and *reasoning* to the aggregation of large-scale datasets.
- Chapter 5 presents techniques derived from information theory, capable of scaling the study of sequence variability to populations of tens of thousands of related sequences. We define *peptide entropy*, a measure of antigenic variability, which is applied to the identification of conserved immunogenic epitopes. We also present a novel use of *mutual information*, as a measure of the association of mutations with sequence alignment subsets. The AVANA software tool, developed by the author to support knowledge-enabled information-theoretical analysis of sequence alignments, is also described.

Chapters 6 to 8 detail the methods and results for three applications that implement the Biological Knowledge Mining approaches proposed.

- In Chapter 6 we present the methods, results and conclusions of a large-scale study of influenza A protein sequences, aimed at identifying mutations involved in host adaptation to humans, and at assessing the pandemic potential of H5N1 avian viruses. The large-scale influenza A protein dataset presented in Chapter 3 was analyzed using mutual information (detailed in Chapter 5), to produce a full-genome map of mutations that characterize circulating human-transmissible viruses. The mutations catalogue was then used to produce isolate adaptation signatures, which provide a means of assessing the potential of avian strains to circulate among humans..
- Chapter 7 discusses the application of information entropy to the identification of

conserved regions in viral genomes (detailed in Chapter 5), as part of a rational method for identifying potential vaccine targets (see Appendix A). The results of applying this method on two different viruses (influenza A and dengue) are briefly outlined; full results are detailed in the papers in Appendices B and C.

- Chapter 8 presents a text mining method which selects relevant literature abstracts from large datasets aggregated from a public database. The method makes use of standard data mining software components, and is able to learn from annotations made by an expert user, without requiring specialized computational knowledge. We show that this method can substantially reduce the workload of database, by focusing their work on relevant documents.

In Chapter 9, we present the overall conclusions. The contributions of this thesis are summarized and reviewed, and future research directions are discussed.

1.5 Publication summary

Most of the work presented in this thesis has been published in international peer-reviewed journals and conferences during the course of the candidature period. Within the scope of this work, the present author published four papers as first author, and co-authored three papers. Two additional papers as first author have been submitted for publication, and are under review at the time of submission of this thesis. The content of the publications is described below.

1.5.1 Published manuscripts

- **Miotto O, Tan TW, Brusic V (2005a)** described the structural rule-based knowledge aggregation approach and the ABK software implementation, detailed in Chapter 3.
- **Miotto O, Tan TW, Brusic V (2008b)** further discussed structural rule-based knowledge aggregation, and described the large-scale influenza dataset aggregation

task detailed in Chapter 3, Section 3.4. This paper also discussed the use of semantic technologies in bioinformatics discovery, and described the metadata quality improvement task based on semantic rules, detailed in Chapter 4.

- **Miotto O, Heiny AT, Tan TW, August JT, Brusic V (2008a)** details the mutual information method for identifying characteristic mutations, presented in Chapter 5, Section 5.2 of this thesis, and shows its application to the identification of host range determinants in the influenza A PB2 protein, reported as part of Chapter 6.
- **Miotto O, Tan TW, Brusic V (2005b)** described the reusable text mining method and its application to the study of allergen cross-reactivity, discussed in Chapter 8. The active learning approach is also described in that publication, which used the ABK knowledge aggregation platform to perform dataset construction and analysis tasks.

1.5.2 Manuscripts in preparation

- **Miotto O, Heiny AT, Tan TW, August JT, Brusic V (2009a)** Complete-proteome mapping of human influenza A adaptive mutations: implications for human transmissibility of avian strains (manuscript in preparation). This paper will detail the full-genome analysis of influenza A characteristic sites, and the assessment of avian influenza strains, described in Chapter 6. This work was also presented orally at the International Avian Influenza Conference, Bangkok, Thailand, January 2008.
- **Miotto O, Tan TW, Brusic V (2009b)** AVANA: a tool for analyzing antigenic variability in large sets of protein sequences. *Bioinformatics* (manuscript in preparation). This article will describe the AVANA software presented in Chapter 5, Section 5.3.

1.5.3 Published manuscripts as second author

- **Khan AM, Miotto O, Heiny AT, Salmon J, Srinivasan KN, Nascimento EJM,**

Marques ET, Brusic V, Tan TW, August JT (2006) describes the method for identifying conserved targets for peptide vaccines, discussed in Chapter 7. This method is a collaborative effort, and the author of this thesis contributed to: (a) the dataset preparation and cleaning stage, which uses the ABK platform described in Chapter 3; and (b) the entropy-based conservation analysis, described in Chapter 5 (section 5.1) and Chapter 6 of this thesis. Other elements of the method were contributed by first author Mr. A. M. Khan, and other co-authors. This paper is included as Appendix A of this thesis.

- **Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, Brusic V, Tan TW, August JT (2007)** and **Khan AM, Miotto O, Nascimento EJM, Srinivasan KN, Heiny AT, Zhang GL, Salmon J, Marques ET, Tan TW, Brusic V, August JT (2008)** applied the method described in Khan *et al*. (2006) to large-scale studies of influenza A virus and dengue virus, respectively. These two papers, which present important applications of the information theoretical approaches presented in Chapter 3, are included in this thesis as Appendix B and Appendix C, respectively. Chapter 6 of this thesis contains a very brief summary of some key results that demonstrate the usefulness of the methods described herein. The author of the present thesis contributed to both papers by providing the peptide entropy conservation analysis method and software (AVANA) and participating in the analysis of conservation. In addition, he contributed to the construction of the datasets used in both studies, which were aggregated using the ABK software. The author is greatly indebted to Mr. Khan and Ms. Heiny for applying the knowledge aggregation and information theoretical analysis methods to diverse case studies, thus demonstrating both the utility and generality of these approaches in specific case studies. The success of their studies has opened the way for studies of other pathogens: West Nile virus, Hantavirus, Rabies and HIV studies are currently ongoing, using the ABK and AVANA tools and the methods described therein.

2. BIOLOGICAL KNOWLEDGE MINING

Information is not knowledge
Knowledge is not wisdom
Wisdom is not truth

Frank Zappa

In this chapter, the field of biological knowledge mining is reviewed. We analyze the role of bioinformatics in the biological discovery process, and discuss the advantages offered by computational methods in the post-genomic era, characterized by inexpensive production of large quantities of biological data. We consider the requirements of "second-generation" bioinformatics, capable of evolving biological discovery from a reductionist study of molecular components to a systemic discipline, processing more data from an ever increasing number of sources, through a greater number of tasks than was possible with earlier methods. A number of key obstacles to bioinformatics scalability are identified. Currently available solutions, including some from related fields, are reviewed. Finally, we introduce a conceptual framework for describing and designing scalable biological knowledge mining pipelines, which will be applied to structure our approach to large-scale bioinformatics studies throughout this thesis. In Chapters 3 to 5 of this thesis, this framework is complemented with scalable bioinformatics techniques, while Chapters 6 to 8 describe the application of this framework to real biological discovery problems in selected immunological and virological examples.

2.1 Bioinformatics and Biological Knowledge

2.1.1 Bioinformatics and its role in Biological Discovery

In a definition written for the Encyclopedia of Molecular Pharmacology, Nilges and Linge (2002) have described *bioinformatics* as follows:

> "Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources,

> patient statistics, and scientific literature. Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine."

This definition captures four essential characteristics of bioinformatics: it produces knowledge from data; it uses heterogeneous experimental observations from multiple sources; it has practical applications; and it is interdisciplinary. Bioinformatics analysis extracts knowledge by organizing experimental data and applying computational methods, so that significant patterns can be identified and extracted. While scientists of the past were able to analyze experimental results manually, computational analysis has become indispensable when handling large amounts of experimental data, as are commonly produced by modern automated laboratory equipment. The capability to organize and analyze large and complex datasets, rendering them manageable and comprehensible to human experimenters, is thus a key motivation for the use of bioinformatics (Luscombe *et al.* 2001). The initial successes of bioinformatics were favoured by the information-oriented nature of molecular biology. For many purposes, nucleotide and protein sequences can essentially be reduced to strings of symbols that can be meaningfully analyzed by identifying patterns and similarities, leveraging on computational methods developed for text analysis and linguistics. Information-based analysis is instrumental in deciphering genomes (Attwood 2000), identifying protein coding regions, locating promoters, predicting protein structure, and various other tasks

From a computing perspective, most bioinformatics discovery processes are *data mining* processes, in which a *knowledge discovery* task analyzes large amounts of experimental data, extracting meaningful patterns that can subsequently be used for *predictive* tasks that describe new properties of biological entities, shown in Figure 2-1. The first stage, known as *knowledge discovery in data* (KDD) (Fayyad *et al.* 1996), is generally performed on experimental data collections, usually retrieved from *primary databases*, where they have been deposited by experimenters. KDD aims to identify patterns and rules "hidden" within the

data, which cannot be determined by visual inspection. Biological KDD based on Fayyad's framework has been described by Brusic and Zeleznikow (1999). For example, sequences that are shown experimentally to regulate protein expression can be compared to reveal conserved sequence motifs, which identify protein-binding targets (Stormo 2000). The results of KDD often help to elucidate biological mechanisms. Whenever possible, these results are expressed in a human-interpretable form and collected in *secondary databases* (sometimes known as *knowledgebases*), such as the PROSITE motif database (Hulo *et al.* 2008). In the second stage of the biological data mining process, the acquired knowledge is used to mine new data, to make *predictions* about the properties of this data. For instance, a newly sequenced genome can be scanned to recognize known motifs, thus locating new putative binding sites (Morgan *et al.* 2007). The application of acquired knowledge usually requires the construction of prediction software systems, frequently implemented using *machine learning* algorithms, capable of learning patterns from data (Brusic *et al.* 1998). Since prediction software is typically built on simplified models, its results must be considered putative, and are subject to experimental verification *in vivo* or *in vitro*. This limitation does not undermine the usefulness of these predictions: on the contrary, it positions bioinformatics as a strategy that is complementary to laboratory experiments. By predicting candidates from a vast number of possibilities, *in silico* methods can dramatically reduce the number of required laboratory experiments, lowering research costs and accelerating discovery. These advantages have earned bioinformatics a place as "equal and essential" partner for the future of biomedicine (Benton 1996), a role that has been recognized in related fields, such as drug discovery (Debouck and Metcalf 2000).

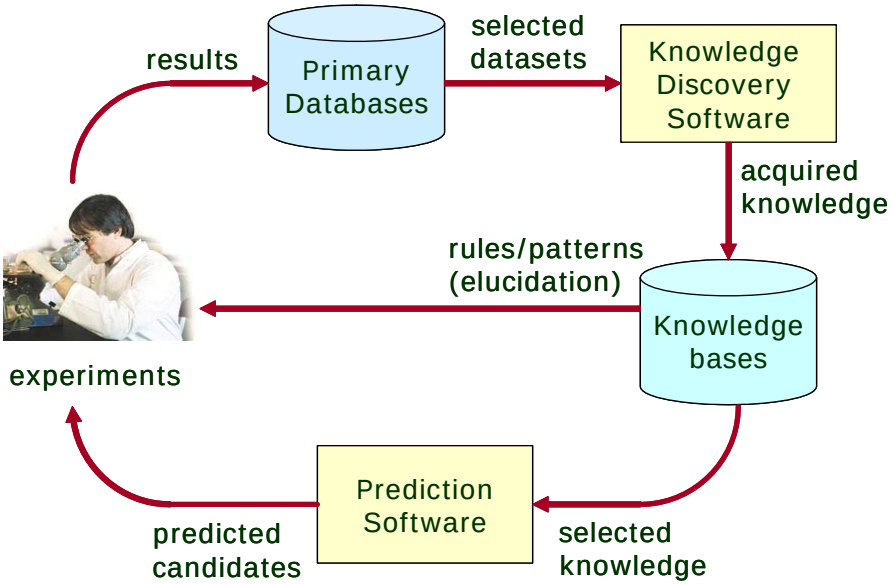


Figure 2-1: The bioinformatics data mining process and its relationship to experimental biomedical research.

Typical KDD tasks may need to combine information of different types, such as sequence data and protein expression levels. Often, they use derived knowledge as well as primary experimental data, blurring the distinction between primary databases and knowledge bases. The world's largest primary nucleotide sequence database is GenBank (Benson *et al.* 2008), which receives a continuous stream of submissions, largely driven by requirements for sequence depositing, stipulated by many important journals (Noor *et al.* 2006). GenBank's records, however, are not limited to raw nucleotide sequences: they contain sequence-level annotations to provide conceptual translations, coding sequence boundaries, functional information, and so on, many of which are produced by bioinformatics analysis. The extent to which acquired knowledge has to be included is constantly increasing, being driven by the variety of analysis tasks to be performed.

2.1.2 Data, Information and Knowledge in Biology

Since this thesis will introduce *knowledge mining* that is distinct from *data mining*, we need to provide working definitions for the terms *data*, *information* and *knowledge*, which are often used without drawing formal distinctions. In particular, usage of the word *information* tends

to overlap with that of both *data* and *knowledge* (Boisot and Canals 2004). In knowledge management, these three terms are often defined in the context of the DIKW (Data, Information, Knowledge and Wisdom) hierarchy (Zeleny 1987):

- ***Data*** comprises raw recordable values, such as the results of measurements. The creation of data may be completely automated, without human intervention. Examples of data in biology are nucleotide sequences generated by a sequencer, or intensity values obtained by scanning a microarray slide. Typically, data has syntax (i.e. a *data format*) but no structure.
- ***Information*** is data that has been transformed to facilitate its use: formatted, structured, published, and so on. An example of biological information is an experimentally-derived peptide sequence, stored in a database and accompanied by descriptive data, such as organism, strain, date, clinical condition, and so on. Information often has complex structure, a context, and relationships to other information.
- ***Knowledge*** may be defined as the interpretation of available information, in the light of a given *context*. Knowledge is generally viewed as deterministic- in other words, it can be derived from the analysis of existing information, at least in principle. However, the derivation of knowledge often requires complex expert input, and may be difficult to automate (*machine learning* approaches, such as those described in Chapter 8, attempt to bridge this gap by building predictive that incorporate existing knowledge). The analysis of multiple pieces of information is usually required when producing knowledge. Conversely, the same piece of information can generate multiple types of knowledge, through the application of different analysis processes. There are many examples of knowledge generation in biology: one example is the assignment of a putative function to a protein, by detecting its sequence similarity to other proteins of known function in a database; another example is the derivation of binding motifs from the analysis of multiple binder sequences.

- ***Wisdom*** is the extrapolative and non-deterministic product of accumulated knowledge, which affects *future* use of the acquired knowledge. Wisdom synthesis requires consciousness and cultural values such as morality and ethical codes. As a result, it is generally accepted that wisdom is a social product that cannot be derived by machines. Some models (Ackoff 1989) distinguish between wisdom and ***understanding***, the latter consisting of the human analysis of *past* knowledge, used to construct wisdom that determines *future* actions.

Data, information and knowledge can be described as three levels at which we structure our perception of the underlying reality (i.e. the biological entities and processes). Chun *et al.* (2000, pp 29-32) describe a "Data-Information-Knowledge Continuum" in which *signals* (such as scanned microarray fluorescence levels) are successively structured into data, information and knowledge, increasing both *order* and *human agency* at each structuring stage. Data is characterized by *physical structuring* (for example, the storage of a nucleotide sequence as a string of characters) resulting from syntactic and lexical choices; information is the result of *cognitive structuring* (e.g. assigning meaning to the data items in database records); knowledge requires *deductive structuring* (e.g. the application of rules that generates new facts). The acquisition of structure is accompanied by the acquisition of context: in order to build a structured representation of reality, we are required to choose a point of view from which the information and knowledge are valuable. Context is determined by purpose, and the same data can therefore be structured and contextualized differently to suit different purposes. For example, describing the function to a protein sequence does not produce any phylogenetic knowledge about the organism that produced it; however, analyzing the same sequence against its homologues in different species can produce such knowledge. Thus, cognitive structuring is contextual: although the same sequence data is utilized for two purposes, each requiring different relationships and annotations. As a result of these context dependencies, organizations with diverse knowledge often describe the same reality with different information architectures, database structures, information encodings and representations.

Such differences clearly pose serious obstacles to information sharing and exchange, and make information and knowledge difficult to reuse in different contexts. It is important to underline that these obstacles are not necessarily the product of organizational problems or lack of standards, nor are they specific to the biomedical domain: they arise because information and knowledge are intrinsically contextual in nature (McCarthy 1987). Contextual dependencies have profound repercussions on the management of shared biological knowledge. To share experimental information and derived knowledge, bioinformaticians have produced a multitude of databases, varying in size, scope and purpose. However, such information and knowledge are often reused in a context which is different from the original intention or scope. This context mismatching causes gaps in the structuring continuum, producing information heterogeneities that are major obstacles to the biological discovery process. These heterogeneities are among the key motivators of the knowledge aggregation approaches and knowledge mining described in this thesis, and will be discussed in detail in this chapter.

2.1.3 Metadata

In previous sections, biological data was described as the outcome of experimental measurement, such as genomic nucleotide sequences, mass spectrometry readings, or protein expression levels. In most cases, raw data alone is insufficient for analysis unless additional descriptive data is provided. For example, protein expression levels measured by a microarray assay will yield no knowledge unless they are accompanied by information about organism, tissue type, disease condition, and so on. Such descriptive information is known as *metadata* ("data about data"). Metadata plays an essential role in bioinformatics analysis, by providing a *context* for interpreting data. Different meanings can be ascribed to the term metadata, arising from different usage perspectives (Lundy 1984): at system level, metadata may describe the encoding of data; in databases, it defines the meaning of data fields; in database records, it may describe the conditions under which the data is valid, under which measurement were recorded, and so on. Within this broad spectrum of uses, two broad classes

of metadata can be defined (Bretherton and Singley 1994):

- ***Control metadata***, such as database schemas, is intended to aid machine processing of data. It captures lexical, syntactic and structural aspects of data interpretation, and is built into the processing software, preferably as a result of implementing standards.
- ***Guide metadata***, such as descriptions of experimental conditions or literature references, is intended to help humans or machines *reason* over the data. Although guide metadata is often encoded in natural language, this is not a requirement, since its presentation to human users is often mediated by machines. The essential feature of guide metadata is its role in decision-making tasks: logical operations can be performed on metadata in order to store, retrieve, sort, combine, analyze, and present the relevant data (Long JM 1986). Guide metadata is frequently encoded in data records, alongside the data it describes: for example, a GenBank sequence record contains the nucleotide sequence (data), as well as taxonomy identifier, literature references, and so on (metadata). Therefore, guide metadata can be viewed as "just data" and we consider data and guide metadata as one when discussing structural aspects of information. In this work, however, we will distinguish between data and metadata in terms of the role played in the analysis process.

Both classes of metadata are essential for performing bioinformatics tasks, although they affect different aspects of the knowledge discovery process. Mismatches in control metadata between information producers (such as databases) and information consumers (such as analysis tools) produce information heterogeneities, which impede the correct decoding and usage of the data, as will be discussed later in this chapter. Errors or omissions in guide metadata, on the other hand, affect the analysis and decision-making process, impairing knowledge acquisition. For example, in a comparative analysis of viral protein sequences from different hosts, records with missing metadata will cause sequences to be discarded, while errors will assign sequences to the wrong set. In both cases the accuracy of the analysis

results is diminished, and therefore their significance.

Because of the impact of guide metadata deficiencies, it could reasonably be expected that descriptive metadata should be submitted to database storage with the same attention to detail as the main data. Unfortunately, this is not the case: we will show in Chapter 3 that metadata errors and omissions plague primary databases such as GenBank, making the construction of large metadata-rich datasets a very arduous task. This state of affairs may largely be due to a lack of appreciation of the importance of metadata for the reuse of the deposited data. Even journals that require sequence data to be deposited prior to publication (Noor *et al.* 2006) do not normally stipulate metadata quality requirements. To compound the problem, metadata errors and omissions in primary sources are almost impossible to correct by third parties, and therefore errors tend to propagate (Bidartondo 2008). The problem of guide metadata recovery is a central issue of this thesis, and will be discussed in Chapter 3 and 4, where it will be addressed by novel rule-based methods. Since guide metadata is the class of metadata most frequently discussed in this work, the word "guide" will often be omitted- for example, the phrase "sequence metadata" should be interpreted as "guide metadata about a sequence".

2.1.4 Digital Repositories of Biological Information and Knowledge

Experimental data is frequently made public upon publication of results, fulfilling two objectives: reproducibility of the analysis task, and repurposing of the data for different type of analysis. Digital repositories of biological data (*biological databases*) collect and store data submitted by multiple research studies, which can be aggregated and examined in different contexts. Within these databases, biological data (such as nucleotide sequences) is normally accompanied by descriptive metadata, which details experimental information, relationships to other records, and provenance information such as a reference to a relevant publication, which may be important to assess the trustworthiness and quality of the data. The number of publicly accessible biological databases is growing continuously: the Molecular Biology Database Collection, a yearly-updated catalogue of biological databases, described over 1000

high-quality databases at the end of 2007, an increase of 11% since the previous year (Galperin 2008). The list encompasses a highly diverse range of database type (shown in Table 2-1), reflecting the diversity of intents of the many research organizations that manage these repositories.

Not all of the databases listed are repositories of experimental data: many are value-added knowledge repositories, in which data from other databases is selected, aggregated, summarized and analyzed. Thus it is useful to distinguish between two broad categories of biological databases: generic and specialized (Brusic and Koh 2004). Metaphorically, the difference between these two categories is similar to that between a wholesale market and a specialized boutique. A boutique supplies a small range of products, selected for their quality and relevance, and provides significant support for choosing the correct product. The products available in a boutique may also be obtainable from the wholesale market, but they are far more difficult to find among a huge range of other products, and may not be available in the desired form, size or quantity. However, the variety on sale at the wholesale market may be advantageous when a single boutique does not offering all the products we need.

Generic databases such as GenBank (Benson *et al.* 2008), UniProt (UniProt Consortium 2008), PubMed (Wheeler *et al.* 2008b) and the Stanford Tissue Microarray Database (Marinelli *et al.* 2008) are the wholesale markets of biological data. Because of their emphasis on coverage, these databases allow the retrieval of comprehensive datasets (for example, "all dengue virus sequences for a given organism") and "needle in a haystack" searches (for example, finding homologues of a sequence in multiple organisms). Generic databases need to handle a high volume of new submissions, and therefore can only offer limited centralized curation; as a result, individual records often present errors, inconsistencies and omissions (Brusic and Koh 2004). In addition, because of their general purpose nature, records in such databases support a limited set of descriptive annotations.

Database Type	Count
Inter. Nucleotide Sequence Database Collaboration	*3*
Genes, motifs and regulatory sites	*39*
Structure, introns/exons, splicing	*24*
Transcriptional factors	*48*
Nucleotide Sequences	**114**
RNA Sequence, Structure and Functions	**54**
Sequences	*13*
Properties	*14*
Localization and targeting	*21*
Motifs and active sites	*21*
Domains, classification	*33*
Databases of individual protein families	*58*
Protein Sequences	**160**
Annotation terms, ontologies, taxonomy	*21*
General genomics	*36*
Viral genomes	*21*
Prokaryotic genomes	*55*
Unicellular eukaryote genomes	*14*
Fungal genomes	*27*
Invertebrate genomes	*47*
Model organisms, comparative genomics	*49*
Human genomes, maps and viewers	*19*
Human proteins	*19*
Genomics	**308**
Arabidopsis thaliana	*18*
Rice	*14*
Others	*42*
Plant Databases	**74**
Organelle Databases	**21**
Small molecules	*12*
Carbohydrates	*9*
Nucleic acid structure	*10*
Protein structure	*56*
Molecular Structures	**87**
Enzymes and nomenclature	*13*
Metabolic pathways	*11*
Protein-protein Interactions	*42*
Signaling pathways	*11*
Metabolic and Signaling Pathways	**77**
Microarray Data and Gene Expression	**41**
Proteomics	**11**
Other Molecular Biology	**30**
Immunological Databases	**21**
Human genetics	*10*
Polymorphisms	*25*
Cancer genes	*20*
Gene-, system- or disease-specific	*42*
Human Genes and Diseases	**97**

Table 2-1 - Biological databases in the Molecular Biology Database Collection
Data is summarized from Galperin (2008). A small number of databases were listed in multiple categories.

Specialized databases tend to have a narrower, well-defined scope: they may concentrate on specialized topics, such as specific pathogens (e.g. PlasmoDB (Stoeckert *et al.* 2006)), cellular components (e.g. Mitome (Lee *et al.* 2008)), class of proteins (e.g. TOPDB (Tusnády

et al. 2008)), to name just a few. Because of their narrower scope, these databases contain a far smaller selection of records than generic databases, and are usually curated by experts in the area of coverage of the database. Dedicated curation effort results in high-quality, carefully edited records that may include specialized detailed annotation and links to related resources. Specialized databases often offer features that extend the usefulness of their data, such as integrated analysis tools that support the common relevant analysis tasks. As an example, the VISTA genome database provides a variety of comparison tools for identifying similarities in genomes of different species (Frazer *et al.* 2004). Many specialized databases are *secondary* repositories, in that they aggregate data retrieved from *primary* repositories of experimental data such as GenBank, often augmenting it with analysis results and data from other sources. In many other cases, however, specialized databases are managed by research groups that produce relevant data, which may not be available elsewhere. Although the creation and management of "boutique" databases by expert research groups is beneficial for data quality, it does present a number of disadvantages. Since many biomedical research groups are not supported by dedicated IT infrastructures capable of designing and implementing complex databases systems, poor support of standards and deficient data structures often characterize smaller biological repositories. In addition, most biological databases are implemented *ad hoc*, using unique access mechanisms and data formats, which translate into difficulties in querying, retrieving and interpreting the data automatically. Indeed, several databases provide web interfaces that are clearly intended for human browsing rather than automated retrieval, probably based on the assumption that data of interest is normally collected manually by researchers.

2.1.5 Dissemination of Biological Knowledge through Text

The "gold standard" method of transferring knowledge in the biomedical research community is through publication in peer-review journals or conferences. From a scientific perspective, peer-reviewed publication is an integral part of the scientific process. It provides quality control, standard dissemination channels, and recognition of scholarship (Ruben 2003).

Publications in print also provide a very expressive medium: they support unstructured information, such as images, diagrams, logical discourse and descriptive natural language, as well as structured information in the form of tables and citations. The unstructured format of free text allows diverse kinds of knowledge to be combined in a single document, with a freedom of expression that electronic data repositories cannot currently match. At best, current state-of-the-art knowledge management technologies can capture complex knowledge as a set of simply structured statements using controlled vocabularies (*ontologies*), but are unable to encode all the nuances supported by human language. In natural language text, *explicit* knowledge is routinely supplemented by *tacit* knowledge, which is implicit and assumes shared experience and mindset (Nonaka and Takeuchi 1995), and therefore usually absent from databases. An important category of tacit knowledge is that of the *rules* we apply on information to generate new knowledge (André 2002). Although several approaches for making rules explicit (mathematical notations, workflows, logic languages) exist, they are generally difficult to apply, and it is therefore more common for rules to be built into software tools than to be expressed in reusable terms.

The conversion of human knowledge (such as publication contents, experimental data or user knowledge) to an electronic shared form can be performed on both explicit and tacit knowledge (Table 2-2). Most knowledge transfer to databases consists of *dissemination*, the encoding of explicit knowledge into a suitable shared schema. For example, the annotation of UniProt protein sequences with domain information derived from structural analysis represents explicit knowledge encoding. The *externalization* of tacit knowledge is a more complex process, requiring manual curation by highly qualified researchers who read academic publications and synthesize facts contained within the text into a machine-consumable electronic form. This conversion process is usually limited to specific portions of the total knowledge contained in any given publication. The process is dictated by the curator's specific objectives and the knowledge types supported by the target database. Some databases, such as OMIM (McKusick 2007), choose to present knowledge extracted from literature as free text, thus creating specialized literature digests. Although this approach

allows facts to emerge explicitly, it does not solve the problem of providing the information in a format that can be processed by analysis tools.

From \ To	Tacit Knowledge	Explicit Knowledge
Tacit Knowledge	Socialization	Externalization (capture and storage)
Explicit Knowledge	Internalization (reuse)	Combination (dissemination)

Table 2-2 – Knowledge Conversion Matrix (Nonaka and Takeuchi 1995)

The extraction of knowledge from peer-reviewed publications has been revolutionized by technological advances in digital publishing and by the growth of the Internet. All leading academic journals now provide indexed and searchable electronic versions of published papers, which can be downloaded online. New publication channels have resulted in continuous growth of the journal sector (Gooden *et al.* 2002), along with a dramatic reduction of the time needed for researchers to access the articles they require, as well as wider and more equitable access to journals (Kmietowicz 2001). In the electronic publishing world, subscriptions managed by libraries, corporate resource centres, and content brokers control access to Web resources. The scientific community is applying increasing pressure on publishers to make journal content freely available, to improve research dissemination. A number of major agencies have instated public access policies, demanding that papers funded under their research grants must be made freely available shortly after publication; these agencies include the National Institutes of Health (NIH 2008), the Howard Hughes Medical Institute, the Wellcome Trust, and other major European funders (Doyle et al. 2003). At the same time, emerging open access publishers have successfully introduced new business models that transfer publication costs to producers rather than consumers (Delamothe and Smith 2004). The trend indicates that the availability of full text content will continue to increase over the next few years, leading to almost universal free access to published research.

This has profound implication for the conversion of text-encoded knowledge to machine-readable form: free access to full text will enable automated agents to search with more accuracy for relevant research, facilitating database curation. These new opportunities have fuelled much research in automated biomedical text mining, which could drastically reduce the work of database curators (Erhardt *et al.* 2006; de Bruijn and Martin 2002). At present, many text mining tools obtain their input data from freely available abstracts from the PubMed database (Wheeler *et al.* 2008b), which currently indexes over 17 million articles. In the future, it is likely that PubMed will retain its primary role as an index resource, while text mining tools will retrieve full text data directly from journal sites. However, text mining tools are almost invariably bespoke applications that are problem-specific and very complex to build. Chapter 8 of this thesis presents a knowledge mining application that analyzes PubMed abstracts, to facilitate curation of biological databases. Rather than extract facts from the text, the application selects the most relevant documents to be read and analyzed by a human curator. Such document *classification* task can significantly reduce the search effort making it manageable to find relevant knowledge amongst hundreds of publications. The approach presented here is unique in that it contains built-in problem-specific software is needed, and a database curator can customize the application without programming expertise.

2.1.6 From Laboratories to Repositories

We have surveyed and summarized the diverse arsenal of knowledge repositories available for further bioinformatics analysis, and have described how data, information and knowledge flow from experimental laboratories to primary databases, journal publications and specialized knowledge bases. The knowledge flow (summarized in Figure 2-2) involves multiple factors and participants, and provides multiple paths for knowledge to reach its destination repositories. Every passage in this flow may present challenges in knowledge access, retrieval, transformation and representation, all of which can introduce gaps and errors in the resulting databases. This is clearly of central importance to the knowledge discovery process, since bioinformatics analysis depends on the correctness and completeness of the

data and metadata used in the analysis. In the following sections, we will discuss several key problems, which must be addressed in knowledge mining tasks.

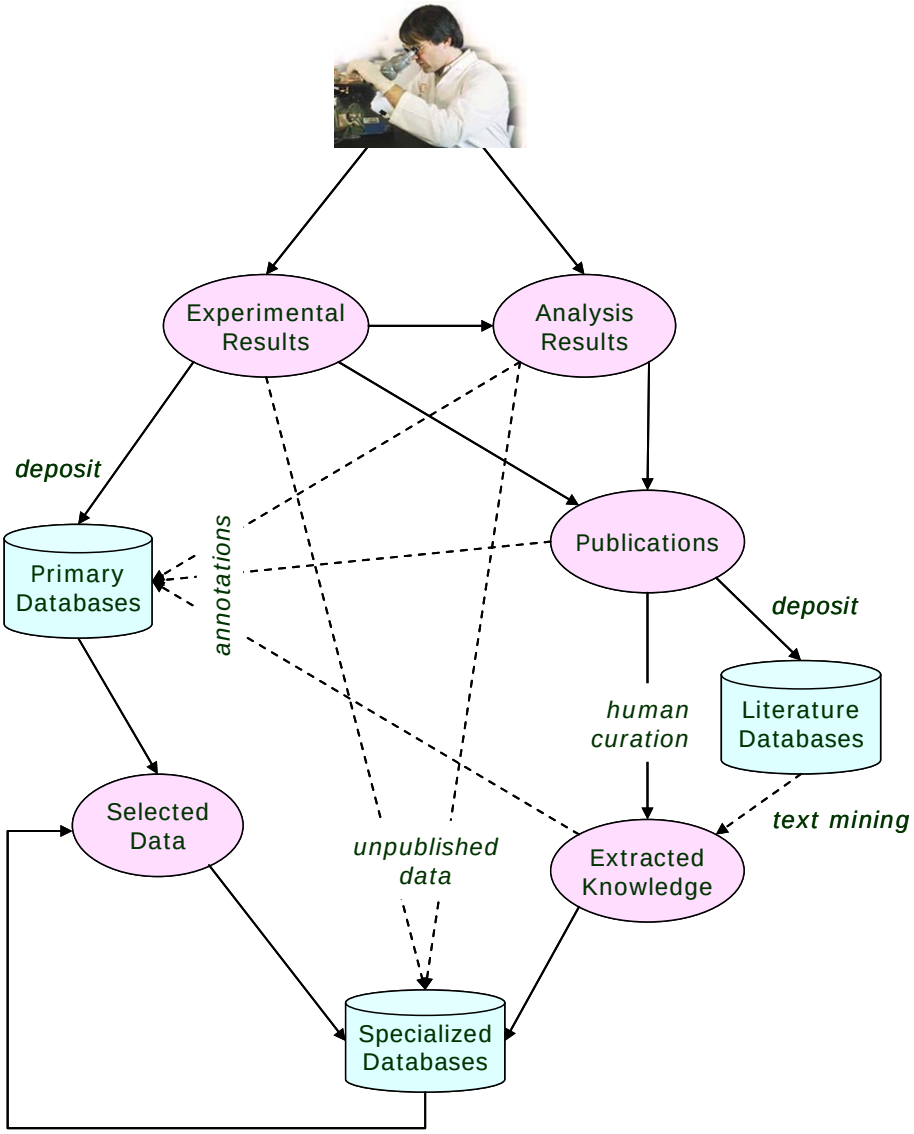


Figure 2-2: Knowledge flow from laboratories to repositories.
Data is produced and analyzed by experimental researchers, who publish their results and conclusion in biomedical papers, which are indexed by literature databases. The experimental data is usually deposited in primary biological databases at the time of publication. Specialized databases tap on both published and unpublished data, collecting selected deposited data from primary databases and expanding annotations. Knowledge in specialized databases is often derived from literature, either manually by a human curator, or automatically by text mining. To a smaller extent, automatic annotation and human curation are used for curation of primary databases.

2.2 Opportunities in Bioinformatics Knowledge Discovery

2.2.1 Bioinformatics in the Post-Genome Era

The Human Genome Project has been a watershed event in the history of biomedical research, not only because it provided the first full reference human genome sequence, but also because of the great technological impetus that accompanied this project. Like NASA's missions to the Moon, which spurred great advances in semiconductor technologies and fuel cells, the colossal challenge of assembling the first human genome has encouraged the development of new methods in biotechnology, instrumentation, and computing, reducing enormously the cost of sequencing. At the same time, the availability of the full genome has revealed a wide range of scientific discovery possibilities: comparative genomics against other species, mapping of human polymorphisms and haplotypes, identification of disease-linked mutations (Austin 2004). One tantalizing prospect is the ability to link features of human genomes with those in genomes of other organisms involved in disease: for example, by simultaneously analyzing genomic data from malaria parasite *Plasmodium*, its arthropod vector *Anopheles*, and its human hosts (Hoffman *et al.* 2002). However, the identification of genetic factors associated with disease requires the analysis large numbers of genomes to achieve statistical significance (Forton and Kwiatkowski 2006). Thus, the completion of the first human genome has paved the way for the sequencing of thousands of other genomes, including humans, mammals, animal models, and other organisms and viruses, including pathogen sequences. This might have seemed an unrealistic goal just a decade ago, but the pace of advances in sequencing technologies leaves little doubt that it will be achieved. Additional pressure to reduce the cost of sequencing comes from the emergence of personalized genomics and personalized medicine (Wheeler *et al.* 2008a), a new field with high economic potential. Current technology trends are promising: a new sequencing approach introduced by Solexa (Bennett *et al.* 2005) has recently lowered the cost of sequencing by two orders of magnitude, suggesting that affordable personalized genomics will soon become a reality. Alongside the development of rapid and inexpensive genomic sequencing, comparable advances have been

made in microarray technologies. Not only has array miniaturization improved to yield extremely high densities, but the variety of applications has also been extended to encompass comparative genomics, sequencing, methylation and others alongside the more traditional protein expression profiling (Cowell and Hawthorn 2007).

Until now, the growth in the volume of data has not fundamentally changed the nature of knowledge flow in bioinformatics. Laboratory-to-repository knowledge sharing mechanisms, via primary data repositories, peer-reviewed publications and value-added knowledgebases, will remain the main data sources for bioinformatics analysis. We anticipate that the scale on which knowledge sharing takes place will change dramatically in the near future. Data records will expand in size from the few kilobytes needed to store a single sequence, to several megabytes, gigabytes, or even terabytes required to encode whole genome sequences, expression levels from millions of microarray spots, or detailed images of pathology specimens. It is reasonable to predict that new secondary databases will emerge to cater for new types of data, and for the study of new biological entities and processes. The bioinformatics analysis pipeline will therefore need to scale up in multiple dimensions: by devising new ways to store, transfer, encode and analyze higher data volumes; and by enabling the aggregation of information and knowledge from a growing multitude of diverse data sources. As the focus shifts on comparative studies, studies of multiple entities, and functional studies, the availability of metadata will hold the key to properly controlled analysis and set selection, for the different types of biomedical data to be analyzed (van Vlymen and de Lusignan 2005). Thus metadata storage, retrieval, quality and representation will grow in importance, as will the capability of tools to make use of descriptive metadata in analysis, and flow knowledge to other tasks. These are very significant new challenges, which require a mindset shift in bioinformatics: although analysis algorithms will continue to play a central role, the management of large quantities of data, encoding of detailed metadata and connectivity to heterogeneous systems, are the domain of software engineering and systems engineering, rather than computer science (National Research Council 2007; Parker *et al.* 2003). Metadata plays a central role in the work presented in this thesis: the knowledge

aggregation approach presented in Chapter 3 produces datasets which are annotated with metadata aggregated from multiple data sources. In Chapter 4 we show that this metadata can be encoded and propagated along the analysis pipeline using semantic technologies, which allow the metadata to be enriched using rule-based reasoning. Finally, the information theoretical approaches presented in Chapter 5 use metadata to select subset alignments and conduct comparative analyses and meta-analyses. The AVANA tool, which implements these approaches, is metadata-enabled as it accepts annotations of the input sequences.

2.2.2 Bioinformatics in the Post-Sequence Era

Another important paradigm shift, catalyzed by the rapid increase in the volume of available data, is the emergence of analytical tools that consider biological *systems* rather than individual components. The high cost of experimental molecular biology, and the resulting low data volume, has traditionally limited bioinformatics to the study of individual molecular components, such as DNA or protein sequences. Such small scale approaches have produced remarkable results, enabling the elucidation of basic cellular and molecular mechanisms, and sometimes identifying causes of disease. These successes have encouraged a *reductionist* view of biology, according to which different aspects of biological mechanisms can be studied separately, thus reducing the systemic complexity of the problem. Reductionism has been applied successfully in the physical sciences, and the deciphering of the physiochemical nature of fundamental processes such as genetic replication, transcription and translation has led some to believe that the reductionist approach could answer all biological questions, as summarized by Francis Crick: “an organism is essentially nothing but a collection of atoms and molecules” (Crick 1966). This philosophical view has been rejected by *autonomists*, who believe that biology cannot be rationalized in the same way as physics, but on the contrary can only be explained in functional terms (Dupre 1986). Both sides have valid points: undeniably, reductionist methods have produced breakthroughs, such as models of the building blocks that enable the identification of the function of proteins, or of genetic determinants in some diseases; however, these may be “low-hanging fruits”, and long-term

progress may be dependent on methods that take a broader perspective. Most diseases have complex causes, involving a myriad of contributing factors, which interact in non-obvious ways. Even if we accept that, in principle, biological systems can be explained in biophysical and biochemical terms, we currently have neither the sophisticated models nor the computational means to "make the leap" (Ogbunugafor 2004) between simple molecular observations and complex phenomena such as disease. The new field of *systems biology* has emerged to bridge this gap, by studying how biological entities (DNA, RNA, proteins, regulators and small molecules amongst others) interact to produce complex phenotypic results. The study of biological systems is expected to yield intermediate models that account for the interaction of molecular components, their assembly, their regulation mechanisms, and so on, as well as novel strategies for the rational development of therapies that can control or interfere with disease-related processes (Kitano 2002). Biological models can be applied at multiple levels: molecular, cellular, organ/tissue, organism, and groups of organisms (Motta and Brusic 2004). The applications presented in this thesis focus primarily on modeling molecular entities from data assembled pertaining to multiple pathogens.

The *post-genome* era is therefore also a *post-sequence* era, in which bioinformatics analysis is characterized not only by growing volumes of data, but also by the increasing diversity of data types involved in analysis tasks, and by the growing numbers of analysis tasks that will have to be applied. This "second-generation" bioinformatics, as we term it in this work, will shift its focus from entity data in primary repositories, to knowledge-rich secondary databases (knowledgebases) that allow multidimensional information to be combined in the same study, and the resulting knowledge will pave the way for more complete models of cells and organisms (Kanehisa and Bork 2003). To cater for multiple different types of analyses, the arsenal of tools used in bioinformatics is also destined to grow, and individual tasks will need to be combined within complex analysis pipelines. At the National Institutes of Health, this emerging paradigm has been dubbed "*digital biology*", that is, the combination of large-scale scientific data integration, multi-scale modeling (the ability to study different systemic aspects of biological phenomena) and networked science (the

ability to "wire together" multiple knowledge tasks) (Morris *et al.* 2005). Achieving highly sophisticated integration, modeling, and networking is a challenge to the growth of the field of bioinformatics. This thesis addresses several aspects of this quest: the aggregation of knowledge from heterogeneous sources; the construction of large-scale datasets of sequences; the recovery of critical non-obvious information (descriptive metadata); the use of metadata in comparative analyses and meta-analyses; the need for knowledge representation capable of transferring knowledge across multiple tasks; the use of information theory for efficiently-computable metrics; the identification of useful information in free text.

2.3 Evolving Bioinformatics Scalability

The research questions of the post-genome era will require the study of large datasets consisting of whole genomes, sampled across large populations of individuals; the combination of increasingly diverse data types of knowledge, extending beyond plain nucleotide sequences, including comprehensive descriptive metadata to support decision-making; and the harnessing of multiple methods and algorithms in complex analysis pipelines and meta-analyses. At present, bioinformatics analysis tools only partially meet these requirements: with the exception of a handful of research organization with the technology capability to implement large-scale bespoke systems, life scientists are limited to performing small-scale analyses. Such "bioinformatics in-the-small" may be characterized as being within the comprehension of one person and focused on a single aspect or components; "bioinformatics in-the-large", one the other hand, spans component boundaries and tends involve multiple participants (Parker *et al*. 2003). Such a scale shift requires changes at several levels, including technology, information handling, and working practices, amongst others. The following are examples of typical needs: the manual collection and curation of datasets, typical of small-scale studies, is not viable in studies involving thousands of records; in analysis pipelines, derived knowledge must be transferred from one analysis tool to the next, rather than each stage being an end to itself; analysis tools must be able to support decision-making by utilizing metadata, rather than obliging the user to create multiple version

of the input data prior to analysis. We have identified three dimensions pertinent to bioinformatics scalability:

- ***Integrative scalability***, the ability to use knowledge of different types, from different sources, in different forms, and integrate it into the analysis process;
- ***Quantitative scalability***, the ability to analyze large volumes of data in ways that are efficient, fast, affordable, and accessible to researchers;
- ***Hierarchical scalability***, the ability to organize the discovery process in multiple analysis stages, driven by the researcher's knowledge and intentions, transmitting the knowledge produced by a task as input information to other downstream tasks.

In this section, we have considered each of these dimensions, and discussed the current challenges faced by researchers as they attempt to scale up bioinformatics analysis tasks. Rather than highlighting the issues faced by large change-driving organizations that benefit from extensive IT support, we have focussed on the needs of the vast majority of biological researchers needing bioinformatics analysis, who have neither an extensive IT infrastructure, nor sufficient programming knowledge to create customized applications.

2.3.1 Biological Knowledge Aggregation

Knowledge Aggregation, a core concept of this thesis, has been defined as "the problem of taking information from multiple heterogeneous sources and aggregating it into a unified knowledge base" (Zeng and Fikes 2005). Here, the term *knowledge base* should be interpreted as encompassing any structured aggregated dataset that can be queried or analyzed. The purpose of knowledge aggregation is to extend the usefulness of existing information by increasing its volume, or its dimensionality. The need for biological knowledge aggregation is the driving factor for integrative scalability of bioinformatics. Aggregation of multiple sources may take different forms, such as:

- Sources contain essentially the same data types, but different sets of records (e.g. patients in different hospitals)

- Sources contain conceptually similar data types, but with different structure and scope (e.g. a gene sequence may have structural annotation in one database, with only a single literature reference, while another may detail splicing sites and a comprehensive set of literature reference s)
- Sources contain different data types, but complementary records (e.g. a database of patient's medical data and a database of patient's accounting records)
- Sources contain different data types, but cross-referenced records (e.g. a database of patients and a database of illnesses)

In general, each data source can be assumed to maintain its own record structure, which is consistent for all its records, but from that of other sources, as discussed in the following section. A knowledge aggregation task will define some target structure (*schema*), and map extracted information from the diverse sources to this target schema. In most cases, source data must be transformed structurally, selecting only information of interest; records from multiple databases may need to be merged when they represent the same entity (for example, the same sequence, or patient); conflicting values from multiple source records may need to be reconciled; and data imputation (i.e. "filling in" gaps in aggregated data) may be required when complementary databases do not have equal coverage.

2.3.2 Challenges in Integrative Scalability

Integrative scalability depends of the ability to obtain, interpret, aggregate and use information from multiple sources. Bioinformatics databases number in thousands, and present little homogeneity: they encode different types of data, for different purposes, with different levels of details, using different structures, accessible using different mechanisms. These differences (*heterogeneities*) make large-scale data gathering and preparation insurmountable obstacles for many life scientists, who often construct datasets by cut-and-paste extraction from a database's Web interface to a spreadsheet. Issues related to *database integration* affect many domains of computing, and are widely studied in computer science. Heterogeneities can be classified in four categories (Sheth 1999; Ouksel and Sheth 1999):

- ***System heterogeneity*** refers to differences in mechanisms of storage, query and retrieval provided by databases, and affects how users obtain the right information;
- ***Structural heterogeneity*** refers to differences in internal organization of information records within databases, and affects how users find values within records;
- ***Syntactic heterogeneity*** refers to differences in the encoding of data fields within information records, and affects how users extract desired values from records;
- ***Semantic heterogeneity*** refers to differences in purpose and meaning of data fields within information records, and affects how users interpret extracted values.

Structural, syntactic and semantic heterogeneities involve the content of information records, and can therefore be considered *information heterogeneities*, in contrast to system heterogeneities which involve the management and access of information records, without considering their content. Each of the heterogeneities listed presents unique challenges, which may in turn require specific solutions.

2.3.2.1 System heterogeneity

Biological data sources may be accessed either locally (a copy of the whole database can be installed on a local computer), or remotely (there is only one copy of the data, and it is accessible over a network, usually the Internet). Local copies of databases, such as are available for GenBank or UniProt, allow fast access to the data and do not require implementing network protocols. However, their large size, lack of built-in querying facilities, and need for regular updates discourage their user unless a sophisticated computing infrastructure is at hand. Remote databases typically provide query and retrieval facilities, but may present access difficulties when implemented using different operating systems, programming languages, and database management systems (DBMS). Most current repository implementations use relational database management systems (RDBMS), which support the powerful SQL standard database query language (Khandheria and Garner 2007). However, SQL is too powerful and too low-level to be used as a robust query mechanism on a

public network. Leveraging on the experience of eCommerce and other computing domains, the vast majority of biological databases have adopted Web technologies as the standard interface mechanism (Figure 2-3, see Newsome *et al.* (1997) for an example), as this allows users to interact with the databases using standard Web browsers that implement the HTTP protocol, without the need to install specific client software (Markowitz *et al.* 1997). The Web-based application mediates between the user and the database by providing a query form for specifying search criteria. These criteria are translated to database queries (typically using SQL), and the results are converted into user-readable HTML pages that can be viewed by the user (Garcia-Remesal *et al.* 2004). There are countless variations on this architecture: some databases provide persistent management of query results, other integrate relevant analysis tools, which can be controlled via query interfaces. In addition, some databases return results in formats other than HTML: proprietary formats such as those used by GenBank or UniProt; standards such as FASTA; or ad-hoc formats such as comma-separated values (CSV).

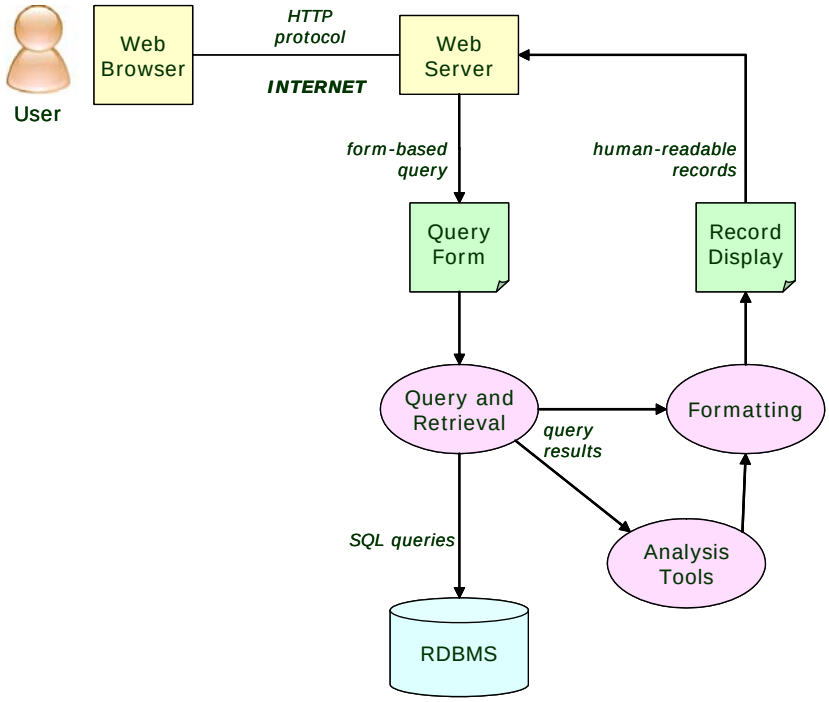


Figure 2-3: Basic structure of a typical Web-accessible biological database
The simplest data repositories do not provide analysis tools. More sophisticated systems, however, may provide result management, and return results in a variety of formats.

The use of a Web infrastructure hides difference in operating system and programming

language, but this is not sufficient to overcome system heterogeneities. Most repositories implement this common architecture differently: form fields and selectable query values are different for different databases, as are the HTML layouts used to display results, and form submissions mechanisms may also vary, sometimes utilizing scripts and hidden parameters. In addition, the dynamics of query execution differ between databases: some execute queries synchronously, while others may use queues to process results; error handling behaviour may vary; and some databases may require subscriptions to access data. In most cases, applications are only designed for interactive users: they use human-readable forms, and produce human-readable results, which are obstacles to automated data collection and aggregation.

2.3.2.2 Database Integration Approaches

Several technology approaches to biological database integration have been proposed (Wong 2002; Hernandez and Kambhampati 2004), most of which have also emerged in other domains of computing. These solutions usually address both system heterogeneities and information heterogeneities (as will be discussed in later section), and can therefore be viewed as architectures for integrative scalability. **Federated databases** (Heimbigner and McLeod 1985; Sheth and Larson 1990) are meta-databases: a *common data model* (CDM) is defined, which integrates the schemas of the participating databases. The central federation system decomposes queries as appropriate, and delivers them to the component data sources, mapping query results to the CDM. Thus, a federation system such as Kleisli (Chung and Wong 1999; Wong 2000) requires information about individual data sources to be used to form queries. Typically, database federations maintain common control metadata, but not a centralized database, and thus data is retrieved from its original sources at every query. Some authors distinguish between various degrees of federation, characterized by their architectural integration (Karasavvas *et al.* 2004), which may be loose- for example, NCBI Entrez (http://www.ncbi.nlm.nih.gov/Entrez/), consist of a Web portal with integrated search and cross-referencing, and does not attempt to combine source data schemas. **Data warehouses** (Schönbach *et al.* 2000; Widom 1995) are databases constructed and maintained from data

retrieved from multiple sources (typically primary databases). Warehouses replicate the source data, mapping it to fit to a centralized CDM (the warehouse's data schema), and managing it independently of its original sources. Differently from database federations, warehouses need not query data sources synchronously, and therefore there is no requirement to translate submitted queries. Furthermore, local storage of data promotes higher performance (Wong 2002). Many secondary biological databases are data warehouses, which collect publicly available information on a given theme, annotating it with metadata and additional knowledge from various sources. There is considerable flexibility in data warehouses design, and toolkits are available to support their creation from data in the main primary data repositories (Koh *et al.* 2004; Lee *et al.* 2006). Even with such support, the design and implementation of data warehouses require a high level of computing skills, and there is limited flexibility to accommodate design changes due to new requirements, such as the support of additional databases, or additional fields, at a later stage. **Mediators** (Wiederhold 1992) are a strategy for improving the flexibility of database federations. Mediators accept user queries in a *generic query language* (GQL), and access data sources by means of components known as *wrappers*. Database wrappers perform all necessary translations: they convert generic queries into database-specific queries; they manage the interaction dynamics with the source database; and they transform results to the desired form. Mediators do not create centralized replicas of the aggregated results, and do not require architectural integration with the original data sources. One of the key advantages of mediator systems is their inherent extensibility: in principle, wrappers for new data sources can be added without modification to the overall architectures. However, mediators do not usually provide local storage and therefore, like federated databases, may suffer from poorer performance.

A comparison of the described approaches is shown in Table 2-3- fundamentally, these approaches offer different trade-offs. Federated database offer highly powerful, specific querying and superior reliability, at high organizational cost. For users with limited IT infrastructure and know-how, a mediator architecture appears to be a clear choice.

	Federated Databases	Data Warehouses	Mediators
Query Expressiveness	Very powerful because of detailed built-in knowledge of sources	Powerful, but limited to the warehouse structure	Limited to translations of generic queries
Query Performance	Can be poor if the federated databases are distributed	Can be very fast	Limited by the performance of remote systems
Control over data schema	Full control and reliability, at high organizational cost	Full control	No central schema, depends on wrappers for translation
Schema Flexibility	Very expensive to change schemas	Schemas can be changed by administrators	Highly flexible, may be changed by end users
Extensibility	Requires engineering effort to add new sources	May be extensible, depending on architecture	Generally very extensible, with plug-in wrappers
Infrastructure requirements	Requires IT infrastructure and organization	Requires servers, RDBMS, admin privileges	Suitable for low-resource environments, peer-to-peer networks
IT Expertise requirements	Can only be implemented by IT experts	Generally manage by IT administrators	Most suitable for end-user management

Table 2-3: Comparison of Database Integration Approaches

2.3.2.3 Structural heterogeneity

When information from multiple sources is aggregated, there may be significant structural differences between schemas (data record organizations) used by the source databases. In a database, schema design is strongly dependent on the purpose and perspective for which the database is designed. For example, designers of a small warehouse of DNA sequences have to choose what type of annotations are to be supported, among the many possible (open reading frames, introns and exons, protein products, promoter sites, polymorphisms, function, genomic location, *etc.*). Since the support of annotations requires laborious curation and maintenance, only a subset of annotation sources will typically be selected. Structural heterogeneities arise even when the same fields are supported from multiple sources, since most complex information can be modeled in multiple ways, and therefore encoded under

different structures. To complicate matters further, data can be represented using different data structuring paradigms (Table 2-4). Besides the strictly *structured* information in RDBMS databases and the *unstructured* data found in text, the Web has popularized *semi-structured data*, such as that encoded in XML documents and Web pages. Semi-structured documents are structured internally, but linked to each other at coarse granularity level- each document can therefore be thought as an interlinked "micro-database".

Data Structuring	Characteristics	Examples
Unstructured	Document-based Minimal internal structure (or none)	Biomedical free text abstracts NMR scan images
Structured	Entities with attributes Relationships between entities Schema and metadata define structure	RDBMS tables Spreadsheets
Semi-structured	Document-based Has an internal structure, often self-describing Relationships between documents	XML files GenBank records

Table 2-4: Data Structuring Paradigms: Structured, Unstructured and Semi-structured

Several approaches have addressed structural heterogeneities, and we will discuss briefly some of the most important ones. First, the problem can be tackled at source, if source databases exchange data records using agreed **data standards**. Although it is unlikely that any two databases use the same internal structure, standardization bodies and industrial committees can define formats for encoding data to be interchanged between databases, simplifying integration tasks. For example, a data warehouse that aggregates information from multiple databases will be able to reuse rules for transforming results to its internal schema, if all databases use the same standard. At the same time, source databases can retain their own internal data structures, invisible to the client warehouses, and mapping query results into a standard-complying format at query time. Semi-structured documents are well suited to such standards, since they produce structured snapshots of query results, which are self-contained and do not require the inclusion of other data. The Extensible Markup Language (XML) is a suitable technology for defining biological data exchange standards (Achard *et al.* 2001), because of its semi-structured nature, and its wide support on the Web

platform. XML is highly versatile, allowing the definition of markup tags that are problem-specific (DNA sequences, taxonomy, protein interactions, *etc.*) but can be parsed and manipulated by generic off-the-shelf XML software. The abundance of ongoing XML-based standards efforts (Brazma *et al.* 2006; Strömbäck *et al.* 2007) reflects both the success of XML and the reluctance of the bioinformatics community to adopt universal formats. The multiplicity of exchange formats forces aggregators to implement multiple transformations, undermining the utility of standards. Furthermore, weakly supported formats are unlikely to be maintained over time, as demonstrated by abandoned proposed standards, such as AGAVE (http://www.agavexml.org/). Low standard adoption may be due to limited technological know-how, to the evolving needs of the field, and to the fact that standards may appear overly complicated or oversimplified for specific purposes (Brazma *et al.* 2006) Paradoxically, the most popular *de facto* standard for sequence interchange is FASTA, which is trivially simple and supports no metadata, but is easy to process. In spite of these problems, interchange formats remain important for information aggregation, because they insulate aggregators from the internal structures of source databases. Even in the absence of standards, semi-structured documents encoded in XML (or in the native formats used by GenBank or UniProt) are an excellent choice as input for database wrappers.

After results are retrieved for aggregation, a **schema mapping** is usually applied, to adapt retrieved records to the target structure (e.g. a data warehouse's schema). The mapping task is typically performed within a database wrapper. Schema mappings are necessary even when source and target schemas represent the identical data types, since relationships between fields may be differently (Figure 2-4). Many problems need to be addressed when designing schema mappings, such as field hierarchy restructure, aggregation of data from multiple tables, splitting of records in small-granularity objects, interpretation of source schemas semantics, finding correspondences between entities. Härder *et al.* (1999) presented a thorough review of these and other important issues in schema mapping. Several tools support schema mappings using different technical approaches, including graphical mapping interfaces (Hernández *et al.* 2001), automated schema analysis (Castano and De Antonellis

1999), and machine learning (Doan *et al.* 2001). XML has strong standard support for schema mapping, consisting of XPath (Clark and DeRose 1999), a query language for extracting information using structural paths, and XSLT (Clark 1999), a transformation language for defining rules for structural conversion.

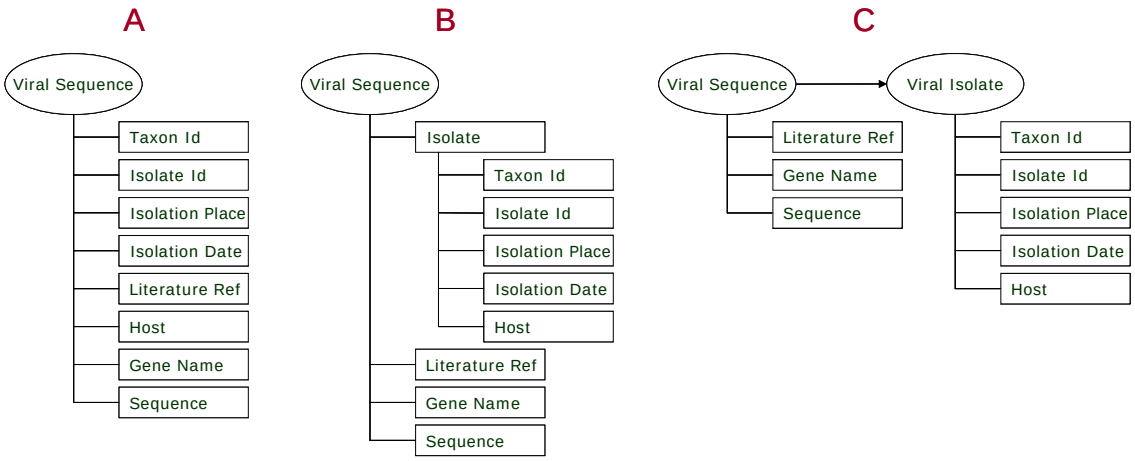


Figure 2-4: Three examples of different data structures encoding the same information. (A) All annotations are individual fields in the same sequence record; multiple sequences from the same isolate will need to duplicate annotations such as Isolate Id. (B) The same information can be encoded in a more structure fashion, grouping related attributes. This is an example of *semi-structured* record. (C) Information on the experimental sample may constitute a separate type of record, which is referenced by the sequence record. This *structured model* minimizes value duplication.

Automated schema mapping is a common feature of aggregator systems, especially in data warehouses and systems with integrated schemas, but it is not a requirement. An alternative approach is represented by **non-transparent** mediators, which retain the original record structure (Karasavvas *et al.* 2004) and have no built-in transformation rules. The user must therefore explicitly specify restructuring rules, making this approach well suited to aggregation systems that allow *ad hoc* selection and transformation of fields, typically via an interactive interface.

An alternative to schema mapping is to "deconstruct" the input document, converting it to a simply-structured **knowledge representation** format which is still capable of expressing the structure of the input information. A suitable emerging knowledge representation standard

is the Resource Description Framework (RDF) (Beckett 2004), a technology developed to support the Semantic Web. RDF and other semantic technologies are discussed in detail in Chapter 4, where they are proposed as a knowledge representation platform to support bioinformatics knowledge pipelines. In short, RDF uses an extremely simple "triple" structure ("subject-predicate-object") to express statement ("facts") which collectively convey equivalent information to that contained in semi-structured documents. RDF only uses the "triple" structure, so all RDF data is identically structured, making aggregation a trivial task. In addition, RDF data is easily extendable, allowing new data types and attributes to be defined easily, thus overcoming some of the limitations of XML (Wang *et al.* 2005).

The ABK knowledge aggregation platform presented in Chapter 3 of this thesis uses a non-transparent mediator system, in which a user-friendly visualization of the original XML source document allows to specify the specification of *structural rules* (structural transformations expressed in XPath) by direct user interaction. This approach allows the user to extract knowledge without requiring a formal description of the source's data structure. Rules are applied to all documents from the same source, providing users with immediate visual feedback of the rule effectiveness. Thus, extraction rules can be defined *by example* without requirements for programming skills, or in-depth knowledge of the sources, making the aggregation of structurally heterogeneous records accessible to life scientists.

2.3.2.4 Syntactic Heterogeneity

Data field values are extracted from a data source, and they may need to be converted to suit the data format expected by the target schema. Of the different transformations possible, the simplest involve a *lossless conversion* between formats- for example, a time stamp in textual format, such as "August 8, 2008 8:08 pm" may be converted to a Unix integer representation (the number of milliseconds since 1 January 1970), a popular date format in computing. This type of conversion may be supported by programming languages or software libraries, and does not pose particular problems. However, many data conversions are *lossy*, in that some precision may be lost in the conversion process- for instance, a source document may express

an experiment's sampling date using a full date, but the target system may only record the year. Conversely, *expanding precision* of such data requires making some arbitrary decisions (for example, "1984" may be translated to "1 July 1984 00:00" if a full date is required by the target system). In some cases, these conversions are non-trivial, and data loss may be significant- for example, with fields that record geographical locations. For an avian influenza sequence, a database may encode the isolation place as "Qinghai Lake, China", while a different database may only report the country name, or encode the geographical position using coordinates, such as "36º 54' N, 100º 08' E". Conversions between these three different forms require special tools, such as gazetteers or GIS software (Janée *et al.* 2004). In addition, reducing the value to retain the country name alone produces a large precision loss due to the size of China, as can be appreciated when converting the country name to a set of coordinates. Another conversion that requires special processing is the substitution of a literal value with a reference- for example, the values "Cow" and "Bos Taurus" may be translated to "9913", an identifier in the NCBI Taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy/). Such translation may require the use of specialized software, aided by a dictionary and/or an **ontology** (a controlled vocabulary, which may include synonyms), but it can resolve ambiguities, and enable selections and queries that leverage on taxonomic knowledge (e.g. "find all sequences isolated in mammals").

Transformations are further complicated when data is encoded in text form (which is frequent in major databases such as GenBank), since the format of human-readable text is typically unconstrained. Thus, a text-based "isolation date" field may contain values "1998" or "Patient infected in December 1998", and special techniques must be used to extract the year value. Several text processing approaches are available: **ontologies** and **dictionaries** can be used to identify keywords within free text; **regular expressions** can identify text character patterns and specific lexical formats; and **text mining** may extract structured information from text by learning from existing examples (Li *et al.* 2005). Considerable work is being put into text mining techniques, in attempts to recover structured information from the vast quantities of free text available in biomedical literature- see (Hunter and Cohen 2006) for a

recent review of the state-of-the-art.

The ABK approach described in Chapter 3 of this thesis processes information from data sources by applying *text filters* which can use either *regular expressions* to extract information from patterned text, or *user-defined dictionaries* to recognize problem-specific keywords and their synonyms. ABK alerts users about value conflicts to be resolved (for example, if the term "Turkey" is interpreted as a country rather than the host organism, and conflicts with another country name).

2.3.2.5 Semantic Heterogeneity

When discussing schema mapping, we make an implicit assumption that fields in the source and target databases have a direct correspondence, and need only be restructured and transformed. However, there are often *semantic mismatches* between databases: difference in the meaning or interpretation of the fields, arising from differences in purpose, objectives and perspective. Semantic heterogeneity is perhaps the most insidious of all information heterogeneities, since its resolution often requires sophisticated reasoning. It is a widely studied aspect of database integration (Garcia-Solaco *et al.* 1996), whose key aspects have been catalogued by Doan and Halevy (2005) as follows. First, the semantics of data have to be interpreted by humans, usually based on documentation or clues from control metadata, such as field names. Inconsistencies may arise because documentation is incomplete or incorrect, and the designer of the data source is not at hand to verify interpretations. Such heterogeneities may be observed even between records within the same databases, when data submitters do not work with the schema designers. For example, a large-scale aggregation of dengue virus protein records from GenPept revealed that the field "isolation_source" was inconsistently used, containing values as diverse as "Samoa" (record BAC77216), "Homo sapiens" (AAT85667) and "isolated in 1993" (AAN74539). These inconsistencies pose two problems: the data needed must be sought in a different field, while the value extracted may be needed for a different target field. The second aspect of semantic heterogeneity is that semantic clues may be incomplete or ambiguous: for example, in records about bacterial

infections, a field named "organism" may be interpreted as either the infecting pathogen or the infected host. Third, there may be multiple fields in a record that are candidates for the extraction of the target value, and one must consider the goodness of a semantic match of a field relative to other fields. In Chapter 3, we discuss how metadata in GenBank influenza records may be distributed over multiple fields, and encoded in different forms (for example, the year of isolation may be encoded as a number, or as part of an influenza isolate identifier). The final aspect to be considered is that semantic matching is subjective, in that depends on the intention and objectives of the aggregation task, and therefore may require the user to make explicit choices about the handling of source data. An additional important problem, identified by Zeng and Fikes (2005), is the need to decide which data source takes precedence when complementary records are used to reconstruct missing data, which raises the issue of *trust* in data sources. The same problem applies when records from different sources refer to the same entity, but contain conflicting values.

Many solutions have been proposed to address semantic heterogeneity, and they fall in two broad categories: *rule-base* and *learning-based* systems (Doan and Halevy 2005). Rule-based systems require the definition of mapping rules based on schema metadata, which are either configurable or embedded in the application code, and then applied mechanistically to perform transformations. In contrast, learning-based systems use machine learning or statistical methods to derive rules based on the content of data values in a number of example records. The "fuzzy" nature of this matching approach is suited to data that is inconsistently encoded, or where the structure of the source data is very different from that of the target data. Some researchers has underlined the importance of using past matches in assessing the accuracy of matching rules or machine learning algorithms (Do and Rahm 2002).

The ABK platform presented in Chapter 3 of this thesis leverages on XPath rules defined to overcome structural heterogeneity, allowing multiple rules to contribute to the result, to account for encoding inconsistencies. Thus, for any target data field, it may be necessary to define multiple rules for multiple source schemas. These rules can be prioritized by the user, and inconsistencies are identified and highlighted to the user for manual resolution. In

Chapter 3, we analyze the contribution of multiple rules in a real-world metadata extraction task from more than 90,000 records from GenBank and GenPept.

2.3.3 Challenges in Quantitative Scalability

Quantitative scalability issues arise when the increase in data volume demands an increase in computing resources, to the point that changes to the computing infrastructure are necessary. In many cases, periodic hardware and software upgrades provide the necessary computing power upgrades to face new data volumes, and new analysis tasks. However, not all bioinformatics analysis scales linearly. The execution time and memory usage of multiple sequence alignment (MSA) algorithms, for example, can grow dramatically with the number of sequences, and different MSA tools impose tradeoffs, which may reduce accuracy or processing time if data volume is increased without adding computational resources (Edgar and Batzoglou 2006).

A common to quantitative scalability solution, particularly in commercial domains such as e-Commerce, is to increase computational power by upgrading the computing infrastructure. A distinction is often made between *vertical* scalability- extending the power of machines by adding processors or memory- and *horizontal* scalability- distributing computing workload over a number of machines (Michael *et al.* 2007). While a diverse array of approaches is available from hardware and software vendors, bioinformatics studies generally favour horizontally-scalable approaches which parallelize the execution of tasks on multiple machines (Henschel and Muller 2007). The benefits of scaling across multiple processors can only be reaped if analysis tasks can be split into multiple subtasks that can be executed in parallel (massive parallelism), which fortunately is a characteristic of many important bioinformatics tasks, such as MSA. Although parallel processing is commonly performed by computer clusters, an emerging trend is to use a grid computing infrastructure, which loosely federates computing resources, distributing the computational load and aggregating results (Andrade *et al.* 2006; Carvalho *et al.* 2005). Some massive-scale grid architectures, such as Folding@home (Shirts and Pande 2000; Zagrovic *et al.* 2002), even

allow volunteers across the internet to aggregate their desktop computers, providing spare computing power during idle periods. Major downsides of grid computing are the paucity of grid-enabled analysis tools, and the high level of technical know-how required to set up and manage computing grids. Although grid-enabled versions of common tools (*e.g.* Trombetti *et al.* 2007) may gradually become available, most life scientist may be some time away from benefitting from these technologies.

Quantitative scalability of bioinformatics analysis is clearly important to biologists that cannot count on a large IT support infrastructure, and thus rely on desktop computer hardware or departmental servers or their analysis. Choosing an approach with low-order computational complexity (i.e. whose time and memory requirements grow moderately as data increases) is therefore important. Some tasks may be supported by a variety of tools with different computational complexity characteristics (Edgar and Batzoglou 2006). For analysis tools with higher order complexity, it may still be possible to adopt a divide-and-conquer strategy, by splitting the data into smaller sets that can be analyzed separately, sand subsequently merging the results. Identifying suitable algorithms and strategies for all possible types of bioinformatics analysis is outside the scope of this work. In Chapter 5 of this thesis, we have presented a number of techniques for diversity analysis and for comparative studies of multiple sequence alignments, implemented on a desktop platform. Because of their statistical nature, these information-theoretical approaches scale very well, with computational requirements growing linearly with the number of analyzed sequences. The semantic data restructuring described in Chapter 4, on the other hand, illustrates a task with high-order complexity (semantic reasoning) which required a divide-and-conquer approach to be executable within a realistic timeframe.

Finally, quantitative scalability applies to growth in the number of records, and also in the number of *dimensions* of the data. We distinguish therefore between the number of *instances* to be analyzed, and the number of *features* of these instances that need to be included in the analysis. Data mining approaches have been particularly effective at combining many variables into pattern discovery, and are therefore suitable approaches for

high-dimensional data.

In Chapter 8 of this thesis, we present a text mining tool which is generic, since it does not embed domain-specific or problem-specific knowledge. To retain this generality, all words present in the text have to be treated as separate features, and therefore instances may contain thousands of features that need to be considered simultaneously. We found the machine learning algorithms used to be very efficient at handling such large sets of attributes, and note that specific claims in this sense have been made for support vector machines (SVM) (Joachims 1998)

2.3.4 Challenges in Hierarchical Scalability

2.3.4.1 Constructing Bioinformatics Workflows

To address more complex research questions, complex analysis processes are required. Such analysis processes are made of multiple analysis tasks, working in concert to produce a final result, which can be organized in a variety of patterns. Task execution is *parallel* when tasks are independent, and their results are combined in a meta-analysis; for example, the protein conservation analysis described in Chapter 7 of this thesis identifies conserved regions in a number of co-circulating lineages separately, and then selects regions which are found in each of the lineages. *Serial* task execution occurs when one task analyzes the results of a previous task, and therefore the two tasks are executed in sequence; for example, the knowledge aggregation task that produces a dataset of all influenza sequences is executed before the alignment task, which is then followed by conservation analysis, and so on. Tasks may also be *recursive*, in that they may use their own results as an input; the active learning text mining task presented in Chapter 8, for example, uses prior classified results in order to improve subsequent classification. “Second-generation” bioinformatics studies may require complex *workflows* that apply several of these patterns, raising a number of challenges: first, how to organize workflows so they can be specified and constructed with maximum flexibility? Second, how to make information and knowledge flow easily from one task to the next?

Third, how to make these tasks controllable and easily understood by biological researchers, allowing them to inject their choices and knowledge into the workflow?

The construction of complex processes from smaller computational tasks is a widely studied aspect of system integration, a branch of computer engineering. Workflow is an important component of business computing, particularly in enterprises that need to use heterogeneous systems across multiple organizational units. The Service Oriented Architecture (SOA) is the approach of choice for organizing such workflows. SOA describes processes in terms of task-oriented components called *services*, which are loosely integrated, and invoked by applications that execute processes through service orchestration (see Erl 2005, for an introduction). SOA services are most commonly implemented as *Web Services*, software components that can be accessed using the Web infrastructure (HTTP and XML), and are invoked using standard protocols. Researchers have constructed specialized web services that encapsulate bioinformatics analysis tools (Neerincx and Leunissen 2005), and there is growing focus in the integration of these services into workflows (Romano *et al.* 2005; Garcia Castro *et al.* 2005). Recently introduced end-user tools such as Taverna (Oinn *et al.* 2004) support the design and control of service-based workflows through powerful graphical interfaces. However, this field is young, and significant problems still stand in the way of universal adoption.

First, constructing a workflow with a tool such as Taverna is a complex task for biologists with limited computational skills, because of the intricate data-oriented "wiring" between services, which demands substantial technical background. However, such early implementations are making important contributions, and it is likely that more intuitive paradigms will emerge as the field matures.

Second, the selection of a suitable task, and of suitable parameters, requires in-depth understanding of the algorithms and tools implemented by services. This problem is not unique to Web Services: similar choices must be faced when using today's standalone analysis tools, and many researchers routinely use services such as BLAST (Altschul *et al.* 1990) without understanding the significance of their parameters, relying on defaults to suit

their problem. Although the Web Services community has developed standard for Web Services description (WSDL; http://www.w3.org/TR/wsdl) and for their discovery and integration (UDDI; http://www.uddi.org/pubs/uddi_v3.htm), these standards describe services at a very low level and are therefore only suitable for programmers that need to integrate service calls in their software.

The third problem is that data flowing from one service to the next may require transformations. If one thinks of bioinformatics tasks as biological data sources, and the results of task execution as equivalent to query results, then previously discussed issues of information heterogeneity apply. Currently, many bioinformatics tools store results in non-standards formats (Wiley and Michaels 2004), and therefore the lack of standards for biological knowledge representation affects all stages in the knowledge discovery pipeline. Not surprisingly, some ongoing work in web service integration uses technologies that aim to solve information heterogeneities, such as semantic technologies based on XML, RDF and ontologies. The BioMoby project (Wilkinson and Links 2002) has produced an ontology of bioinformatics data types, able to describe the inputs and outputs of bioinformatics web services at a higher level than is possible using WSDL and UDDI, and supported by code libraries. Beyond integration with workflow tools, the ambitious long-term aim of BioMoby is to standardize *semantic web services* (McIlraith *et al.* 2001), with descriptions in machine-understandable form, so that *reasoning software* will be able to select and invoke appropriate services, flowing results between them, without requiring user intervention. At this time, however, we can only view this effort as a promising future direction. In Chapter 4 of this thesis, we have discussed the role of semantic technologies in encoding and transferring knowledge, and present an example of how bioinformatics tasks can leverage on the reasoning capabilities provided by these technologies.

2.3.4.2 Integrating the User

Bioinformatics has been described as a discipline where a cultural gap has formed between three cultures- those of biologists, computer scientists, and engineers, mainly due to

differences in their vocabularies, perceptions of requirements, and the scarce appreciation of the mutual efforts needed to understand each other's domain (Benton 1996). Over a decade after this assessment, there has been only limited progress on this front (Kumar and Dudley 2007). Although bioinformatics is a multi-disciplinary field, there is little doubt that biomedical research is its *raison d'être*, and that biologists who need computational analysis will continue to outnumber bioinformaticians. Thus, a growing number of life scientists are tackling the analysis of large datasets, often equipped with little more than a standard desktop personal computer, and limited programming knowledge. It should come as no surprise that these users, may be uncomfortable with the programming libraries and command-line tools favoured by computer scientists. They typically prefer user-friendly graphical or Web-based tools with low entry barriers (Kumar *et al.* 2008). Although these users lack IT expertise, they possess extensive domain knowledge, and are often better equipped than computer scientists or engineers for specifying and controlling computational pipelines. Thus, bioinformaticians should strive to provide life scientists with easy-to-use building blocks for constructing analysis processes, and intuitive mechanisms for injecting the user's own knowledge to control analysis tasks.

To our knowledge, no comprehensive study of suitability of interaction mechanisms has been conducted specifically targeting biologists. Although such study is beyond the scope of this work, some important factors been identified in the course of this thesis. First of all, it is desirable to adopt user interface mechanisms that are familiar to biologists in their day-to-day usage of computers. For example, most biologists employ spreadsheet tools such as Microsoft Excel to manage small-scale experimental data, and are thus familiar with editable tabular interfaces and drag-and-drop mechanisms. Second, input data formats using plain text (such as comma-separated values, or FASTA), are preferred to more structured data formats (such as XML documents). Third, life scientists are able to read structurally complex output (such as records displayed by Web-based databases such as UniProt) but have less tolerance for "technical" formats intended for machine processing, such as native or XML formats, as they may perceive structure and metadata to "swamp" interesting data (Figure 2-5).

In the work presented this thesis, special emphasis was placed on user interaction mechanisms that would be acceptable to life scientists, allowing them to control the analysis tasks without programming. The ABK knowledge aggregation software, presented in Chapter 3, does not present XML structured data in its "tagged" form (Figure 2-5C). Rather, data is presented using a specially-designed user interface component that displays a hierarchy of name/value pairs, offering two advantages: its layout separates control metadata from data, and allows the user to define extraction rules by interacting directly with the desired data. In addition, the dictionaries used by ABK value filters to overcome syntactic heterogeneity are encoded in plain text format, and can be easily extended and customized by the user. The results of the knowledge aggregation tasks are presented in a familiar tabular form, similar to that of common spreadsheets, from which the user can cut-and-paste the data. ABK is also capable of exporting the aggregated dataset in plain text CSV format (human readable), or in RDF format (machine readable). The AVANA tool, presented in Section 5.3, accepts metadata in plain text CSV format, allowing the user to construct ad-hoc metadata files using a spreadsheet application such as Excel. The metadata selection is intuitive, performed by the user by selecting values in list boxes. Finally, the RATMAT text mining tool presented in Chapter 8 allows simple user customization of the text mining task: keyword classes can be created from lists of words and regular expressions, and the expert user can prune the list of features (classification terms) at classification time. These features, which will be presented in detail in the relevant chapters, reflect the focus of the present work on the production of bioinformatics tools that are not only powerful, but also intuitive for biologists to use.

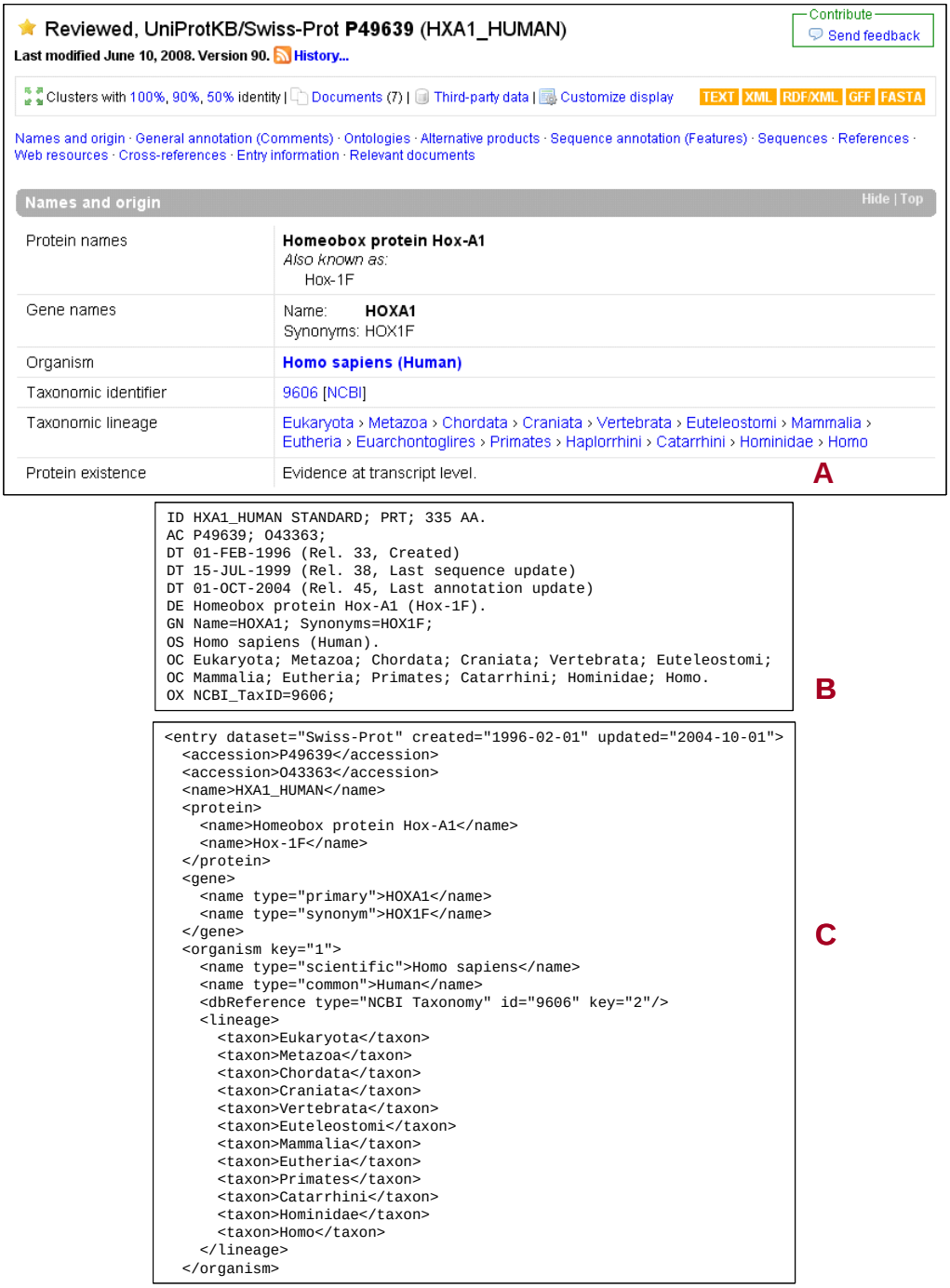


Figure 2-5: Different representations of the UniProt record P49639
The same information is shown (A) as seen when browsing the EBI UniProt web site; (B) in the Swiss-Prot native format; and (C) in the Swiss-Prot XML format.

2.4 Towards Biological Knowledge Mining

2.4.1 What is Knowledge Mining?

Second-generation bioinformatics will be driven by user knowledge, and will have to scale in

multiple dimensions, in order to fully utilize the vast quantities of diverse knowledge that are becoming available in the post-genome era. The predominant concept in *knowledge discovery* is *data mining*- the discovery of significant patterns in large volumes of data. Underlying this concept is the assumption that patterns solely result from combinatorial properties of the underlying data, and patterns will thus emerge from the analysis of data attributes. However, biological data is rarely viewed from a single perspective, and the same data may be analyzed for different purposes, yielding different kinds of knowledge. In addition, the concept of data mining appears limited when modeling processes made up of multiple, cascaded tasks.

Recent work by Ryszard S. Michalski, a founder of the field of machine learning within computer science, addresses these concerns, laying the foundation for a new direction that he named *knowledge mining* (Michalski 2003). Michalski described data mining tasks as computationally complex in their pattern analysis operations, but conceptually simple in how they use the derived knowledge. Since patterns must not only be discernible, but also relevant to the analysis task in hand, certain limitations in data mining must be overcome:

- the user must be allowed to express *analysis goals* which must be understood by the analysis system, and are used to drive the knowledge mining tasks;
- *background knowledge* should be injected into the analysis task, where it is used to extract new knowledge from data, or refine existing knowledge;
- pattern discovery should be *incremental*

Thus, the "traditional" data mining paradigm:

$$\text{DATA} \rightarrow \text{PATTERN}$$

is transformed to a knowledge mining paradigm:

$$\text{DATA} + \text{PRIOR\_KNOWLEDGE} + \text{GOAL} \rightarrow \text{NEW\_KNOWLEDGE}$$

In this mapping, new knowledge from one stage of the analysis can be transferred (perhaps selectively) as prior knowledge to subsequent stages, thus supporting the incremental nature of the knowledge mining paradigm.

In Michalski's vision, users will not use data mining tools directly. Rather, knowledge

mining systems will be able to *understand* (in an artificial intelligence sense) user's goals expressed in a suitable form, and automatically select a number of data mining approaches, apply them, and then reason over the results. The meta-analysis of the results, driven by the user's goal, will then determine the next stage of analysis and this process will be repeated in subsequent stages (Kaufman and Michalski 2005). Although Michalski's group produced prototypes of *inductive database* systems capable of processing derived knowledge (*theory formation*) (Kaufman and Michalski 2003), we are still a long way from practical and generic implementations of the knowledge mining paradigm. A number of gaps to be bridged were identified (Michalski 2005):

- a suitable method for goal representation, intuitive enough for the user to define, but precise enough to be used to drive the analysis process;
- suitable knowledge representation and management (such as a knowledge base) to allow reuse of discovered knowledge from multiple tasks in downstream analysis;
- an intelligent multi-strategy data mining system that can select and invoke numerous data mining analysis tools; and
- a suitable knowledge visualization tools for the user to interact with the output of the knowledge mining process.

2.4.2 Applying Knowledge Mining Principles to Bioinformatics

It is not the focus of this thesis to advance the field of knowledge mining from a computer science perspective. We are, however, interested in the close match between the objectives of knowledge mining as defined by Michalski, and those of second-generation bioinformatics, in particular with respect to hierarchical scalability needs. The identified knowledge mining gaps are very similar to gaps identified in post-genomic bioinformatics: the need to capture user goals is related to integration of user's knowledge discussed in Section 2.3.4.2, while the issue of knowledge representation and management is identified in Section 2.3.4.1. Thus, knowledge mining can be used to describe second-generation bioinformatics at least

conceptually. We have found no evidence of previous attempts to do so and to our knowledge this is the first attempt to use knowledge mining for definition of second-generation bioinformatics concepts.

In this thesis, we have defined **biological knowledge mining**, the application of knowledge mining to biological data; we have used knowledge mining concepts to *model* large-scale bioinformatics tasks. In other words, we have used knowledge mining to describe second-generation bioinformatics tasks and map the knowledge flow; rather than implement fully automated systems. Such a project would be ambitious and unrealistic, given the current state of technology. We will use a simple modeling notation which captures the knowledge mining paradigm (DATA + PRIOR_KNOWLEDGE + GOAL → NEW_KNOWLEDGE) as shown in Figure 2-6.

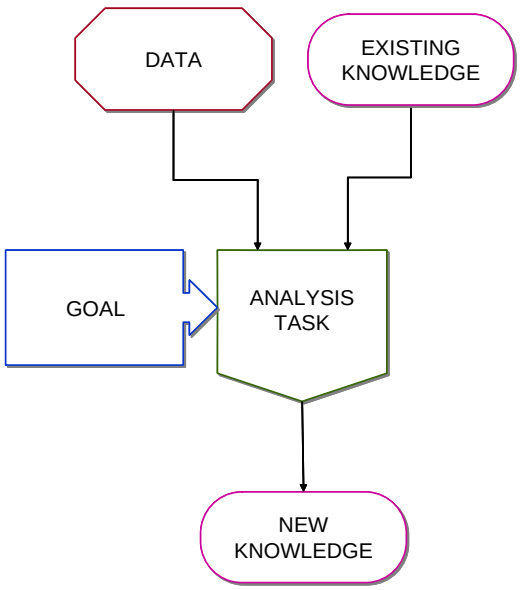


Figure 2-6: Graphical notation for describing biological knowledge mining components

This notation will be used to describe components of bioinformatics workflows. In the case studies presented in this work, we have found that the notation is suitable for both high-level views (in which the components are large-granularity workflows) and low-level views

(in which the components are single data mining tasks). To compose multiple tasks into knowledge mining workflows, we have connected components so that the new knowledge generated by a task becomes the existing knowledge that affects the next. This modeling notation supports the cascading of tasks to analyze multiple types of data; or to refine the analysis of one type of data (for example, as a meta-analysis); and the recursion of tasks in which derived knowledge is injected to refine the result (Figure 2-7). Each of the real-world knowledge mining tasks presented in Chapter 3 (aggregation of large-scale viral sequence datasets), Chapter 6 (evaluation of human-to-human transmissibility of avian influenza viruses), Chapter 7 (identification of potential epitope-based vaccine targets) and Chapter 8 (active learning text mining of biomedical abstracts) makes use of multiple tasks, arranged according to these patterns. The specific details are shown in the individual chapters. In each task, the produced models have helped us to identify data mining components, and methods for injecting user's knowledge and intentions. In addition, the models describe the knowledge flow, and therefore highlight knowledge representation requirements in the workflows. The models were found to be very versatile formalizations of the knowledge mining process, and we believe they can be applied to most knowledge mining tasks.

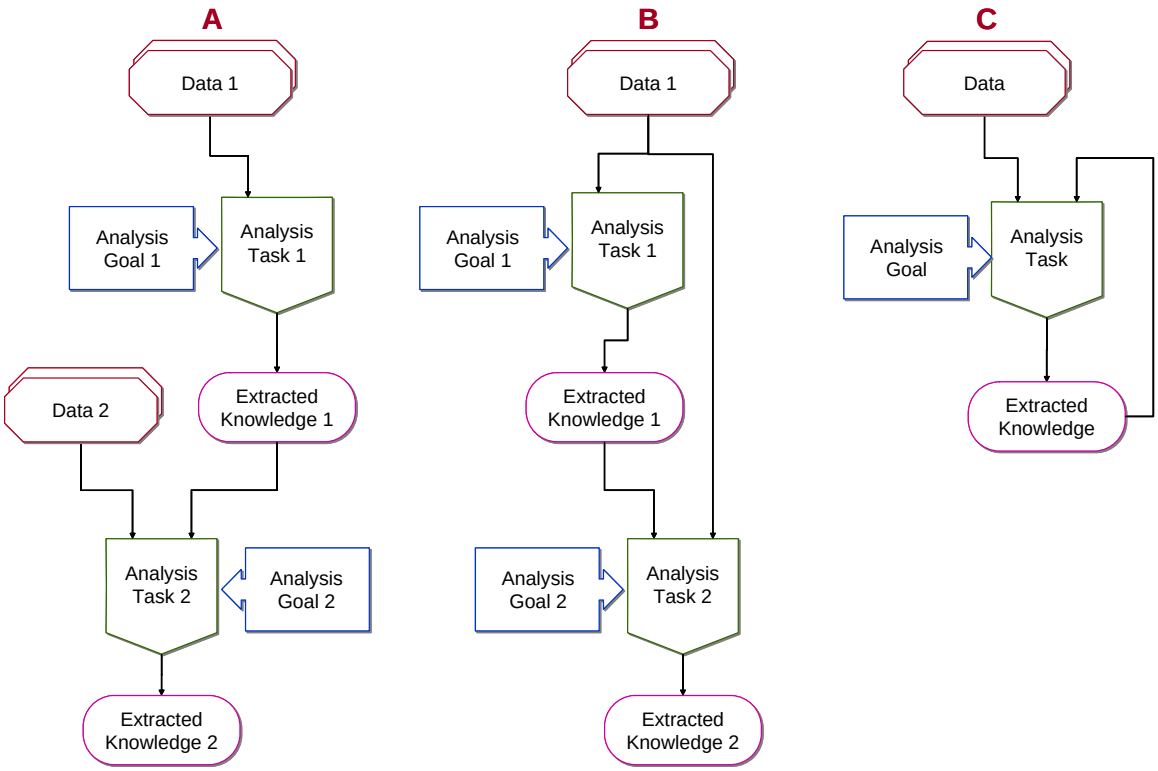


Figure 2-7: Three modelling patterns for biological knowledge discovery pipelines. In these three examples, knowledge flows between analysis tasks: (A) from upstream to cascaded task, used to analyze new data; (B) from upstream to cascaded task, used to refine the same data; and (C) feedback through the analysis task, to refine the extracted knowledge.

2.5 Conclusion

We have presented the opportunities and challenges set before bioinformatics practitioners in the current "post-genome" era. An unprecedented growth in volume and availability of biological data, driven by low-cost biotechnologies, presents new opportunities for large-scale computational analyses. As a result, bioinformatics analysis tasks will have to grow in multiple dimensions: to integrate more diverse information sources; to handle a greater number of data records and greater data dimensionality; and to apply an increasing number and variety of algorithms to the knowledge discovery process. This scaling up of bioinformatics tasks brings forward many challenges. Heterogeneities in the database systems and the information they deliver are the major obstacles to the integrative scalability of bioinformatics, and they should be addressed by solutions that combine various standards,

intelligent processing activities (such as data mining), and user input. The high volume of data will pose difficulties for the growing number of biologists who need to use analysis tools, but who lack programming knowledge, and who have access to limited information technology infrastructure. The availability of scalable algorithms and easy-to-use distributed computing will therefore be critical to these users. The construction of complex analysis workflows and their application through multiple data mining tasks present major challenges, in spite of significant standardization efforts around bioinformatics web services. The complexity of these infrastructures is an important factor, as is the lack of interchange formats for piping knowledge from one stage to the next. Semantic technologies, which allow generic representation of knowledge so it can be processed by machines, promise important advances in this area, but are still not developed to their full potential. Finally, an important and pervasive underlying consideration is that bioinformatics will ultimately be used by biomedical researchers, not by computer scientists or engineers. Biologists are sophisticated and extremely knowledgeable users, but they mostly possess very limited IT skills; solutions that do not take these needs into account are doomed to failure.

The concept of *knowledge mining* has been introduced to extend that of data mining: knowledge mining does not rely exclusively on patterns embedded in the data, but also uses prior knowledge and user goals to drive the discovery process. Knowledge mining therefore supports the flow of knowledge across multiple tasks, which makes it a suitable paradigm for modeling "second-generation" bioinformatics processes. A simple modeling notation will be used throughout this thesis to describe knowledge mining task components, capturing the knowledge flow and user goals at each stage.

The remainder of this thesis will develop the theme of second-generation bioinformatics analysis processes. A number of technical strategies are proposed to overcome many of the obstacles to achieving scalability of knowledge mining process. The platform for aggregating biological knowledge (ABK) combines several approaches to overcome multiple heterogeneities: a mediator architecture with database wrappers, structural rules for the extraction of data from XML documents, and rule prioritization. It also supports output using

semantic technologies (RDF and OWL), which facilitates knowledge transfer and supports reasoning- which we will show to an advantage for the curation of large data sets. Finally, we will present a set of information-theoretical methods for the analysis sequence diversity, which scale well to large data volumes, enabling thousands of sequences to be analyzed on standard desktop hardware.

In Chapters 6 to 8 of this thesis we combine these techniques and strategies into knowledge mining workflows, demonstrating that they can be used to answer challenging real-world biomedical research questions. From a large-scale sequence dataset comprising all available influenza A proteins and their metadata, we have applied an analysis method based on mutual information, to identify molecular factors involved in the adaptation of this virus to human-to-human transmissibility. The results of this analysis enabled us to assess the pandemic potential of H5N1 avian influenza viruses. The same dataset was also used to investigate the presence of conserved potentially antigenic peptides in the influenza genome, to be investigated as potential vaccine components. The same task was also applied to a similar dataset of dengue virus sequences, demonstrating the generality of the analysis pipeline. Finally, we have shown that the ABK supports another type of analysis, that of biomedical text mining. A reusable knowledge mining pipeline, built from standard data mining algorithms, was applied to identify relevant documents from the analysis of generic features (text words), easily customizable by users. The results obtained from these case studies are evidence that biological knowledge mining is both viable and useful. The case studies presented in this thesis are therefore important contributions towards the establishment of this field in bioinformatics.

3. RULE-BASED AGGREGATION OF HETEROGENEOUS KNOWLEDGE

In Chapter 2 we presented the conceptual approach for biological knowledge mining (Section 2.4), which supports the design of multi-stage bioinformatics processes, in which knowledge flows across tasks. Most large-scale studies begin with the construction of a dataset which, in a knowledge mining process, typically comprises both raw data and descriptive metadata. Often, datasets must be constructed by aggregating information from multiple data sources. As discussed in Chapter 2, Section 2.3, they heterogeneity of biological data makes knowledge aggregation a highly challenging task. Current approaches at automating this task, reviewed in Chapter 2, require in-depth technical knowledge of the data sources, and/or a high level of specialized IT skills. This forces most biomedical researchers to aggregate data and metadata manually, which is only feasible for relatively small datasets (up to tens or hundreds of records).

In this chapter, we present a novel *knowledge aggregation* approach which enables biomedical researchers with limited IT skills to construct large-scale datasets from multiple sources. This approach overcomes system and information heterogeneities through a combination of innovative techniques and standard technologies, which are made available to end users through intuitive "biologist-friendly" user interface. End users inspect source records, and identify the data they need through point-and-click mechanisms. User selections are translated into *structural rules* that are applied consistently to other records, so that data extraction can be automated from the inspection of a limited number of records even when the source records are inconsistently encoded. Multiple structural rules are prioritized, and their output is filtered so that the aggregated values are in the form needed by the user. We have implemented this knowledge aggregation approach as a desktop tool, called Aggregator of Biological Knowledge (ABK), which is described in this chapter. ABK runs on standard desktop computer hardware, allowing users to handle tens of thousands of input records interactively, through a familiar spreadsheet-like interface.

We have used ABK to apply our knowledge aggregation approach to the construction of a dataset of over 90,000 influenza A proteins records from public databases, described in this chapter. The dataset, which comprised a number of descriptive metadata annotations for each sequence, was constructed in two weeks, included manual verification and imputation of missing data. The manual construction of this dataset, on the other hand, would have been far too laborious and expensive to carry out with a small research team. In this chapter, we measured the extent of heterogeneity in the public database records aggregated in this task, and showed that our knowledge aggregation approach is very effective at recovering metadata that is inconsistently encoded. The influenza A dataset was used in large-scale studies of viral host adaptation (Chapter 6) and of conserved immunogenic sequences (Chapter 7), which revealed important immunological and virological results, demonstrating that our knowledge aggregation approach is a key enabler of bioinformatics discovery.

3.1 Requirements for a Knowledge Aggregation platform

The requirements for a practical Knowledge Aggregation platform must consider scalability requirements (see Section 2.3), as well as key human factors such as availability and usability. The following requirements for ABK were identified:

- **Access mechanism independence**. To address system heterogeneity, the ABK platform should support the capability to connect to a wide variety of databases, irrespective of their data access and retrieval mechanisms (for example, Web-based, SQL-based, etc.).
- **Query mechanism independence**. The ABK platform should transform user-specified queries as appropriate for the target database, without requiring users to know specific query languages.
- **Extensibility**. The range of available databases should be extensible without changes to the core architecture and code of the ABK platform.
- **Data structure independence**. To address structural heterogeneity, ABK should be able to process input data with arbitrary structures, without any in-built knowledge about any

specific database being built into the software.

- **Data syntax independence**. To address syntactic heterogeneity, ABK should provide simple mechanisms for allowing users to transform the extracted knowledge lexically and syntactically into the form they ultimately require. These mechanisms must be customizable by the end user wherever possible.
- **Semantic flexibility**. To address semantic heterogeneity, ABK should not require data field to be interpreted in a consistent fashion. Rather, it should be able to use multiple fields in multiple records as the sources of a result, and resolve any emerging conflicts when aggregating these source values.
- **Data Management**. The ABK platform should support the storage and management of the aggregated knowledge, and the capability to select and export this knowledge as needed by the user.
- **Usability**. ABK should be deployable and manageable by any biological researcher without requiring significant expansion of their technology infrastructure, or any programming knowledge.
- **Scalability**. ABK should be able to handle large datasets, comprising hundreds of thousands of database records.
- **Versatility**. ABK should be expandable to allow the addition of analysis tools for specific purposes. In combination with the data management tools, this capability allows the analysis process to scale hierarchically without exiting the ABK platform.

3.2 Defining a generic, reusable and versatile Knowledge Aggregation approach

To meet the requirements outlined in the previous section, we combined a number of technological approaches: an extensible mediator architecture for querying and retrieval from data sources; a generic XML-based result processing system; and a user-friendly mechanism for specifying data extraction rules. All these approaches strongly support the generality of the ABK system, which is designed independently of any source database, encoding schema,

or analysis requirement.

3.2.1 Mediator framework

Our approach of choice is to use a *non-transparent mediator framework* (Karasavvas *et al.* 2004), capable of reconciling conflicts in data modelling and encoding that exist between the data source and the ABK system- namely, systemic, syntactic and structural heterogeneities. This approach does not require any database-specific knowledge to be embedded into the ABK system- by contrast, a *data warehouse* requires the selection and transformation of specific data fields from the source database. The framework supports an extensible set of wrappers (database clients) for search and retrieval. These wrappers are not part of the ABK system, but are organized as plug-ins that can be installed independently by the user.

The wrappers embed all database-specific code, so that new databases can be integrated without changes to the central ABK system. This separation of database-specific concerns is achieved by defining a simple, generic interface which the wrappers must adhere to. This interface accepts a generic user-specified query, and returns a result set from which XML documents that match the query can be retrieved one by one. Thus the wrapper acts as an intermediary through which all databases can be accessed as XML document repositories whose documents are returned to the central ABK system for presentation, data extraction and aggregation. This approach hides much of the complexity of interacting with the remote database, since the wrapper a) translates the generic query into a query in the specific format required by the target data source; b) handles interaction with the remote data sources, hiding any database-specific synchronization mechanisms (e.g. how many records can be retrieved with a single HTTP call); and c) it can transform data encoded as database-specific non-XML formats into XML documents that can be presented to the user.

3.2.2 XML-based structural rules

Aggregated results from multiple databases are encoded according to different schemas, but need to be processed in a homogeneous way. The main obstacle is that the desired knowledge

must be extracted from query results using a mechanism independent of the source database's schema. Performing data extraction in the database wrapper would allow schema-specific processing, but would also introduce a semantic gap because the wrapper developers may interpret the data differently from the end user. The extraction of knowledge from an arbitrary schema, without leveraging on any schema-specific knowledge, may seem an impossible task. However, the task can be achieved by using generic mechanisms that can *navigate* a data structure, without knowing its semantics. This principle is widely used by tools that perform *screen scraping* on Web pages, such as aggregators of news and commercial information. These tools can collect information from multiple Web sites: for example, they may collect the prices of a given CD from several eCommerce sites, by extracting the price tags from their Web pages. These applications have no embedded knowledge of the target eCommerce site semantics. Rather, they support the definition of rules for extracting the desired information from a particular *location* in the source page, based on the assumption that the pages displayed by a given site will be identically structured for different products. In other words, they do not attempt to "understand" the page content (which would be almost impossible with HTML encoding), but merely to identify where the information is most likely to be.

A similar approach is applied to the processing of query results returned by the ABK wrappers, using a mechanism known as structural rules. The central idea is to retain a record from a remote source in its native form (without restructuring), and to provide a uniform mechanism for specifying structural rules that identify the location where the desired data is stored. The Extensible Markup Language (XML) (Bray *et al*, 2006) has several features which make it particularly suitable as an encoding platform for supporting this mechanism. An XML document is structured, and consists of a hierarchy of elements which can represent a wide variety of schemas. XML is thus capable of structuring records from any database, because it is independent of the schema semantics. The document structure is self-descriptive, since structural tags are encoded within the document. Such tagging removes the need for *a priori* knowledge of the schema: a valid XML document can be parsed using a standard off-the-shelf software library, which extract a hierarchical data structure (known as a DOM tree)

that can be queried structurally. XML has a standard syntax for querying document structure, known as XPath (Clark and DeRose 1999). A structural rule consists of an XPath query, specifying the path used to reach the desired value in a document- that is, the hierarchical path from the DOM tree root to the value location. It is possible to extract any XML document field by providing an XPath structural rule, regardless of the specific schema (set of XML tags) that has been used to encode the data. A key advantage of XML is that it is freely available and widely supported. Many larger public databases, such as GenBank (Benson *et al.* 2008) and UniProt (UniProt Consortium 2008), provide XML versions of their records, which can be handled directly by the ABK system. This is not always the case for many smaller databases, whose results may require conversion from their native representation to a suitable XML format; such conversion can be performed by the appropriate database wrapper.

3.2.3 Definition of structural rules by example

Although XML and XPath provide powerful generic mechanisms for implementing structural rules, the syntax of XPath rules is not sufficiently intuitive for non-technical users to specify them manually. Similarly, the syntax of "raw" XML documents can be verbose and difficult to analyze and to derive such rules. To allow users to specify XPath-based structural rules, we devised a user-friendly interface mechanism for specifying structural rules by example, while visually inspecting a document. Briefly, we present XML documents using a specially designed user interface component which presents the content as a hierarchy of name/value pairs, as shown in Figure 3-1. This novel component provides a compact display that evidences the data, so it can be intuitively selected by the user with a point-and-click gesture. The user gesture is automatically translated to an XPath rule; for simplicity, ABK structural rules support a limited subset of the XPath grammar, onlly allowing path constraints based on value matching. This restriction allows users to remove path ambiguities, while keeping the interface paradigm simple. The rule definition process can therefore be summarized as follows: the user opens an XML records for which no value has been extracted; finds the field that contains the value; selects the value, specifying constraints if needed; and finally

visualizes the extracted value. Once a rule is defined, it is automatically applied to all documents from the same database, as they are assumed to use the same XML schema.

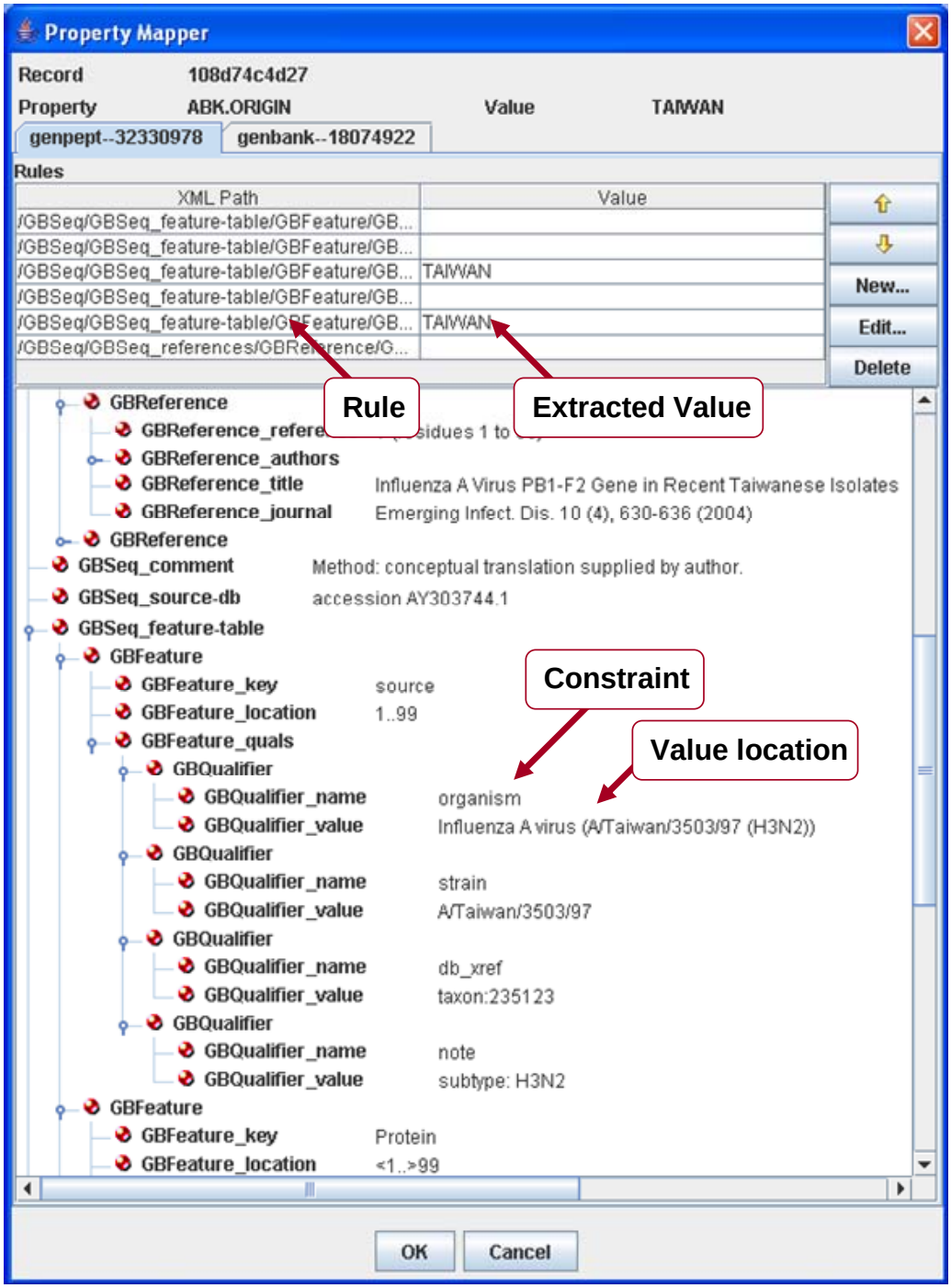


Figure 3-1: The ABK Record Viewer, showing an GenBank XML record
The XML record is presented as a tree of name/value pairs. Although the structure is specific to the originating database, XML labels make it understandable to biologists. The user interface allows the specification of the extraction rules, by selecting desired value and constraints as indicated. In this example, five separate rules were specified for geographic location, two of which returned consistent results.

Note that ABK does not attempt to interpret the semantics of the data field from which the data is selected, since information is often embedded within semantically unrelated fields. For example, the influenza isolate naming conventions specify that the sampling location and year must be embedded in the isolate name. Therefore, a user may choose to specify rules to extract year and country information from the "isolate name" field (see Table 3-1); to extract the final value, ABK is capable of processing the selected field values, by means of *filters* and *user-defined dictionaries*.

3.2.4 Filters and Dictionaries

The extraction of property values (for example, the host organism or isolation year of a viral sequence) often requires some transformation of the results of structural rules: for example, the desired property value may be embedded in free text. To overcome this source of syntactic heterogeneity, ABK uses *value filters*: plug-in modules that perform string processing tasks on XPath query results. The configurable filters currently provided by ABK suit a variety of tasks, and new filters can be added easily. The current set includes dictionary filters (capable, for example, of producing the value "CHICKEN" when encountering the string "bantam", see Figure 3-2), regular expression filters (capable to recognize formatted strings, such as NCBI identifiers "ABB12345.1"), and date parsing filters, able to recognize years in 2- and 4-digit formats.

In ABK, user-defined dictionaries are generally task-specific, for a variety of reasons. Firstly, the input data may contain values that are recognized and meaningful in the context of the current aggregation task, but are not universal. For example, the abbreviation "TY" is commonly used in influenza isolate names to indicate a turkey host, but may be interpreted as a yeast transposon in a genomics dataset. Second, a specific taxonomic granularity of the extracted values may needed by the downstream analysis tasks. For example, the host organism categories "Avian", "Human" and "Others" may be sufficient in a study of the transmissibility of avian influenza to humans, while studies of avian influenza may need to organize avian sequences by bird orders (as shown in Figure 3-2). In either case, a dictionary

that produces species names may be inappropriate, as it would make the downstream selection of groups (e.g. “Galliformes”) more complex. Finally, it is advantageous to restrict the dictionary to the actual values observed in the aggregated dataset, to avoid false positives. Although the dictionary shown in Figure 3-2 may appear incomplete and patchy at first sight, this is simply a reflection of the value distribution in the aggregated influenza dataset, for which this dictionary has been designed. Although it is possible to construct a “standard” dictionary comprising a complete tree-of-life taxonomy, such dictionary would mostly consist of irrelevant entries, and would be hard to customize by the user. Worse still, it would be prone to producing errors for reasons that may not be immediately obvious. Using a sparsely populated dictionary, a user would be forced to inspect records that do not yield a value, and identify the desired literal value; in such cases, however, a comprehensive dictionary of species would have a high match probability amongst the irrelevant entries, particularly in cases where values are incorrectly or ambiguously entered. For example, if a submitter entered the common bird name “oystercatcher” erroneously as “oyster catcher”, a match to “oyster” would be found in a full animal taxonomy, while a custom dictionary would return no value, forcing the user to verify the record.

```
AVIAN, BIRD
AVIAN (CHICKEN), CHICKEN, CK, CHICK, SILKY CHICKEN, SCK, CHICKENS, BANTAM, CO
AVIAN (TURKEY), TURKEY, TY
AVIAN (GALLIFORMES), PARTRIDGE, PHEASANT, PH, QUAIL, QA, QUAILS, GROUSE, PEAF
AVIAN (COLUMBIDAE), PIGEON, FERAL PIGEON, DOVE
AVIAN (ANATIDAE), DUCK, DK, MUSCOVY DUCK, DUCKS, GARGANEY, MALLARD, ANAS PLAT
AVIAN (CHARADRIIFORMES), GULL, BLACK HEADED GULL, BLACK-HEADED GULL, GULLS, G
AVIAN (CICONIIFORMES), HERON, GREY HERON, EGRET, EG, STORK, OPENBILL STORK, O
AVIAN (STRUTHIONIFORMES), OSTRICH, EMU, RHEA
AVIAN (PASSERINE), PASSERINE, BLACKBIRD, BLUEBIRD, CROW, FINCH, NIGHTINGALE,
AVIAN (PSITTACINE), PSITTACINE, PARROT, MACAW, CONURE, PARAKEET, BUDGERIGAR
AVIAN (FALCONIFORMES), FALCON, PEREGRINE FALCON, HAWK, EAGLE, CRESTED EAGLE
AVIAN (ROLLERS), ROLLER, ROLLERS
AVIAN (PROCELLARIIFORMES), SHEARWATER, PUFFIN, PETREL
HUMAN, HOMO SAPIENS
MAMMAL (EQUINE), EQUINE, HORSE
MAMMAL (SWINE), SWINE, SW
MAMMAL (CARNIVORA), CANINE, DOG, FELINE, CAT, LEOPARD, TIGER, MINK, SEAL
MAMMAL (ARTIODACTYLA), CAMEL
MAMMAL (CETACEA), WHALE
ENVIRONMENT
```

Figure 3-2: A fragment of a user-defined dictionary for value filtering
Here we show a fragment of a user-defined dictionary used to extract host organisms information from a free text value. If one of the strings listed is matched, the filter yields the value at the start of the corresponding line as the result. This dictionary, used for aggregating influenza A records, was edited by the end user in the course of the curation task. The organization of entries is determined by the user, and reflects the needs of the analysis task. The dictionary is generally organized according to taxonomic order, but certain important hosts, such as human, chicken, turkey, swine and equine, are classified separately because of their special significance in influenza research.

3.2.5 Conflict Resolution

Users are allowed to specify multiple structural rules for extracting a given value from the same XML schema, and these rules are organized in order of priority. The support of multiple properties allows a large degree of semantic flexibility, such that several database fields can be inspected to determine the extracted value. In most documents, only certain rules will produce values; the final value extracted from a document is that of the highest-priority rule that yields a value. However, if values from other rules conflict with this "winning" value, this conflict is flagged in the ABK system, and highlighted by the user interface (see Figure 3-5). Since ABK supports the association of a record to multiple source documents (from multiple databases), the user can also specify database priority, thus establishing an

order for processing documents and identifying the overall “winning value”. Conflicts resulting from differing values extracted from multiple documents are also automatically highlighted to alert the user.

3.3 ABK architecture and components

The ABK system has been implemented as a desktop-based software tool, installable on all Java-enabled platforms (such as Windows and Linux). The ABK software is freely available from the author, on request. It implements the technological specifications described in Section 3.2, and it provides record management facilities, a plug-in tool architecture for extending the software with analysis tools, and a simple spreadsheet-like user interface for visualizing the aggregated extracted information. End users control three major subsystems, as shown in Figure 3-3, and in expanded form in Figure 3-4. The Data Collection subsystem accepts user queries, executing them through the mediator framework. The query results (XML documents) are then handled by the Data Management subsystem, which applies structural rules and aggregates the results. Finally, the Data Analysis subsystem allows analysis tools to interact with the stored data, and augment it with further results.

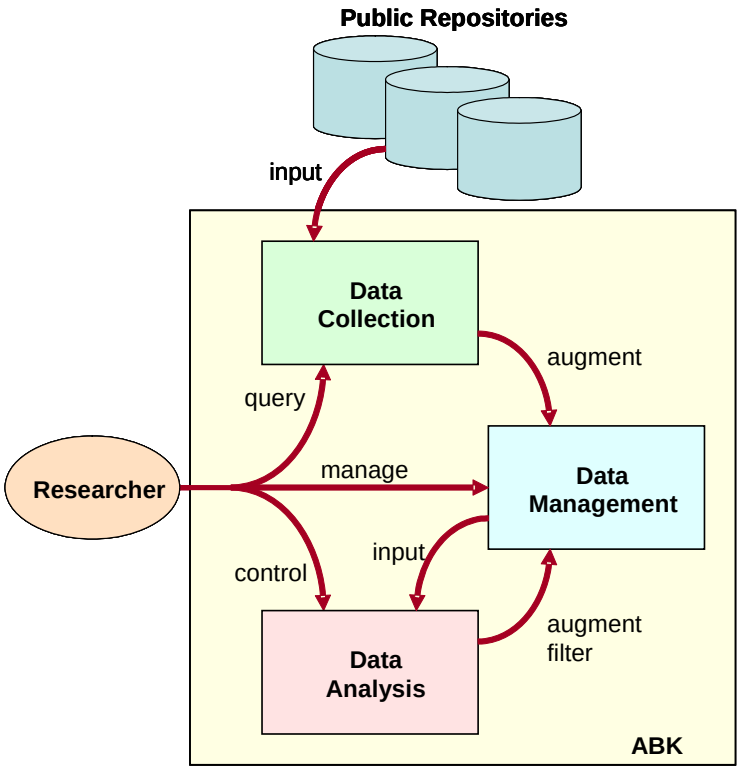


Figure 3-3: Architecture of the ABK system

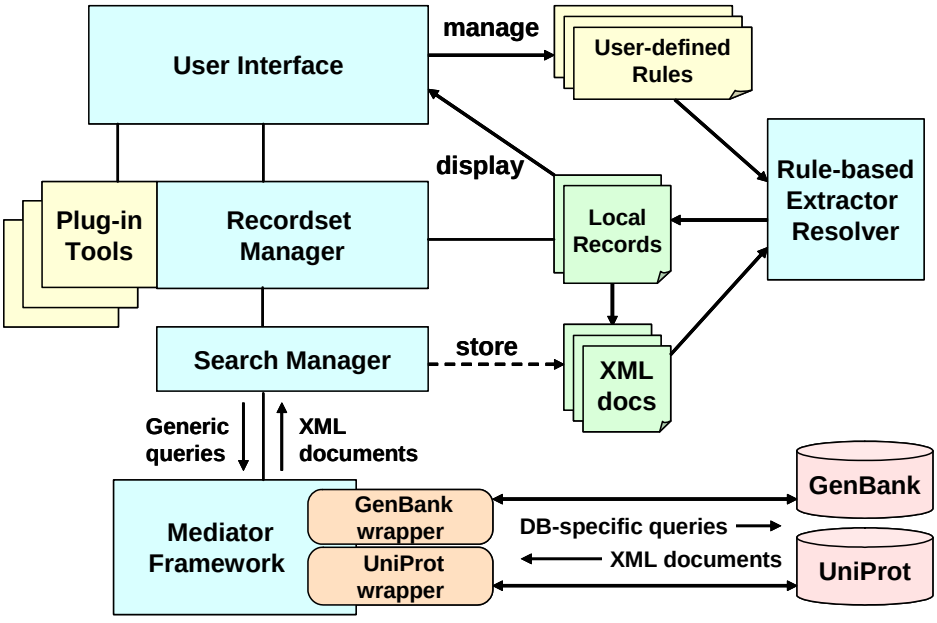
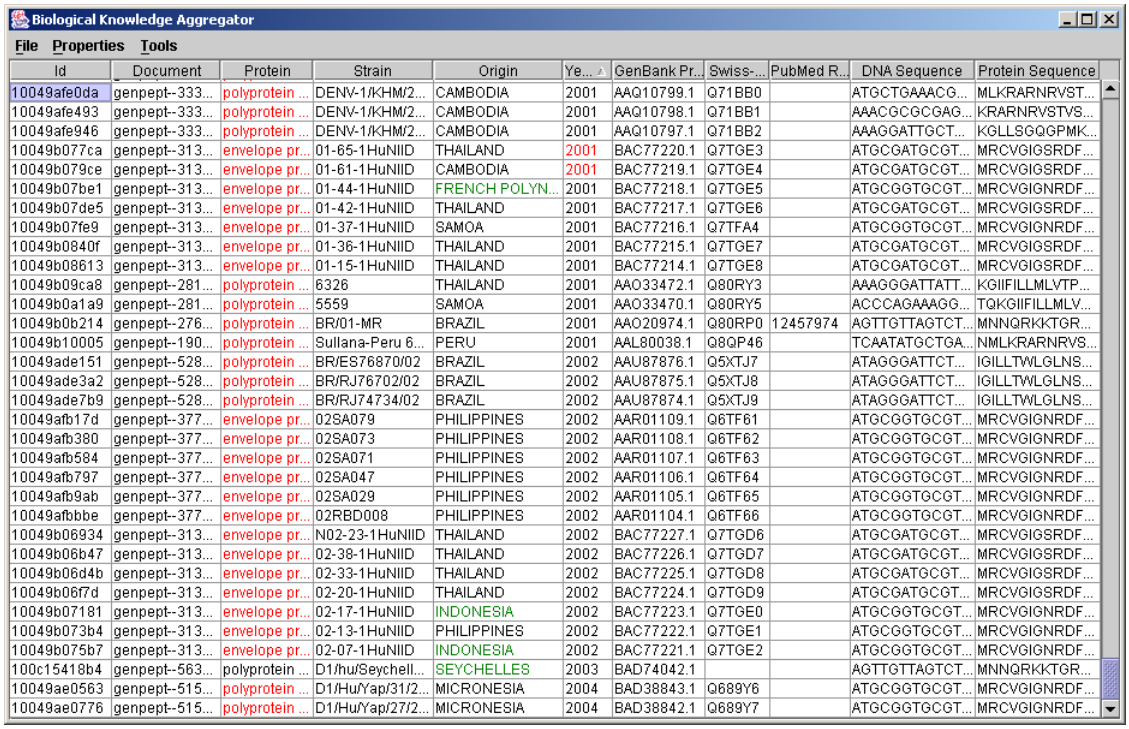


Figure 3-4: Detailed Architecture of the ABK system

The *mediator framework* supports an extensible set of database wrappers for search and retrieval (only two example wrappers are shown in Figure 3-4). Generic user-specified queries are translated into data source-specific queries, and submitted by the wrappers to the remote data sources. XML Documents retrieved by the mediator framework are stored locally by the *recordset manager*, which constructs collections of local records. Documents from various sources which describe the same entity may be aggregated (for instance, the protein product of DNA sequence records in GenBank is often described by records in both UniProt and GenPept databases). To create a new field in a local record set, users specify structural rules by example, using the *record viewer* shown in Figure 3-1. The structural rules are subsequently automatically applied by the *rule-based extractor and resolver*, and the resulting aggregated value, after text filtering, are displayed through a graphical spreadsheet-like user interface (see Figure 3-5). The resolver identifies conflicts among the field values extracted from different documents, and among values extracted from the same document using different rules. Such conflicts are highlighted by text in red font colour (see Fig. 3-5) in the user interface. This alerts user to rapidly check and, if necessary, reconcile the value after manual inspection, or reject the record altogether. Finally, *plug-in tools* are the analysis tools that can be applied to the whole dataset, or to a subset of records. They are useful for connecting to external systems, and are generally application-specific.



Id	Document	Protein	Strain	Origin	Ye...	GenBank Pr...	Swiss-...	PubMed R...	DNA Sequence	Protein Sequence
10049afe0da	genpept--333...	polyprotein ...	DENV-1/KHM/2...	CAMBODIA	2001	AAQ10799.1	Q71BB0		ATGCTGAAACG...	MLKRARNRVST...
10049afe493	genpept--333...	polyprotein ...	DENV-1/KHM/2...	CAMBODIA	2001	AAQ10798.1	Q71BB1		AAACGCGCGAG...	KRARNRVSTVS...
10049afe946	genpept--333...	polyprotein ...	DENV-1/KHM/2...	CAMBODIA	2001	AAQ10797.1	Q71BB2		AAAGGATTGCT...	KGLLSGQGPMK...
10049b077ca	genpept--313...	envelope pr...	01-65-1HuNIID	THAILAND	2001	BAC77220.1	Q7TGE3		ATGCGATGCGT...	MRCVGIGSRDF...
10049b079ce	genpept--313...	envelope pr...	01-61-1HuNIID	CAMBODIA	2001	BAC77219.1	Q7TGE4		ATGCGATGCGT...	MRCVGIGSRDF...
10049b07be1	genpept--313...	envelope pr...	01-44-1HuNIID	FRENCH POLYN...	2001	BAC77218.1	Q7TGE5		ATGCGGTGCGT...	MRCVGIGNRDF...
10049b07de5	genpept--313...	envelope pr...	01-42-1HuNIID	THAILAND	2001	BAC77217.1	Q7TGE6		ATGCGATGCGT...	MRCVGIGSRDF...
10049b07fe9	genpept--313...	envelope pr...	01-37-1HuNIID	SAMOA	2001	BAC77216.1	Q7TFA4		ATGCGGTGCGT...	MRCVGIGNRDF...
10049b0840f	genpept--313...	envelope pr...	01-36-1HuNIID	THAILAND	2001	BAC77215.1	Q7TGE7		ATGCGATGCGT...	MRCVGIGSRDF...
10049b08613	genpept--313...	envelope pr...	01-15-1HuNIID	THAILAND	2001	BAC77214.1	Q7TGE8		ATGCGATGCGT...	MRCVGIGSRDF...
10049b09ca8	genpept--281...	polyprotein ...	6326	THAILAND	2001	AAO33472.1	Q80RY3		AAAGGGATTATT...	KGIIFILLMLVTP...
10049b0a1a9	genpept--281...	polyprotein ...	5559	SAMOA	2001	AAO33470.1	Q80RY5		ACCCAGAAAGG...	TQKGIIFILLMLV...
10049b0b214	genpept--276...	polyprotein ...	BR/01-MR	BRAZIL	2001	AAO20974.1	Q80RP0	12457974	AGTTGTTAGTCT...	MNNQRKKTGR...
10049b10005	genpept--190...	polyprotein ...	Sullana-Peru 6...	PERU	2001	AAL80038.1	Q8QP46		TCAATATGCTGA...	NMLKRARNRVS...
10049ade151	genpept--528...	polyprotein ...	BR/ES76870/02	BRAZIL	2002	AAU87876.1	Q5XTJ7		ATAGGGATTCT...	IGILLTWLGLNS...
10049ade3a2	genpept--528...	polyprotein ...	BR/RJ76702/02	BRAZIL	2002	AAU87875.1	Q5XTJ8		ATAGGGATTCT...	IGILLTWLGLNS...
10049ade7b9	genpept--528...	polyprotein ...	BR/RJ74734/02	BRAZIL	2002	AAU87874.1	Q5XTJ9		ATAGGGATTCT...	IGILLTWLGLNS...
10049afb17d	genpept--377...	envelope pr...	02SA079	PHILIPPINES	2002	AAR01109.1	Q6TF61		ATGCGGTGCGT...	MRCVGIGNRDF...
10049afb380	genpept--377...	envelope pr...	02SA073	PHILIPPINES	2002	AAR01108.1	Q6TF62		ATGCGGTGCGT...	MRCVGIGNRDF...
10049afb584	genpept--377...	envelope pr...	02SA071	PHILIPPINES	2002	AAR01107.1	Q6TF63		ATGCGGTGCGT...	MRCVGIGNRDF...
10049afb797	genpept--377...	envelope pr...	02SA047	PHILIPPINES	2002	AAR01106.1	Q6TF64		ATGCGGTGCGT...	MRCVGIGNRDF...
10049afb9ab	genpept--377...	envelope pr...	02SA029	PHILIPPINES	2002	AAR01105.1	Q6TF65		ATGCGGTGCGT...	MRCVGIGNRDF...
10049afbbbe	genpept--377...	envelope pr...	02RBD008	PHILIPPINES	2002	AAR01104.1	Q6TF66		ATGCGGTGCGT...	MRCVGIGNRDF...
10049b06934	genpept--313...	envelope pr...	N02-23-1HuNIID	THAILAND	2002	BAC77227.1	Q7TGD6		ATGCGATGCGT...	MRCVGIGSRDF...
10049b06b47	genpept--313...	envelope pr...	02-38-1HuNIID	THAILAND	2002	BAC77226.1	Q7TGD7		ATGCGATGCGT...	MRCVGIGSRDF...
10049b06d4b	genpept--313...	envelope pr...	02-33-1HuNIID	THAILAND	2002	BAC77225.1	Q7TGD8		ATGCGATGCGT...	MRCVGIGSRDF...
10049b06f7d	genpept--313...	envelope pr...	02-20-1HuNIID	THAILAND	2002	BAC77224.1	Q7TGD9		ATGCGATGCGT...	MRCVGIGSRDF...
10049b07181	genpept--313...	envelope pr...	02-17-1HuNIID	INDONESIA	2002	BAC77223.1	Q7TGE0		ATGCGGTGCGT...	MRCVGIGNRDF...
10049b073b4	genpept--313...	envelope pr...	02-13-1HuNIID	PHILIPPINES	2002	BAC77222.1	Q7TGE1		ATGCGGTGCGT...	MRCVGIGNRDF...
10049b075b7	genpept--313...	envelope pr...	02-07-1HuNIID	INDONESIA	2002	BAC77221.1	Q7TGE2		ATGCGGTGCGT...	MRCVGIGNRDF...
100c15418b4	genpept--563...	polyprotein ...	D1/hu/Seychell...	SEYCHELLES	2003	BAD74042.1			AGTTGTTAGTCT...	MNNQRKKTGR...
10049ae0563	genpept--515...	polyprotein ...	D1/Hu/Yap/31/2...	MICRONESIA	2004	BAD38843.1	Q689Y6		ATGCGGTGCGT...	MRCVGIGNRDF...
10049ae0776	genpept--515...	polyprotein ...	D1/Hu/Yap/27/2...	MICRONESIA	2004	BAD38842.1	Q689Y7		ATGCGGTGCGT...	MRCVGIGNRDF...

Figure 3-5: ABK provides a spreadsheet-like presentation of extracted data
Each row in the display corresponds to a single record, and the columns represent the extracted properties. Values displayed in a red font indicate that there is a conflict between the results of multiple rules. Green fonts indicate fields where conflicts have been manually corrected.

3.3.1 Applications of ABK

In the period of candidature covered by the present thesis, the ABK framework has been used by the author and by other researchers at the Department of Biochemistry for several studies. The following is a summary of this application work:

- Large datasets of viral sequences and accompanying metadata were assembled for use in immunological and virological studies. Data sets were produced by the author and other collaborators for influenza A virus, dengue virus, rabies virus and hantavirus. For each virus, domain-specific vocabularies were developed for the precise identification of host organisms and protein names.
- A proof-of-concept plug-in tool was implemented to support the phylogenetic analysis of extracted viral sequences, and the visualization of the resulting trees.
- A set of plug-in tools were developed for the analysis of text tokens, identification of

text patterns, creation of text occurrence matrices and the rapid annotation of biomedical abstracts.

- A corpus of biomedical abstracts for allergy research was created and annotated.
- A plug-in tool was created for aggregating records from different databases by matching identifiers.

In the following sections of this chapter, we have described specific knowledge aggregation tasks, to illustrate the utility and performance of ABK, and the level of heterogeneity in public databases that this approach can overcome.

3.4 Curation of a large-scale influenza protein dataset

We used the ABK platform to construct a large-scale datasets of influenza A sequences and their accompanying metadata through the application of XML-based *structural rules*. The task involved retrieving as many protein sequences as possible from public databases, and extracting several metadata properties that are later used to perform a variety of comparative analyses. Since full manual curation of such a large volume of records is prohibitively expensive, we devised an approach to make this analysis viable, in which most of the metadata curation task is automated. After applying structural rule extraction, many records still had incomplete or conflicting metadata, and had to be manually corrected. Even with the help of productive tools, two expert curators were required to work intensively for two weeks to manually complete and verify the annotations. We used this metadata to analyze the performance of structural rules, and quantify the extent of semantic heterogeneity and inconsistencies, both within and between the two popular databases GenBank and GenPept (Benson *et al.* 2008).

3.4.1 Task Requirements

The construction of the influenza dataset was conducted as the data preparation stage of a multi-stage study. The study aimed at determining a number of genetic, evolutionary and

immunological properties of the influenza A virus, by analyzing as many protein sequences as possible. Eleven large separate multiple sequence alignments (MSA) were created, one for each of the proteins expressed by this virus. The alignments were analyzed to identify:

- Alignment positions where adaptation to a given host (such as human) produced specific differentiation from the natural avian form of the virus. This knowledge was used to assess the potential for transmissibility to humans of avian influenza targets, as described in Chapter 6.
- Protein regions that were conserved (i.e. did not mutate) for certain subtypes of the virus (e.g. H5N1, H3N2) over given periods of times. This knowledge is used in the context of identifying potential conserved vaccine targets, as described in Chapter 7.
- Mutations associated with specific geographical areas or periods of time.
- Alignment positions that co-evolve in different proteins (i.e. when one of the positions mutates, the other mutates simultaneously).

The analysis tasks described can only be automated if sequences are accompanied by *descriptive metadata*. Metadata is "data about the data" (Swedlow *et al.* 2006), for example provenance information about the amino acid sequences that constitute the main data. Our study required the following fields: the ***subtype*** of the virus, the ***protein name*** for the sequence, the ***isolate name*** (used to associate multiple proteins for studying co-evolution), the ***host*** organism from which the virus was isolated, and the ***year*** and ***origin*** (country) of isolation.

The study data was retrieved from the two major public databases at NCBI: GenBank (a nucleotide database) and GenPept (a protein database). In September 2006, over 90,000 relevant records were available, although the actual number of unique sequences was much lower. In most cases, each GenBank record has a corresponding record in GenPept, containing the nucleotide sequence's translation. GenPept often contains multiple alternative versions of its own records, mirrored from other public databases. Although NCBI records provide semi-structured metadata, it is frequently plagued by heterogeneous encoding and

quality issues, as reported in other studies (Karp *et al.* 2001; Brusic, Millot *et al.* 2003). We found that records documenting the same sequence do not necessarily carry the same metadata, and sometimes provide conflicting information. Metadata is frequently missing, and the choice of record field for encoding a given metadata property is often arbitrary and inconsistent. Metadata values can be difficult to extract even when their location can be correctly identified- for example, because of free text embedding, misspellings and inaccuracies, or non-standard granularity (e.g. a city specified rather than a country) (Koh *et al.* 2005). Such issues make this knowledge aggregation problem an excellent case study not only to verify the effectiveness of the ABK approach, but also for quantifying the extent of heterogeneities in public databases.

3.4.2 Task Structure

The workflow of this task can be modeled as a knowledge mining process, as shown in Figure 3-6. The source databases are queried through the mediator framework using a simple taxonomy query for the influenza A virus. The retrieved XML files are subsequently merged when they are found to refer to the same virus proteins, as detected by the identifier cross-referencing tool in ABK. Finally, structural rules processing extracts data and metadata, both to be manually verified within the ABK tool. The output of the task are: a set of sequences for each influenza protein (which are to be subsequently aligned), and a metadata file. ABK can output the sequence sets as FASTA files, while the metadata can either be written as a simple comma-separated value (CSV) file, or as an RDF file using a simple *ad hoc* ontology (an application of RDF-formatted data will be described in Chapter 4).

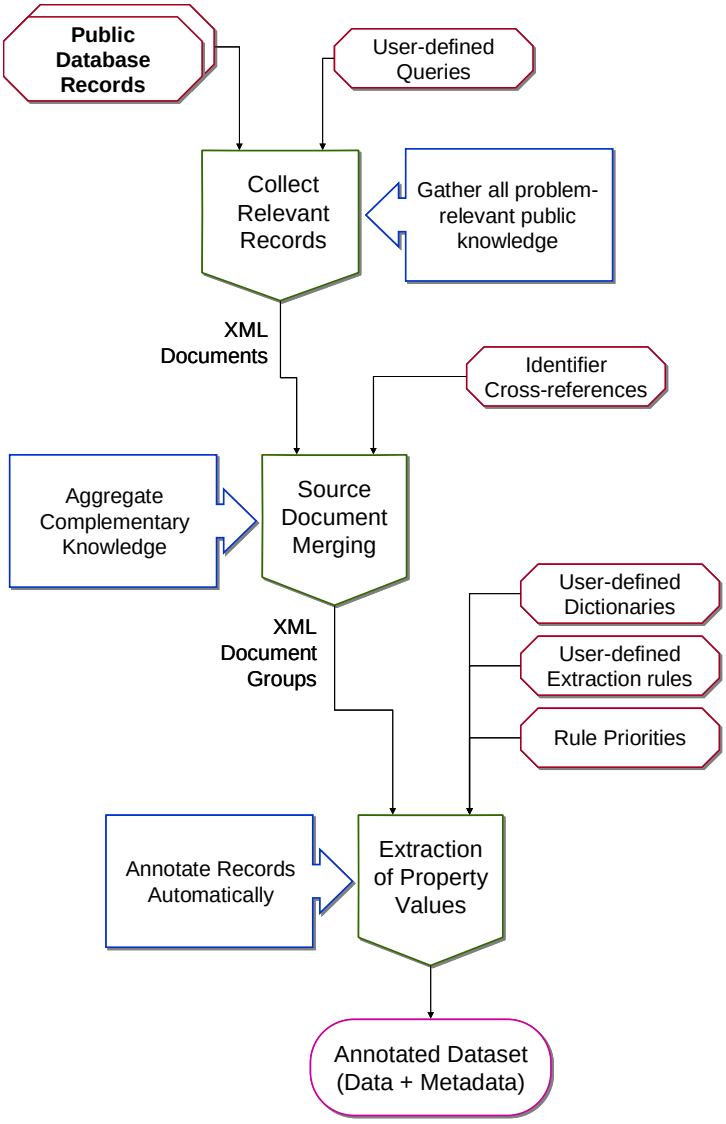


Figure 3-6: Knowledge Mining Model for the Biological Knowledge Aggregation process.

3.4.3 Methods

Data retrieval was performed by a taxonomy query submitted to the two NCBI sequence databases, retrieving a total of 92,343 documents (39,775 from GenBank and 52,568 from GenPept). These documents were encoded in native NCBI XML format. ABK extracted cross-referencing identifiers from each document, and matched them to identify multiple documents referring to the same sequence. This task reduced the total dataset to 40,169 records, each associated to at least one and no more than three database documents. Each of these records represents a protein sequence from a given isolate (GenBank record have

protein sequence encoded as a feature field).

Metadata extraction was performed by ABK using structural rules. For each metadata property, multiple structural rules can be defined, with an order of priority chosen by the user. Both databases accessed in this study use the same XML schema and thus a common set of structural rules was specified (see Table 3-1). The rules and their priority were determined by an expert curator, based on manual inspection of several representative records. The same curator assigned GenPept a higher priority than GenBank.

For the text filtering stage, two influenza-specific dictionaries were developed: one to extract the host organism information (a fragment of which is shown in Figure 3-2), and one to extract protein type information. In addition, a regular expression filter was configured to match standard influenza A isolate identifiers.

3.4.4 Results

We measured the yield and accuracy of property value extraction from both GenBank and GenPept. Yield was defined as the fraction of documents from a given database that produces a value from structural rules, while accuracy was computed as the percentage of extracted values that matches the manually curated property value (i.e. the property value at the end of full manual curation of the dataset by two independent domain experts). These results are summarized in Figure 3-7. The yield difference between the two databases (approximately 9% for origin and host) indicates that GenBank records have more detailed annotation, justifying the decision to aggregate records from both databases. The two databases provided values with almost identical accuracy (within 1% for most properties), indicating that their priority order was not critical to the outcome of the extraction task. Accuracy ratings exceeded 96%, except for the *host* property, which produced accuracies of 89% for GenPept and 91% for GenBank. Although this might still seem a high level of accuracy, it resulted in some 4,000 host annotations requiring manual correction.

Property **proteinName**	
1	/GBSeq/GBSeq_definition
2	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='gene']/GBQualifier_value
Property **subtype**	
1	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='strain']/GBQualifier_value
2	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='isolate']/GBQualifier_value
3	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='organism']/GBQualifier_value
Property **isolate**	
1	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='strain']/GBQualifier_value
2	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='isolate']/GBQualifier_value
3	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='organism']/GBQualifier_value
Property **host**	
1	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='specific_host']/GBQualifier_value
2	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='strain']/GBQualifier_value
3	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='isolate']/GBQualifier_value
4	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='organism']/GBQualifier_value
Property **origin**	
1	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='country']/GBQualifier_value
2	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='isolation_source']/GBQualifier_value
3	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='strain']/GBQualifier_value
4	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='isolate']/GBQualifier_value
5	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='organism']/GBQualifier_value
6	/GBSeq/GBSeq_references/GBReference/GBReference_title
Property **year**	
1	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='note']/GBQualifier_value
2	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='isolation_source']/GBQualifier_value
3	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='strain']/GBQualifier_value
4	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='isolate']/GBQualifier_value
5	/GBSeq/GBSeq_feature-table/GBFeature/GBFeature_quals/GBQualifier[GBQualifier_name='organism']/GBQualifier_value

Table 3-1: Structural rules employed for the extraction of sequence record properties from GenBank and GenPept.
For each property, the XPath expressions of all relevant structural rules are given in order of priority (lower numbers indicate higher priority). The proteinName property was only extracted from GenPept.

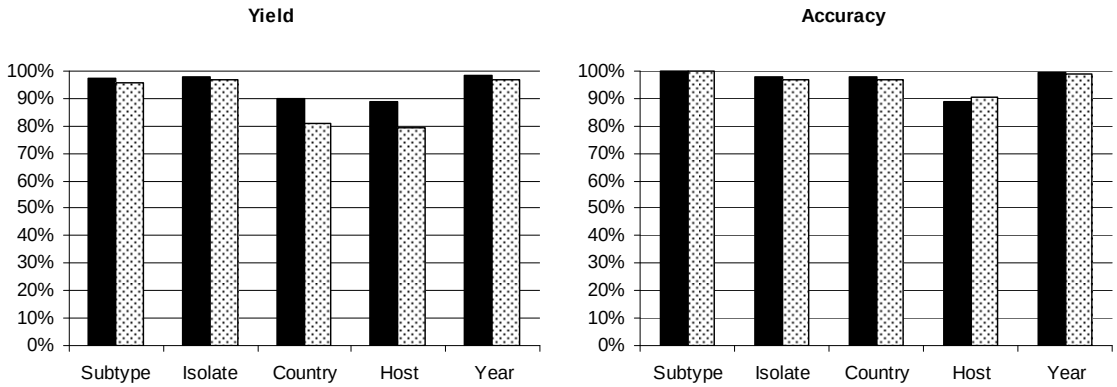


Figure 3-7: Retrieval performances of the NCBI nucleotide and protein databases
Each chart shows 5 pairs of bars, one for each extracted property. The first (darker) bar of each pair shows the performance for the GenBank database while the second (lighter) bar shows the value for GenPept. The first chart shows the percentage of source documents from which a property value could be extracted, while the second graph shows the percentage of accurate values extracted, measured against the manually annotated dataset.

The low accuracy of the host property is related to its low yield (79.5%-88.8%), primarily caused by a high proportion of human influenza sequences, which frequently lack the host annotation. Isolate naming standards are not adequate for automating metadata extraction, since they allow the host to be omitted from identifiers of human isolates, making the extraction of this property very problematic. In this study, we have chosen not to assume that an isolate identifier without host name necessarily implies a human virus. Such an assumption would have produced much higher yields, but also a much higher number of incorrect annotations. In these cases, we resorted to manual curation, which was expedited considerably by the spreadsheet-like interface of the ABK tool.

Each property required a number of structural rules to be applied, each rule defined to extract a relevant value from a different location in the source document. The performance of the structural rules used was analyzed, and Figure 3-8 shows the percentage of documents for which a given rule was the "winning" rule for a given property, i.e. the highest-priority structural rule that produced a value. The performance diagrams display several interesting features. First, they clearly show the extent of semantic heterogeneity in public databases: although the most productive structural rule can be identified for each property, contributions

from other rules can constitute up to 35% of the extracted values. Second, it is evident that a human expert does not always rank the rules by their productivity, but rather by their perceived accuracy. Finally, the charts for properties *isolateName*, *origin* and *year* clearly show that identical rules produced values more frequently from GenBank, although documents from the two databases are identically structured. This clearly indicates that GenBank records are often more thoroughly annotated by submitters. Extraction from many GenPept records frequently has to rely on lower-priority rules, and sometimes does not yield any value at all. This clearly has a negative impact on studies of protein sequences, since researchers may limit their data gathering to the GenPept database, thus omitting significant proportions of the metadata.

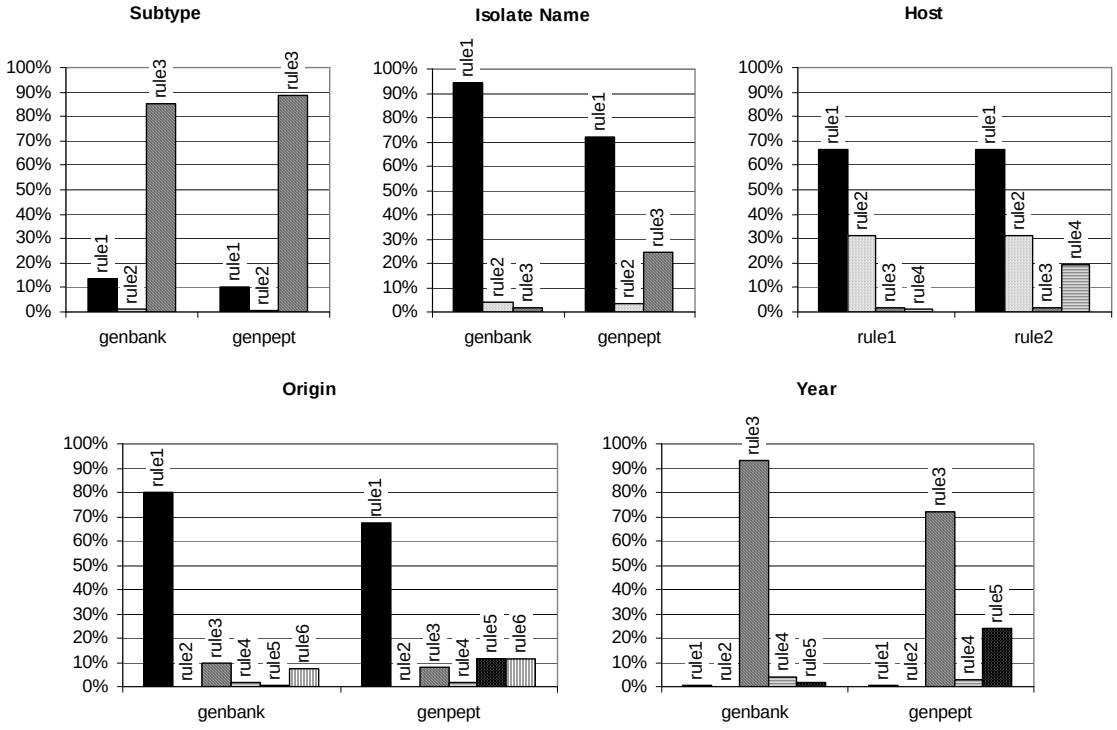


Figure 3-8: Performance of structural rules for five metadata properties
Bars show the percentage of records for which a given structural rule produced the final property value. Rules are numbered according to their priority, matching the priorities shown in Table 1.

3.5 Discussion

Semantic heterogeneity is a serious obstacle in the production of annotated datasets, and a semi-automated approach is currently the only practical solution when studies need to process thousands of records. We have shown that the ABK platform can recover a very high proportion of the necessary metadata, through the application of XML-based structural rules.

Our case study presented relatively humble metadata needs: a small number of highly relevant fields, with little structural complexity. Yet, we have shown that the NCBI databases, arguably the most important primary data sources used in bioinformatics, are incompletely and inconsistently annotated to the extent that meeting even such simple requirements is a major challenge for automated tasks. One might argue that the problem could be solved by simply choosing for each property the most productive source database field, and discarding those records that do not yield a value. The results in Fig. 3-8 suggest that this approach may fully annotate up to 65% of records, which would still form a large-scale dataset. However, such a draconian mechanism would introduce major biases: since large influenza surveillance projects tend to cover specific geographies (e.g. North America), and provide more complete metadata, discarding records based on metadata quality would eliminate mostly records that are submitted by smaller projects, and thus greatly decrease the diversity of the dataset. Such bias would undermine the statistically-supported results of large-scale studies. To put it simply, metadata that is hard to recover is sometimes more valuable than metadata that is easily accessed.

The large proportion of data from influenza surveillance projects should also be considered when reviewing the results of our isolate-based restructuring task. The number of inconsistencies in *isolateName* may appear surprising low (only 5% of the 7,640 unique isolate names), but most of the credit goes to the existence of a standard influenza isolate nomenclature (World Health Organization 1972), and to high reporting discipline and consistency of large-scale project that every year submit large numbers of new sequences isolated in specific geographies. None of the 388 isolate name corrections involved records

from large surveillance projects; the vast majority of the corrected records involved animal sequences, confirming that the techniques used were beneficial for improving dataset diversity.

Structural rule-based extraction can deliver the intelligence necessary to reconstruct metadata for a great proportion of records. Automated recovery of the order of 90-95% makes it possible to complete the annotation process manually for the remaining records lacking metadata. XPath-based structural rules could achieve most of this metadata recovery in this study. Structural rules are a very powerful means for extracting annotations, yet simple to set up even for researchers with low technical skills, and highly generic since they can process data encoded in any database schema.

3.6 Conclusions

In this chapter we have presented an innovative approach to knowledge aggregation, and its implementation in the ABK software tool. The key contribution of the ABK approach is that it empowers real biomedical researchers, without programming skills, to overcome the system and information heterogeneities that currently prevent them from constructing large-scale metadata-rich datasets. The specification by example of structural rules is simple and intuitive, and rules can be prioritized to account for the different reliability of source data fields. Finally, text filters based on dictionaries provide a simple method for user-driven transformation.

We have shown the utility of our approach and tools by constructing a large-scale influenza A protein sequence dataset, including several metadata fields. Our results showed that information heterogeneities in the source data were a very significant obstacle, and that structural rules were very effective at recovering metadata values, minimizing the effort required for manual verification and curation of the dataset. In summary, ABK enabled a team of two researchers to complete within a short period of time the task of aggregating tens of thousands of records, which would have been prohibitively laborious using manual curation, demonstrating the scalability of our approach. In Chapters 6 and 7 we will demonstrate that

the construction of such a dataset enables new discoveries of importance in immunology and virology. Our knowledge aggregation can therefore be considered an important enabler of biomedical discovery.

4. SEMANTIC TECHNOLOGIES FOR BIOLOGICAL KNOWLEDGE REPRESENTATION

The biological knowledge mining approach presented in Chapter 2, Section 2.4 requires *knowledge flow* between the tasks that compose a bioinformatics pipeline. In other words, knowledge (comprising data, metadata, and analysis results) is used as input to analysis tasks, which can then augment the knowledge with the addition of results. Augmented knowledge is then transferred to downstream tasks, which continue the process, further enhancing the aggregated knowledge. This knowledge mining vision, originally formulated by Michalski (2003) requires two enabling components: a suitable *knowledge representation* capable of expressing knowledge in a generic and standard fashion, and *knowledge-enabled tools* capable of interpreting this knowledge representation, and augmenting it with new results.

In this chapter, we reviewed a set of technologies (collectively known as s*emantic technologies*) which have the potential to fulfill the knowledge representation requirements of our biological knowledge mining approach. These technologies are implemented as standards, and therefore can be integrated in any bioinformatics tools. We showed that knowledge encoded using semantic technologies can be easily extended, and therefore can be augmented with new information from analysis tools. In addition, semantic technologies are suitable for processing by standard *reasoners*, programs that can apply semantic rules to modify, restructure and augment knowledge. To demonstrate the utility of this technology solution, we encoded the large-scale aggregated influenza A dataset described in Chapter 3, Section 3.4 using RDF and OWL technologies. By applying relatively simple semantic rules to the dataset, we transformed its structure and improved its quality, inferring many missing metadata values from those of other records from the same isolates. Alternative approaches to this task would require considerable amounts of custom programming. This study is a proof-of-concept of the applicability of semantic technology to our biological knowledge mining approach. This technology stack is still evolving, and its current limitations are discussed in this chapter. Semantic technologies will therefore be the subject of further research, but their

versatility and generality make them very promising candidates to support biological knowledge mining.

4.1 Knowledge Representation in Bioinformatics

The most fundamental need for hierarchical scalability is that of representing the knowledge to be transferred, so that it can be easily consumed by downstream analysis tasks. *Knowledge representation* is a well-known problem space in artificial intelligence, since it is a prerequisite for reasoning over knowledge using machines. Davis *et al.* (1993) assigned five distinct roles to knowledge representation:

- As a *surrogate* of the reality it represents, imperfect and limited
- As a *set of ontological commitments*, since it captures a particular perspective of reality, with a given vocabulary and set of axioms
- As a *fragmentary (*reductionist*) theory of intelligent reasoning*, since it is generally only capable to reason using a small set of beliefs
- As a *medium of efficient computing*, since the purpose of the representation is to perform an analysis task
- As a *medium of human expression*, for users to communicate intentions to machines

All these distinct roles play some part in the bioinformatics discovery process. The knowledge produced by analysis tasks generally describes patterns or properties of biological entities (the reality being represented). These descriptions have a specific perspective, determined by the purpose and context of the upstream task, which is reflected in the vocabulary and structure of the produced knowledge. Downstream tasks that consume knowledge have specific processing (reasoning) capabilities, which also supply a context for the interpretation of knowledge, as well as computational constraints on task execution. Finally, it is human intention that drives knowledge processing, typically in the form of rules.

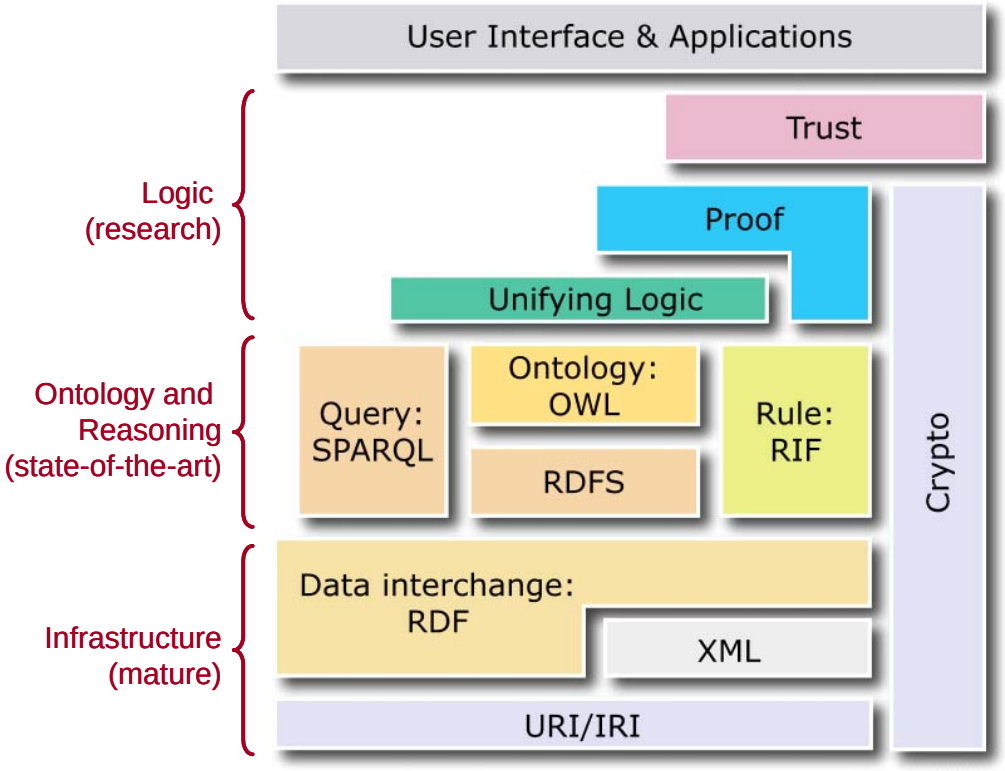


Figure 4-1: Semantic Web "layercake" architectural diagram
This "official view" of the Semantic Technologies layers has been annotated (on the left) to show the state of development and adoption of the various layers.

Source: http://www.w3.org/2001/sw/, accessed 16 May 2008.

4.2 Semantic Technologies

The technology platform needed to tackle the bioinformatics hierarchical scalability should possess the knowledge representation capabilities described in Section 4.1, and be adaptable to the bioinformatics problem space. The current platform that most closely meets these requirements is the collection of standards and technologies known as *semantic technologies*, which are coordinated under a common integrated platform, known as the Semantic Web (Berners-Lee *et al.* 2001). It is envisaged that the Semantic Web will form a complex interlinked network of knowledge sources, traversed by intelligent agents capable of reasoning over knowledge gleaned from multiple sources. Although the Semantic Web holds much promise for biomedical discovery (Wolstencroft *et al.* 2005), it is currently only a

vision (Neumann 2005). However, semantic technologies today form a coherent infrastructure, which is being implemented and adopted progressively. Figure 4-1 shows that different technologies are currently at different stages of adoption by the IT industry: the low-level technologies used for the exchange of information are mature and established, while knowledge representation technologies are currently gaining acceptance. More abstract machine reasoning tasks are in the research phase and can only be prototyped at present.

A detailed description of the full technology stack is beyond the scope of this thesis; an introduction with a biological perspective is provided by Stevens *et al.* (2006). Here, we have highlighted some features of semantic technologies that are particularly applicable to bioinformatics scalability applicable to this work:

- **Structural independence**. The Semantic Web uses XML as the low-level encoding format. This allows records to be formatted as semi-structured files, which can be parsed by standard software, regardless of the vocabulary used, as discussed in Chapter 3.
- **Universal knowledge structure**. Knowledge is expressed using the Resource Description Framework (RDF) (Beckett 2004), which organizes knowledge into a simple sequence of statements, simplifying its structure. RDF knowledge is not structured as tables (as in relational databases), or hierarchical tree structures (as in plain XML documents), but is made up of statements of the simple <*subject*, *predicate*, *object*> form, which are joined into graphs (see Figure 4-2), and can be traversed by reasoning tools.
- **Knowledge extensibility**. RDF supports the *open world assumption*, which assumes that knowledge from any one source may be incomplete. Thus RDF supports the seamless aggregation of knowledge from multiple sources (see Figure 4-3).
- **Reasoning**. The simple structure imposed by RDF allows the application of *semantic rules* that analyze and manipulate the knowledge. These rules can be expressed in a variety of ways, and executed by standard software tools called *reasoners*. Often, semantic rules syntax consists of direct statements, and is arguably simpler to learn than the programming necessary to achieve similar results.

- **Ontologies**. Within the RDF framework, knowledge is represented using ontologies – shared domain-specific models and vocabularies (Bard *et al.* 2004). The OWL standard (McGuinness and van Harmelen 2004) supports the definition of domain classes (for example, sequences, genes, etc.), their properties (e.g. country of isolation), specific instances (e.g. the hemagglutinin protein), and *description logic* (DL) which describe their semantics (e.g. "a sequences can only have a single country of origin"). Like semantic rules, DL are processed by reasoners, and are used to validate the consistency of a knowledge model.

In summary, semantic technologies provide a complete platform for knowledge representation to suit bioinformatics tasks. Ontologies are defined to describe unambiguously specific domains of knowledge (for example, viral sequences), and provide a vocabulary for expressing knowledge as a series of RDF statements, encoded into XML files for portability. The receiving analysis tool can use and transform the knowledge by means of semantic rules, which can be built-in or user-generated. The simple, universal structure of RDF removes the need for structural transformation of information, shifting the emphasis to transforming *meaning* through ontologies. If the source RDF uses a different ontology from that of the tool transformations are possible by means of reasoning rules. New knowledge from analysis tasks can easily augment existing knowledge, since the Open World Assumption permits RDF to combine input and output (Figure 4-3).

ISOLATES table

Id	IsolateName	Country	Year	Host	Subtype
ISO1234	A/Goose/Guangdong/3/97	China	1997	Avian (Anatidae)	H5N1

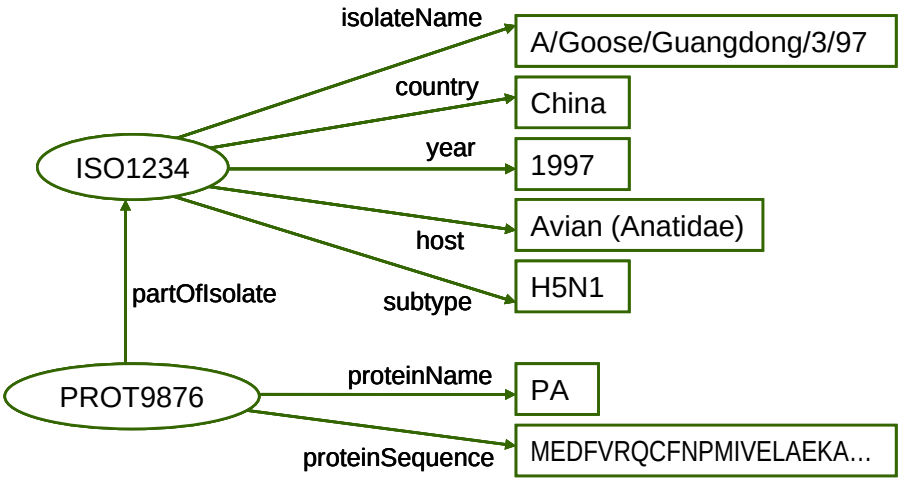
A

PROTEINS table

Id	IsolateId	ProteinName	ProteinSequence
PROT9876	ISO1234	PA	MEDFVRQCFNPMIVELAEKA...

ISO1234	isolateName	A/Goose/Guangdong/3/97
ISO1234	country	China
ISO1234	year	1997
ISO1234	host	Avian (Anatidae)
ISO1234	subtype	H5N1
PROT9876	partOfIsolate	ISO1234
PROT9876	proteinName	PA
PROT9876	proteinSequence	MEDFVRQCFNPMIVELAEKA...

B



C

Figure 4-2: RDF flattening of knowledge structure
In a relational database (A) data is organized in tables, each representing a different record type, with column contains the record properties; the table structures must therefore be known in order to interpret the data. When the same data is encoded in RDF, it is broken down into a series of simple <*subject*, *predicate*, *object*> statements (B), and processed in a uniform way, removing the need to understand table structure. The statements can be used to construct a graph which can be queried and traversed (C).

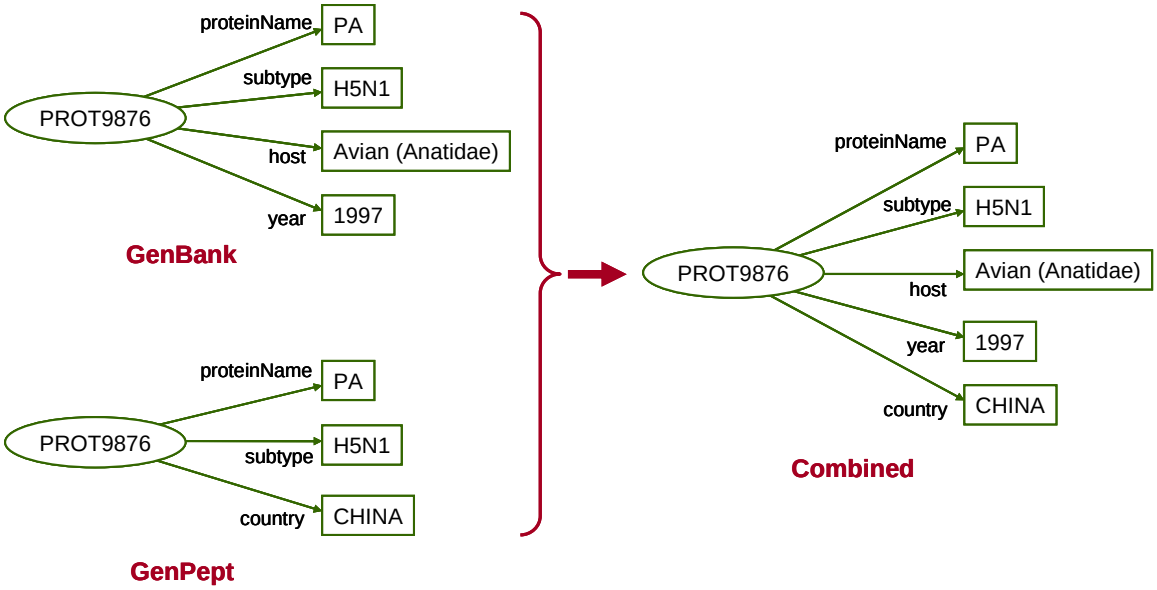


Figure 4-3: Open World Assumption in RDF
The RDF platform makes no assumptions about the properties of any given entity (such as the protein sequence in this figure). We can, therefore, collect RDF statements that describe the entity from multiple sources (two public databases in this example). The combined RDF statements form a new graph, in which information from different sources complement each other, capturing multiple perspectives about the entity.

Applications of the semantic technologies stack are still currently in their infancy. The combination of RDF, ontologies and rules can support the construction of a new class of analysis tools, which will be able to use in multiple forms of information: for example, metadata-aware phylogenetic tools could combine sequence similarity measures with date and country information, to characterize clusters of viral sequences. Only a small number of experimental tools offer semantic capabilities today. The Antigenic Variability Analyzer (AVANA) tool, presented in Chapter 5, is an example of such a metadata-aware tool, although it uses CSV, rather than RDF, as the metadata input format. This choice was motivated by the need to support users who want to create metadata from spreadsheet tools. This reflects the current lack of availability of RDF output from mainstream tools, which limits the widespread adoption of these technologies. We believe this is a temporary obstacle, which will diminish in importance as more tools emerge.

4.3 Improving metadata quality through semantic reasoning

Although semantic technologies are still in the adoption phase, and knowledge representation requires more progress in bioinformatics, it is already possible to show the direct benefits of applying semantic technologies to real-world bioinformatics problems. In the following sections, we will describe a study, performed on the large-scale influenza A dataset. The creation of this dataset is described in Chapter 3. Hereby we have demonstrated that semantic technologies can be effectively applied to improve the quality of descriptive metadata.

In the manual verification phase of the dataset construction, we observed that influenza sequences are often submitted to public databases as genomic sets: up to eleven protein sequences, produced from the same isolate, are simultaneously deposited by the same genome sequencing team. Frequently, only some of the records are fully annotated with provenance information, while the remaining records remain incomplete. ABK structural rules can aggregate annotations from the two databases, but are unable to fill in missing metadata in one protein record with metadata from a different protein record from the same isolate. In the absence of automated solutions such gaps must be filled during manual curation. The presence of such annotation gaps reveals an underlying knowledge representation mismatch between the reality *surrogate* being submitted, and the *ontological commitments* dictated by GenBank and GenPept. These databases manage sequences as the core entities, and it is therefore only possible to attach descriptive metadata to sequence records. In a genomic study, however, much of the metadata describes the whole genome (the isolate) rather than individual component sequences. For example, the values for year and country of origin, or host organism, must be the same for all sequences from the same isolate, since they are isolate properties rather than sequence properties. Rather than enter the same metadata for up to eleven separate entries, many submitters prefer to annotate a single protein record, and omit metadata from the other records. Even when they choose to complete the metadata entry for the whole set, repeated annotations may introduce manual errors, which in turn cause inconsistencies between protein records from the same isolates.

Our study investigated how semantic technologies could be used to apply reasoning on our dataset, to relate multiple sequences from the same isolate, verify their metadata consistency, and fill existing gaps. To address the semantic mismatch introduced by the sequence-oriented organization of the source databases, we developed a reasoning task to manipulate the RDF knowledge graphs, reconstructing the relationships between sequences and isolates. In other words, we were able to treat *isolates* as entities with descriptive metadata, which was derived from the metadata of the component sequences. The resulting model was validated using the description logic of a simple OWL ontology, to assess the quality of the restructured metadata, and determine the amount of manual curation needed. In the final process step, the curated isolate metadata was used to re-annotate the sequence records. Therefore, this process impacts the curation task in three ways: it finds inconsistencies in the extracted metadata; it transfers the manual curation process from sequence records to isolates (fewer in number); and it fills missing sequence metadata from isolate annotations. As we will show in the following sections, the ontology and semantic rules used in this task are remarkably simple, yet they can yield very useful results.

4.4 Materials and Methods

Descriptive metadata for the influenza A dataset (extracted from structural rules as describe in Chapter 3) was encoded in RDF format, using an OWL ontology. This ontology was specifically designed to suit our analysis needs, as no suitable standard ontology could be identified for this particular purpose. In the ontology, each sequence is represented by a resource of type *SequenceRecord*, which may posses any of the properties *proteinName*, *subtype*, *isolate*, *host*, *origin* and *year*, amongst others. Each of these properties is declared both as owl:DatatypeProperty (it can be assigned a literal value), and as owl:FunctionalProperty (it is single-valued, since multiple values would be inconsistent). Another type of object, *IsolateRecord*, is defined to represent individual isolates associated to one or more sequences. To facilitate the semantic restructuring task, properties *subtype*, *isolate*, *host*, *origin* and *year* can also be applied to *IsolateRecord* objects.

The metadata extraction task produced an RDF graph comprising thousands of *SequenceRecord* resources, with associated extracted metadata and references to their source documents (Figure 4-4A). This model reflects the relationships that exists between sequences and their properties in the source database, where records from the same isolate are not connected from each other. We restructured the RDF graph by reconstructing *IsolateRecord* objects associated to the *SequenceRecord*s. Since most sequence properties (except for *proteinName*) are also isolate properties, their values were attached to the *IsolateRecord*, producing a restructured graph (Figure 4-4B). This restructuring task was effected by simple semantic rules, executed by Jena2 (McBride 2002), which is also used by ABK for RDF data storage. For convenience, semantic rules were specified in the rule language of Jena's built-in reasoner. However, the same rules could be defined in other semantic rule languages, such as SWRL (Horrocks et al. 2004), or the future standard language RIF (Hawke 2005).

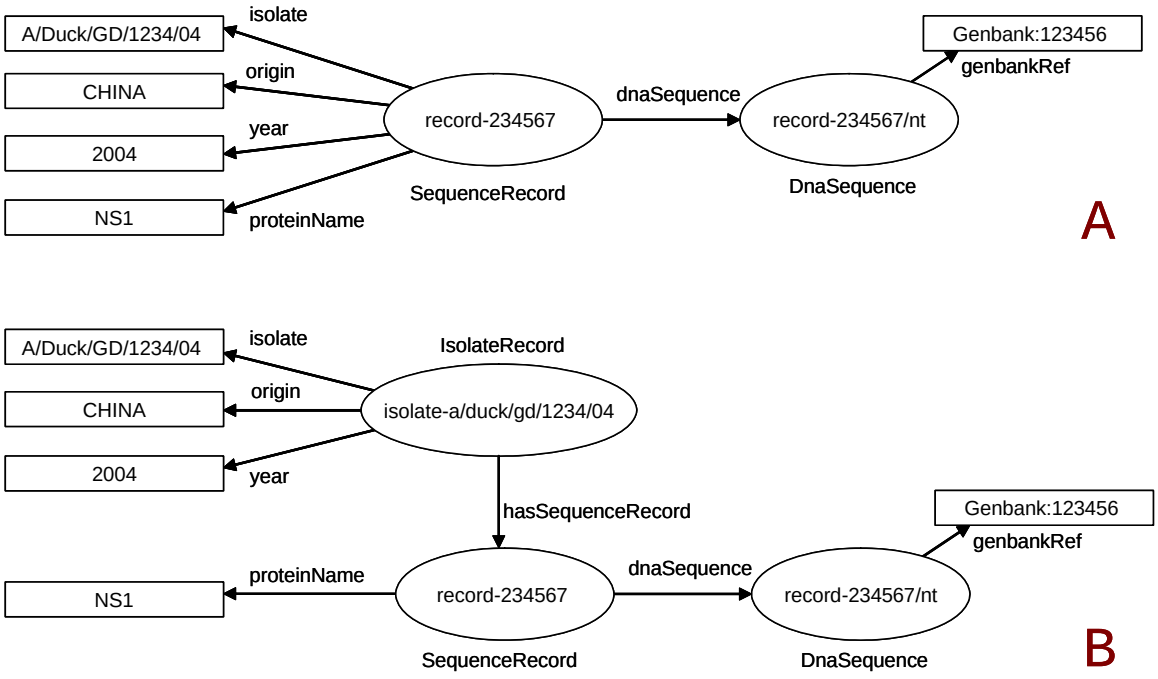


Figure 4-4: Restructuring sequence metadata
Graph A shows the relationship between SequenceRecord resources, their metadata properties, and a source GenBank document, as encoded by ABK in its RDF output. In this example, records belonging to the same isolate have no relationship to each other. Graph B shows the same knowledge, restructured by the introduction of the IsolateRecord resource, and the transfer of isolate-specific metadata.

The two semantic rules used for the restructuring task are shown in Figure 4-5. The first rule identifies *SequenceRecord* objects that possess an isolate name, creates a URI (unique identifier) based on a normalized form of that isolate name, and ensures that an object of type *IsolateRecord* assigned that URI is attached to the *SequenceRecord* object. The second rule copies the desired metadata properties to the *IsolateRecord* object, whenever they are found in a *SequenceRecord*. The `oneOf()` built-in function, which matches a property type against a list, was created using Jena's extension mechanism. Isolate normalization was necessary in rule1, since isolate naming is often inconsistent (for example, "A/HongKong/123/04", "A/hongkong/123/04" and "A/Hong Kong/123/04"). The function `normalizeIsolate()` was thus added to remove all whitespace and special characters (except for slashes) from isolate names, and convert them to lowercase. Although this normalization did not solve all inconsistencies, it resolved naming defects in hundreds of records.

```
[rule1: (?rec rdf:type   vg:SequenceRecord)
        (?rec vg:isolate ?isolateId)
        normalizeIsolate(?isolateId, ?nIsoId)
        uriConcat('urn:abk:isolate:', ?nIsoId, ?isolateUri)
        ->
        (?isolateUri rdf:type vg:IsolateRecord)
        (?isolateUri vg:hasSequenceRecord ?rec)
]

[rule2: (?isolateUri vg:hasSequenceRecord ?rec)
        (?rec ?prop ?value)
        oneOf(?prop, vg:isolate, vg:virusSubtype, vg:year,
                                 vg:country, vg:hostOrganism)
        ->
        (?isolateUri ?prop ?value)
]
```

Figure 4-5: Semantic rules used for the metadata restructuring task

The inferences from semantic rules were validated against the OWL ontology using Jena's OWL DL reasoner, which identified all cases in which the inferred isolate metadata violated the ontology's description logic constraints. It identified all isolates which received conflicting metadata from their sequence records, and therefore were assigned multiple values for their

functional properties, as shown in Figure 4-6. The validation task reported all such inconsistencies, which were then resolved manually by a curator. In the final processing step, another simple semantic rule (shown in Figure 4-7) was executed to re-annotate the sequence records: for every *SequenceRecord* associated to an *IsolateRecord*, the *IsolateRecord* properties were copied to the *SequenceRecord*. This ensured metadata consistency for sequences derived from the same isolate, and transferred all isolate metadata corrections to the sequence records, thus reducing the necessary manual curation effort.

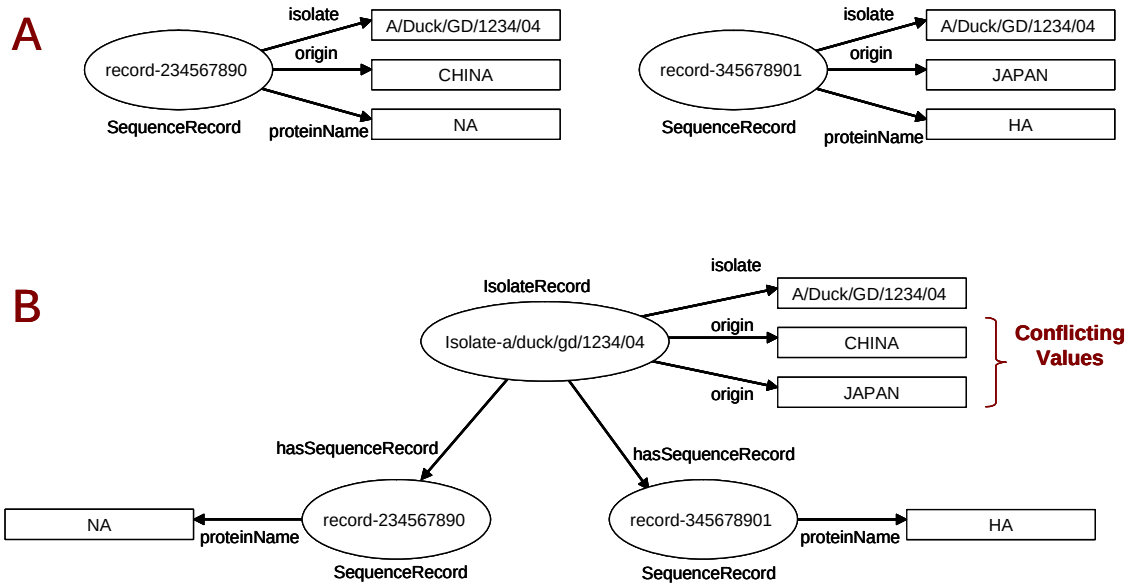


Figure 4-6: Identification of conflicting metadata values
Sequences from the same isolate should have identical value for certain metadata properties, such as *origin*. However, inconsistencies often occur, as shown in (A). Rule-based metadata restructuring transfers the inconsistent values to the IsolateRecord resource, as shown in (B). Since origin is declared as a functional property, an OWL reasoner can identify the inconsistency as a breach of the ontology DL constraint.

```
[rule3: (?isolate rdf:type             vg:VirusIsolate)
        (?isolate vg:hasSequenceRecord ?rec)
        (?isolate ?prop                ?value)
        oneOf(?prop, vg:isolate, vg:virusSubtype, vg:year,
                                 vg:country, vg:hostOrganism)
        ->
        (?rec ?prop ?value)
]
```

Figure 4-7: Semantic rule used for re-annotation of sequence records

4.5 Results

Reasoning was applied to all records that had an *isolateName* property value (38,474 records), producing a total of 7,640 distinct isolate records each being associated to one or more (up to eleven) sequence records. Figure 4-8 shows the distribution of isolates according to the number of sequences linked to the isolate. The predominance of isolates associated to 10 or 11 protein sequences, accounting for about 63% of all sequences, indicates that most sequence records were submitted by full-genome sequencing studies (older genome sets only include 10 proteins, due to the relatively recent characterization of the PB1-F2 protein). At the other end of the scale, about 12.5% of sequences belong to isolates represented by only one or two sequences, usually submitted by studies that focus on one or two proteins (hemagglutinin and neuraminidase are more intensely studied than any other influenza proteins). Several individual sequences could not be associated to the correct isolate, because of errors in isolate name that could not be corrected by our name normalization task (e.g. misspellings). Finally, 4.8% of sequences were associated to isolates with more than 11 protein sequences. This is due to artifacts from sequences used in multiple studies and resubmitted to the databases, sometimes as fragments of the original sequences. Identifying of such duplicates is not a simple task with the rule language we used, because of limited string processing capabilities.

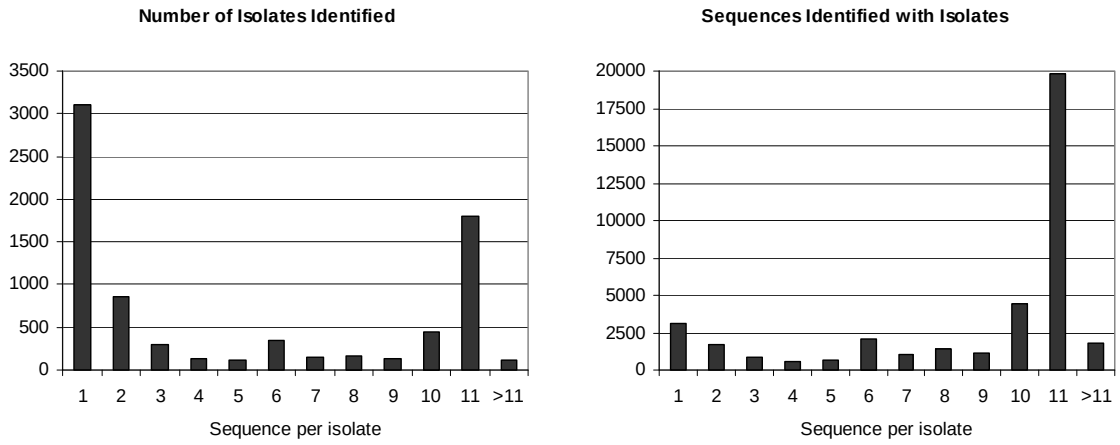


Figure 4-8: Associations of sequences to isolates
The left chart shows the number of identified isolates, according to the number of sequences they are associated to. On the left, we have shown the distribution of sequences according to the number of sequences associated with their isolates.

Isolate metadata, inferred by the reasoner by applying semantic rules, was subsequently validated against the OWL ontology by an OWL DL reasoner, which identified a number of errors and inconsistencies. Multiple variants of *isolateName* were found for 388 isolates, most often due to upper/lower case differences; 98 isolate names contained additional symbols, such as spaces or dashes. For the *subtype* property, 22 isolates were reported as conflicting. In 13 cases, we found that the same name had been used for two separate isolates, which required manual separation; in the remaining cases, one or more sequences were ambiguously annotated and had to be discarded. The majority of the 22 isolates with multiple *host* values contained values of different specificity (e.g. "AVIAN" and "DUCK"), which demonstrated once more the inconsistent standard of annotation. Similarly, 28 of 70 issues identified for *origin* were conflicts between overlapping regions (e.g. "CHINA" and "HONG KONG"). More importantly, the *origin* annotation had to be manually verified for all protein sequences isolates from turkeys, since the host organism was often confused with the country of origin: 181 isolates were inspected and manually corrected. Although corrections were substantial in number and complexity, the advantage of our approach is that isolate metadata corrections are back-propagated to multiple sequences, thus significantly reducing the manual curation effort.

Following manual curation, sequences were automatically re-annotated by semantic rule-based reasoning; the results are summarized in Figure 4-7. Sequence re-annotation affected more than 1200 records, focusing on filling in the gaps and correcting errors. Although the numbers of records may seem small (2-3% of the total), manual curation is time consuming, tedious and error-prone, and these results translate to a significant impact for the curation workflow. It is notable that only 70.1% of isolates were annotated with the *host* property, a lower percentage than available in sequence record annotations. This indicates that full-genome submissions tend to contain more complete annotations, probably because they are produced by large sequencing or surveillance studies, with stringent quality guidelines (Ghedin *et al.* 2005).

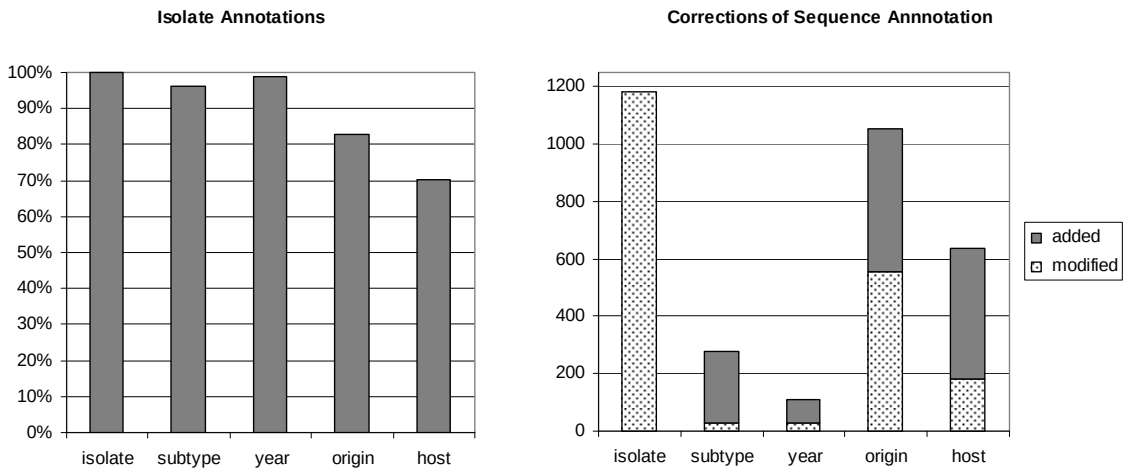


Figure 4-9: Isolate annotation and resulting corrections
The left chart shows the percentage of created IsolateRecord objects with a value for each of the five properties. For the *host* and *origin* properties, the low yield of isolate annotations would indicate that isolates with a full complement of proteins (10 or 11 sequence records per isolate) are generally better annotated than isolates with a small number of sequences. The chart on the right shows the number of property values that were automatically modified (or added, in the case of sequence records for which structural rules did not yield a value).

4.6 Discussion

The contribution of RDF, semantic rules and description logic affects a smaller proportion of records, but produces sophisticated and fully automated results, reducing the effort required for time-consuming and error-prone manual annotation.

Alternative approaches to metadata restructuring and quality validation could be used: the use of a relational database, appropriate queries and string manipulation could reconstruct the viral isolates and identify and correct inconsistencies. Such an approach, however, require a non-trivial programming effort, and significant infrastructure (such as running a database), beyond the skill of most biological researchers. Semantic technologies use a simple, file-based infrastructure, and a very flexible way of defining schemas with RDF and OWL. Although our experiments required a certain amount of programming, all domain-specific functionality was embedded into the ontology and the rules employed, indicating that generic tools could support this class of task, leaving biologists the flexibility of structuring metadata according to their needs. In addition, the relative simplicity with which semantic reasoning

rules are specified adds utility to this approach. Many software applications (such as email clients or network firewalls) provide user interfaces for expressing rules of various kinds. We have shown that it is possible to provide similar intuitive mechanisms to support sophisticated rule-based data preparation and cleaning tasks using molecular database records.

The conversion of primary public data repositories to RDF has been advocated by proponents of the Semantic Web vision and even prototyped for a small number of databases- for an example, see the UniProt-RDF project (Jain 2007). Our results, however, suggests that a straightforward format conversion would not solve the more fundamental semantic heterogeneity issues, whose causes are found in data submission practices that sacrifice quality to achieve greater scalability. Since NCBI sequence records are submitted by researchers, without a curator as an intermediary, different interpretations of metadata field meanings give rise to discrepancies. Even if these process defects were addressed, the metadata structure imposed by large primary data repository is unlikely to match the individual needs of different analysis tasks. This fundamental issue explains the emergence of a vast number of smaller-scale "boutique" databases in recent years, offering richer and more highly curated metadata, while the structure of large primary databases has remained substantially unchanged. Since small specialized databases are often the result of manual annotation of primary sources such as GenBank, the RDF-encoded metadata output of our knowledge aggregation tasks seems highly suitable as a data warehousing product. In other words, it might be more useful to provide simple mechanisms for researchers to make their high-quality metadata available in a versatile format such as RDF, than to try to convert large and mature primary repositories. Such capillary supply of well-curated metadata could fuel a "grassroots" level adoption of semantic technologies, especially once trust and provenance concerns are addressed (Stevens *et al.* 2007). In turn, this could drive the development of analysis tools that understand RDF metadata and can integrate it in the analysis process.

The stack of semantic technologies is developing gradually. The foundation layers such as XML and RDF are solid and well-understood, while reasoning capabilities are still in the various stages of deployment and present early adopters with scalability concerns. We have

found that our simple semantic rules, when applied to tens of thousands of records, cannot be executed on-line within reasonable waiting times on a current fully-featured desktop computer. For certain tasks, we were able to increase performance dramatically with a divide-and-conquer approach, by splitting the input data into separate files of around 6,000 records each. However, this approach is only viable for tasks that do not require reasoning over of multiple interlinked resources. Although performance gains may also be achieved by choosing alternative data storage and programming platform options, scalability issues eventually emerge, given sufficiently complex reasoning demands. At present, more research is needed to address some of these scalability issues: for example, to assess whether a dataset is suitable to be broken down into smaller datasets for a given reasoning task. Ontology complexity is a major area where improvements can be achieved: large ontologies contain a vast number of DL semantics, which cause numerous reasoning operations to be executed even if they are not needed for the final result. “Right-sizing” ontologies, to suit the problem in hand, can mitigate these problems. Scalability issues are a sign of the relative immaturity of the semantic technologies platform, and we expect they will be successfully addressed, as they have been for other integrative platforms.

4.7 Conclusion

In this chapter, we have reviewed semantic technologies as a knowledge representation layer for biological knowledge mining. We have shown that the technology stack comprising XML, RDF and OWL allows flexible and extensible encoding of knowledge, and therefore supports the flow and augmentation of knowledge necessary for biological knowledge mining (see Chapter 2, Section 2.4). The study presented in this chapter showed that semantic rules are a powerful addition to semantic knowledge representation, and are capable of restructuring and extending existing knowledge through the application of simple reasoning task. With a small number of relatively simple semantic rules, we were able to restructure our large-scale influenza A dataset, filling in large proportions of missing data. These results demonstrate that semantic technologies are expressive and powerful, and are therefore good candidates for

the knowledge flow backbone of bioinformatics pipelines. Their ongoing standardization will also mean that tools will be able to integrate these technologies using standard software libraries. However, we have also highlighted that semantic technologies are still evolving and currently present issues of scalability and ontology modelling, which will need further research before full adoption can take place. For these reasons, the knowledge-enabled tools ABK (Chapter 3, Section 3.3) and AVANA (Chapter 5, Section 5.3) presented in this thesis make limited use of semantic technologies.

5. INFORMATION THEORY-BASED SEQUENCE ANALYSIS

In this thesis, we have chosen to demonstrate the power and utility of the biological knowledge mining approach through three real-life research applications in immunology and virology. Two of these applications are studies of the variability of pathogens from different perspectives: a comparative study that identifies systematic differences between groups of sequences (Chapter 6), and a meta-analysis that identifies sequences that are highly conserved in multiple significant groups of sequences (Chapter 7). Both of these applications showcase important aspects of "second-generation" bioinformatics analysis: they require *metadata-enabled tools* to define the groups of sequences to be analyzed, and they must support *quantitative scalability* (see Chapter 2, Section 2.3.3).

We have developed two novel methods based on *information theory*, which we have applied to the studies of viral diversity in Chapters 6 and 7 of this thesis. In the current chapter, we have discussed their mathematical foundations, as well as the limitations and statistical corrections required by these measures The first method (Section 5.1), measures *peptide entropy* a measure of diversity of *n*-mer peptides, which will be used in the identification of conserved, antigenically significant peptides in Chapter 7. Whereas conservation measures usually focus on the conservation of individual residue positions in protein alignment, peptide entropy accounts for the variability of neighbouring positions, providing a more realistic measurement of sequence variability, as viewed from the perspective of HLA molecules. The second method (Section 5.2) uses *mutual information* to measure the relationship between a mutation and the set of sequences in which it occurs most frequently, identifying *characteristic mutations* for specific sequence sets. In Chapters 6, we have applied this method to the identification of mutations that have permitted the adaptation of influenza A viruses to human-to-human host transmission.

Information theory has previously found several applications in bioinformatics (Gatenby and Frieden 2007), and was selected for our applications because of its relatively simple computational requirements. Due to the statistical nature of entropy and mutual information,

the time necessary for computations increases linearly with the number of sequences, while memory requirements remain practically constant. As a result, the AVANA tool (presented in Section 5.3), which implements the methods described in this chapter, was able to handle the analysis of thousands of sequences in real time, using current standard desktop computer hardware, showing that information-theoretic measures are an excellent choice for scalable studies of variability. The AVANA tool is knowledge-enabled, in that it is able to use descriptive metadata in order to organize sequence sets in comparative analysis and meta-analysis. The metadata handling capabilities of AVANA enabled important results to emerge from our applications, demonstrating the power of extending analysis tools with knowledge capabilities.

5.1 Information Entropy

In information theory (Shannon 1948), entropy is a measure of the *randomness* of a given measurement, and thus of its variability. Entropy is defined in terms of a discrete random event x, for which all possible outcomes are included in the set $E = \{e_1, e_2 \ldots e_n\}$:

$$H(x) = -\sum_{e \in E} p_e \log_2(p_e) \qquad \text{(i)}$$

where p_e is the probability of event $e \in E$ occurring. In the context of an alignment of homologous sequences, it is assumed that residues aligned at the same position tend to occupy the same structural position, and entropy can therefore be used to measure the variability of the residues (or peptides) that are present at that position. In this case, equation (i) is modified to measure *peptide entropy* $H(x)$ at any given position x of the alignment as follows:

$$H(x) = -\sum_{i=1}^{n(x)} p(i,x) \log_2 p(i,x) \qquad \text{(ii)}$$

where $p(i,x)$ is the probability of a particular peptide variant i being centered at position x. Here the term *peptide* is used as a generalization, which encompasses amino acid residues (which can be considered peptides of length 1), or amino acid strings of arbitrary length n. The entropy value increases as $n(x)$, the total number of peptide variants observed at position

x, increases. Entropy is also sensitive to the relative frequency of variants, such that it decreases when one variant is clearly dominant (*i.e.* the position is conserved). $H(x) = 0$ denotes a site with 100% conserved residues.

Entropy is measured in *bits*, and its scale is determined by the number of possible outcomes. If the random event *i* has $n(x)$ possible outcomes, the maximum possible entropy value is that associated with total randomness, where each of the outcomes occurs with equal probability. Under these conditions, the entropy value is $H_{max}(x) = \log_2 n$. In a multiple alignment of protein sequences, there are 20 different possible amino acid residues that can be observed in theory at every position; the theoretical maximum residue entropy at any position is therefore $\log_2 20 = 4.233$. When measuring the entropy of peptides of length *m*, there are 20^m peptides that can be formed by combining 20 different amino acids over *m* neighbouring positions, and the theoretical maximum peptide entropy is therefore $\log_2 20^m = m \log_2 20$, that is approximately m*4.322. In practice, these theoretical limits are never reached: extreme variability is unlikely in sets of closely related sequence sets, and the very purpose of alignment algorithms is to minimize position diversity in the alignments.

Information entropy can be interpreted in a variety of ways, and two such interpretations are particularly meaningful for the analysis of alignments of homologous sequences. On one hand, entropy gives a *descriptive* measure of the variability. In this sense, entropy is a measure of the heterogeneity of results that are obtained when sampling a population, with high entropy indicating that any given observation has poor generality (Martín and Rey 2000). On the other hand, entropy can be viewed as a *predictive* property – a measure of *surprisal* of an observation (Tribus 1961). Under this interpretation, low entropy indicates stability, such that future outcomes are likely to be identical to past outcomes; conversely, high entropy means that future outcomes are difficult to predict. The predictive interpretation is interesting from an evolutionary perspective, since low entropy may indicate evolutionary constraints that limit the occurrence of mutations in particular regions of a sequence.

As an enabler of quantitative scalability, information entropy offers improvements in terms of computational speed and memory requirements. Since entropy is a probability-based

measurement, the time taken for computation increases linearly with the number of sequences, while the memory requirements remain constant. As a metric, information entropy combines in a single value two dimensions of diversity: the number of observed outcomes and their relative frequency. The statistical nature of entropy presents an additional advantage: it can be used in alignments that contain partial fragments as well as full-length sequences. If a sequence does not extend over a particular position, it does not contribute to the entropy computation at that position. As a result, entropy values at different positions have different statistical support (i.e. the count of sequences used in the entropy computation).

5.1.1 Residue Entropy and Peptide Entropy

When describing the variability of a pathogen, researchers often focus on single residue mutations. Immunological mechanisms, however, involve short peptides rather than individual residues. In T cell-mediated immune responses, HLAs bind to short peptides, typically between 8- and 20-amino acid long, with 9 amino acids being the predominant length of class I peptides and the core of class II peptides (Rammensee 1995). In immunological studies that investigate pathogen variability, it is therefore appropriate to consider peptide diversity rather than residue diversity. Some studies have attempted to approximate peptide entropy by averaging residue entropy over a sliding window of the desired size (Yusim *et al.* 2002), but such an approach is likely to underestimate the peptide variability measure. The variability of individual residue positions has a combinatorial effect on peptide entropy. In other words, even limited variability at multiple neighbouring positions can produce a very diverse set of resulting peptides, unless the variants at the different positions co-occur (i.e. they are observed in the in the same sequences). In any case, it is clear that peptides must be at least as diverse as their constituent residues, and peptide entropy can therefore never be lower than the residue entropy of any of the spanned positions. When averaging over multiple positions, the effect of high entropy positions will be mitigated by any conserved residues, and therefore the region's diversity will be underestimated.

To obviate these problems, and obtain a true reflection of peptide diversity, true peptide

entropy should be computed at each position. This is done by applying equation (ii), considering each peptide of length *m* occurring at that position as a separate outcome. Since each position corresponds to a single residue, we define the position of a peptide as the position where the central residue in the peptide is located. For a peptide of length *m*, this residue occupies position *m*/2 or (*m*-1)/2 within the peptide, depending on whether *m* is even or odd, respectively. Theoretically, the peptide can be any one of 20^m possible *m* amino acid combinations, and the maximum entropy is therefore $H_{max} = m \log_2 20$. In practice, because of the similarity of aligned sequences, such extreme values are never reached. In our work on several viruses, we found that 9-mer entropy values rarely exceed the range 0.0-5.0, with the highest peaks (≈ 8.0) observed in the most variable regions of the HIV proteome. Considering peptide variants rather than residue variants is a relatively straightforward change, yet we found no alignment tools that supported this computation. The AVANA tool, presented in Section 5.3, was developed to fill this gap.

5.1.2 Alignment Gaps in Entropy Computation

Multiple sequence alignments tend to contain alignment gaps that represent evolutionary events, such as the insertion or deletion of one or more nucleotides, which affect both nucleotide and amino acids sequences. Alignment gaps are introduced and positioned based on scoring schemes, which use gap penalties and mutation matrices to determine the most likely location of the gaps. In the case of protein alignments, both 3-D and secondary structures define the proteins, and gaps must therefore be regarded as notional, since the physical conformation of the protein is not interrupted at the gap point. This poses a challenge for entropy computation: a gap position does not contain an amino acid residue to the used as the centre of the peptide, yet the gap is only notional, and the sequence does contain a peptide at that point in the protein conformation. Therefore, we cannot simply discard gapped sequences from entropy computations at a given position, since this operation removes some of the diversity at that position, artificially lowering the entropy value. Generally, gaps indicate higher variability and tend to be associated with high entropy values. Empirical

methods for estimating entropy at highly-gapped positions have been suggested (Pei and Grishin 2001), but we have found them very approximate, and not appropriate for application with small sequence sets.

As previously discussed, entropy can be computed on alignments that contain sequences of different lengths. When aligned, sequences fragment may start and/or end at positions other than the endpoints of the alignment. To be able to store such fragments in a file, and read them back correctly aligned, alignment tools and sequence editors "pad" the sequence with padding symbols at the start and end positions, so that the all sequences are of the same length. Such sequence padding often uses the dash symbol ('-'). Unfortunately, this symbol is also used to indicate alignment gaps, and therefore several programs do not distinguish between alignment gaps and padding. However, this distinction is important in entropy computation: a padding symbol indicates that the residue at that position is unknown, while a gap signifies that it is known that the residue is not present. In our computations we discard all padding, such that the statistical support at an alignment position is the count of sequences that do not contain a padding symbol at that position. Conversely, gaps are accounted for in our entropy computation, rather than discarded. For any sequence s, the *m*-mer peptide centered at position *x* (*m* is set by the user) is constructed according to the following rules:

1. The central residue symbol corresponds to the residue symbol at position *x* (either an amino acid or an alignment gap).
2. The (*m*/2)-1 symbols preceding position x are determined by scanning the sequence to the left of position *x*, ignoring gapped positions.
3. The (*m*/2) symbols following position x are determined by scanning the sequence to the right of position *x*, ignoring gapped positions.
4. If a padding symbol is encountered either at position x, or during the scans in steps 2 and 3, the sequence is discarded from peptide entropy computation at position *x*.
5. If *residue entropy* (rather than peptide entropy) is being computed, the alignment gap is considered to be the residue at position *x* (i.e. it is treated as an additional amino acid).

These rules ensure that peptides containing a gap are treated distinctly from those that present a valid residue; however, they do not treat gaps as additional amino acids, since multiple gaps are discarded (if we accept that gaps are notional, it makes little sense to distinguish a single gap from a sequence of multiple gaps). Figure 5-1 shows how the presence of gaps in a sequence is handled by our method.

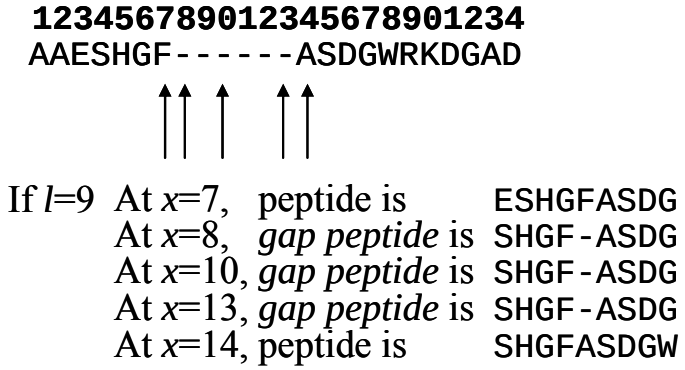


Figure 5-1: Determination of 9-mer peptides at various positions in a gapped sequence, to be used in entropy computations.
The term *gap peptide* is used to label a notional peptide at a gapped position. The presence of gap peptides prevents the artificial decrease in entropy values caused by excluding sequences containing gaps from the entropy calculation.

Even though the above rules ensure a credible estimate of entropy at positions with a moderate quantity of gaps, it is harder to interpret entropy values for positions with a high percentage of gaps, especially when the majority of sequences do not contain a residue at those positions. For residue entropy calculation in particular, a high percentage of gaps will tend to gradually reduce the entropy value. For these reasons, a *maximum gap threshold* is chosen (for example, 50%), above which entropy computation should not be regarded as reliable, and therefore removed from the alignment.

5.1.3 Set size considerations in Entropy Computation

Information entropy is a statistical measure, and its theoretical premises assume that the set over which it is measured is of infinite size (i.e. it is computed from an infinite number of sequences). In practice, sequence sets are generally samples of a larger population, and are

likely to contain a subset of all the variants that are present in the source population. This leads to an artificial reduction of the variability, and hence of the entropy value; in addition, smaller sequence counts increase the entropy estimation error. The effects of *alignment size bias* are especially noticeable for alignments containing fewer than about 100 sequences, and must be accounted for when making direct comparisons between sequence alignments of different sizes. To illustrate this effect, Figure 5-2 shows the distribution of entropy values measured at four different sites (chosen as representative of different entropy levels) in an alignment of Influenza A PB2 proteins. The values were obtained by constructing multiple subset alignments, by randomly subsampling sequences from the master alignment. For alignments with a relatively large number of sequences (250 sequences, graph A), the distribution of the entropy values obtained is narrow, and the four sites are well-separated. As sequence count decreases (509 sequences in graph B and 20 in graph C), the average entropy estimation error increases, to the point that it becomes hard to distinguish between different entropy levels.

Alignment size bias causes entropy to decrease with set size. It has been shown that, for an alignment of N sequences (Figure 5-2), size bias is proportional to $1/N$ (Paninski 2003; Slonim 2005). Leveraging on this relationship, we can correct for size bias, by applying to each alignment a statistical adjustment that estimates entropy values for an infinitely-sized alignment with analogous variant distribution. To obtain such estimate, the alignment is repeatedly randomly sampled to create smaller alignments of varying size, whose entropy can be measured. At each alignment position, the entropy of these subset alignments of size N can be plotted against $1/N$, using a linear regression to extrapolate the entropy estimate for $N \rightarrow \infty$. We have found in practice that this adjustment, applied on alignments of the eleven influenza A protein sequences for different subtypes, produces regressions with a a very high *coefficient of determination* ($r^2 > 0.9$ in most cases) which was used as a goodness-of-fit of the estimates, confirming the validity of this adjustment. The chief advantage of infinite-set extrapolation is that it produces entropy values that can be used for direct comparisons of sequence alignments with different sequence counts (Khan *et al.* 2008). However, that this adjustment

cannot address other sampling bias errors, such as excessive sampling of a narrow pool of sequences, which often compounds errors caused by set size bias.

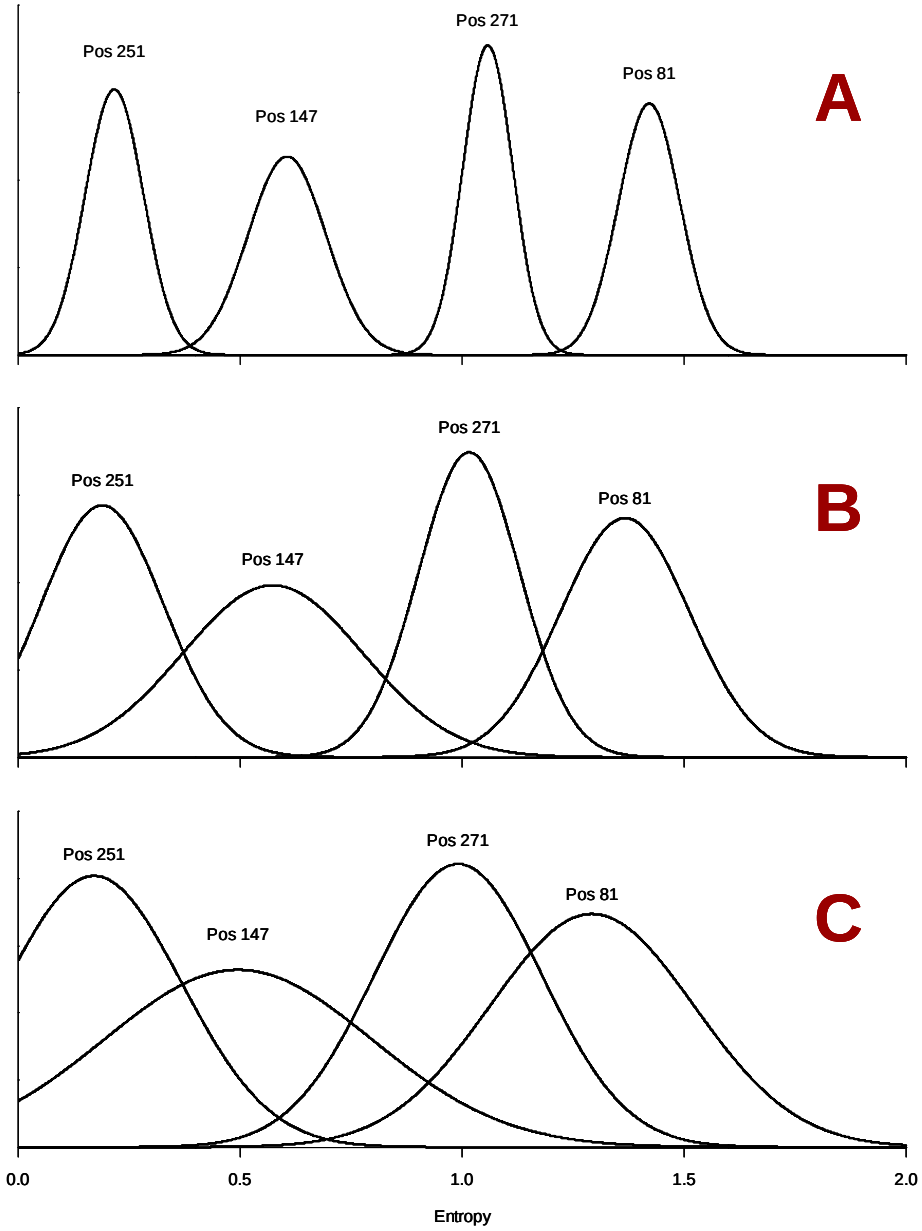


Figure 5-2: Effect of set size on information entropy.
The *probability density* of entropy values at four sites of the Influenza A PB2 proteins is plotted for alignments of decreasing sequence count N (graph A: N=250; graph B: N=50; graph C: N=20). For each graph, we constructed 200 random alignments of the required size from the PB2 master alignment. The mean and standard deviation of measured entropy from these alignments were used to plot the normal probability distributions shown in this chart. The entropy values for different sites are well-separated in large sequence sets (plot A) while the likelihood of distinguishing medium-entropy sites from high- or low-entropy sites drops dramatically at low sequence counts (plot C). The sites were selected based on their equally-spaced entropy values.

5.2 Mutual Information as a Comparative analysis tool

Entropy computations can be combined to determine relationships between pairs of variables (Shannon 1948). When considering two discrete events *A* and *B*, one can measure the *mutual information* (MI) of the two events as follows:

$$MI(A,B) = H(A) + H(B) - H(A,B) \qquad \text{(iii)}$$

where $H(A,B)$ is the *joint entropy* of the two variables, which is computed using equation (i), replacing *E* with the set of all unique pair of values (A,B). MI is interpreted as the *reduction in uncertainty* of the outcome of B when the outcome of A is known, and thus a measure of the dependence between the two variables. It has been shown (Steuer et al. 2002) that MI is 0 for two fully independent variables, while the MI of two variables that are fully co-dependent is determined by the entropy of the variables:

$$0 \le MI(A,B) \le \min\{H(A), H(B)\} \qquad \text{(iv)}$$

MI has been used in mapping of genes and clustering of genetic markers (Dawy et al. 2006). It has also been employed to identify pairs of co-evolving sites in proteins, which produce high MI values when individual and joint residue entropies are combined using equation (iii) (Martin et al. 2005).

5.2.1 Identification of Characteristic Sites and Characteristic Variants

In this thesis we use mutual information to identify mutations that characterize sets of sequences. Specifically, we seek to identify mutations that allow certain influenza A strains the capability to be transmitted between humans. Conservation analysis is often used as a tool for identifying functional residues (Valdar 2002; Nobrega and Pennacchio 2004), and we extend this method to the identification of functional components that confer specific properties to a pathogen population. We expect these mutations to be strongly conserved in virus populations for which the mutation is functionally necessary, and uncommon in populations that are do not exhibit the properties conferred by the mutations (in our case, human-transmissible and avian influenza strains, respectively).

Such functionally important mutations can be found by comparing a *characterized set* of sequences (sequences selected on the basis of a common property), against a *reference set* (the pool of sequences that do not possess this property). This comparison can identify one or more *characteristic sites*: sites that exhibit residues (which we will refer to as *characteristic variants* or *characteristic mutations*) which are common in the characterized set, but rare in the reference set, and are therefore likely to participate in conferring the defining property of the characterized set. The study of mutations by methods based on information theory has previously been explored; these analysis methods tend to be based on identifying sites that exhibit an entropy differential between the two groups. Korber *et al*. (1994) demonstrated the benefits of comparing the *information entropy* of separate alignments of HIV protein sequences, sampled from blood and brain tissues. They identified sites which were highly conserved (lower entropy) in the brain but not in blood, suggesting that the virus had forgone mutations to adapt to brain tissues. Because of reliance on entropy differential, only sites characterised by high diversity in blood isolates were selected; the method is not capable of identifying sites which are conserved in blood isolates, but acquire mutations as a result of tissue adaptation.

Mutual information can be used to identify characteristic sites in sets of aligned sequences. We compare pairs of *homologous alignments* to measure the relationship between the amino acids residues observed at a site, and the alignment in which they are observed. In a pair of homologous alignments, every residue site *n* in one alignment aligns with the same site *n* in the other alignment. In practice, pairs of homologous alignments may be formed by extracting sets of aligned sequences from a master alignment, without further realignment. Thus, variables *A* and *B* in equation (iii) are replaced with the observed residue *a*, and the label *S* of the set (alignment) within which the residue is observed. The MI at a site *x* is therefore computed by:

$$MI(x) = H_a(x) + H_S(x) - H_{S,a}(x) \qquad \text{(v)}$$

$H_a(x)$ is simply the entropy at site *x* for the merged alignment, computed using equation (ii).

$H_S(x)$ is derived from the number of sequences in each of the two sets (n_1 and n_2):

$$H_S(x) = -\frac{n_1}{N}\log_2\left(\frac{n_1}{N}\right) - \frac{n_2}{N}\log_2\left(\frac{n_2}{N}\right) \qquad \text{(vi)}$$

where $N = n_1 + n_2$. Finally, $H_{S,a}(x)$ is given by:

$$H_{S,a}(x) = -\sum_S \sum_{a \in A} p(S,a)\log_2 p(S,a) \qquad \text{(vii)}$$

where $p(S,a)$ is the probability of any given combination of residue and set label (in other words, occurrences of the same amino acid in two different sequence alignments constitute distinct outcomes).

Characteristic sites present different residues in the two sets, highly conserved within each set. Therefore, there is a strong relationship between residues and set labels at these sites, resulting in high MI values. Conversely, sites with low MI (e.g between 0.3 and 0) exhibit similar distributions of amino acid variants in the two sets and are not considered to be characteristics. Since there are exactly two sets, the upper bound of $H_S(x)$ is 1, the maximum entropy for a variable with two outcomes. From equation (iv) we therefore infer that $0 \le MI(x) \le 1$. However, $H_S(x)=1$ only when both alignments are equal in size, and the range of $MI(x)$ decreases as one set becomes larger than the other.

Although a high MI value is the primary requisite of a characteristic site, the selection process must take into account a variety of factors that affect MI. Firstly, it is reasonable to assume that the mutations that characterize a specific subset may circulate in a limited proportion of the larger reference pool. In influenza A, one may expect some mutations that are highly conserved in human-transmissible strains (and thus likely to be involved in conferring transmissibility) to also emerge in the highly diverse avian pool. Therefore, the requirements for conservation in the human set are more stringent than those in the avian set; in addition, the presence of human variants in the avian pool lowers the MI value. These considerations are problem-specific, and require background knowledge of the pathogen's biology. Another factor that affects MI is the presence of "noise" mutations: sporadic random mutations and sporadic episodes of infections transmitted from other hosts, which we

observed in both sets. Finally, characteristic mutations may emerge gradually in the evolution of the pathogen, and therefore the inclusion of historical sequences, sampled before such variants stabilized, tends to lower the MI value.

. To select characteristic sites and characteristic variants, we identified four criteria that help distinguish these sites from the background noise. The choice of threshold values for these criteria is largely dependent on the analysis task selection. The four criteria are:

- A characteristic site s_c must have an MI value above MI_{min}, the MI threshold below which no characteristic sites are deemed to be present.
- If a characteristic variant v_c is present at site s_c with probability $pc(v_c,s_c)$ within the set it represents and $po(v_c,s_c)$ in the other set, the ratio $r(v_c,s_c)=pc(v_c,s_c)/po(v_c,s_c)$ must exceed a minimum frequency ratio r_{min} if $po(v_c,s_c)$ is non-zero. A high r_{min} ensures that the variant is significantly more common in the set it represents.
- The probability $pc(v_c,s_c)$ must exceed a minimum probability pc_{min}. Raising this threshold prevents statistically insignificant mutations from being considered characteristic, even when they are more frequent in one set than in the other.
- At a characteristic site s_c, the probability $pc'(S,s_c)$ of a set S containing variants characteristic of the other set must be lower that the maximum contamination probability $pc'_{max}(S)$. This threshold prevents a site from being classified as characteristic if there is significant cross-contamination of variants between the two sets. Depending on the analysis task, it is desirable to specify a different threshold for each set: for example, the tolerance for human variants present in avian influenza sequences may be greater than the tolerance for avian variants in human sequences, to account for a more diverse pool of mutations in the avian influenza population.

The selection process produces a *characteristic variant pattern*: a catalogue of characteristic sites, each possessing a list of the characteristic variants identified for each of the two sets. A characteristic variant pattern presents in a concise form the systematic differences between a pair of aligned sequence sets, and can be used to derive a *sequence signature* for any

homologous sequence. Sequence signatures comprise only the residues at characteristic sites, and thus provide a concise representation of any given isolate, useful for determining which characteristic mutations it possesses.

5.2.2 Set size ratio considerations in Mutual Information Computations

The variation in MI range, caused by size disparities between sequence sets, poses a challenge to the objective identification of characteristic sites, since selection must rely on absolute MI thresholds. Figure 5-3A shows that, as one set becomes several times larger than the other, MI values decrease at all sites. The effect occurs at all sites, irrespective of the MI value (the six positions were chosen are representative of various MI ranges).

We have devised a statistical correction that compensates for such set size bias. The correction uses a sampling method, which compares the smaller of the two sets to multiple subsets of the larger set and evaluates the mean MI. Each subset is randomly selected and equal in size to the smaller set of aligned sequences. Figure 5-3B shows the effect of applying this statistical correction: MI values remain relatively stable even as set size ratio exceeds 1:10, especially at sites where MI is high. Small sequence counts, however, affect the estimate reliability at very low ratios. These measurements indicate that the sampling correction gives reliable MI results with size ratios up to 1:10. The capability to apply this statistical correction has been built into the AVANA tool.

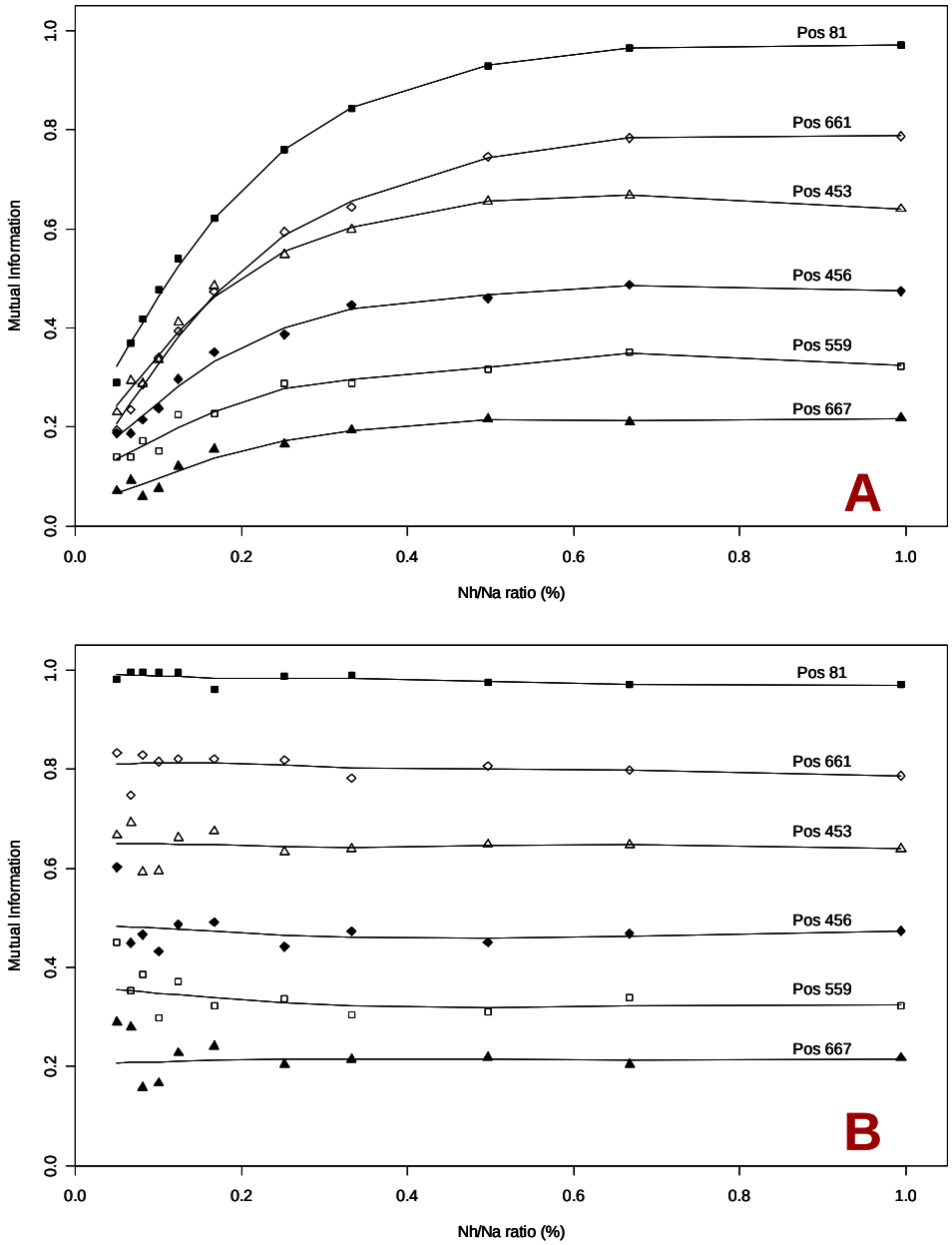


Figure 5-3: Effect of set size ratio on mutual information.
The y-axis represents the measured mutual information (MI) between two sets of influenza A PB2 protein sequences, comprising human and avian sequences respectively. The x-axis represents the size ratio Nh/Na, where Nh and Na are the sequence count in the human and avian sets respectively. A) Changes in MI at selected alignment sites as Nh is varied (Na=719). MI values fall rapidly as the ratio decreases, especially at high MI sites. B) Each data point is computed by averaging the MI obtained by comparing the human set with 200 random subsampled sets of avian sequences with the same sequence count. The estimated MI values remain stable up to a size ratio of approximately 1:10. At very low ratios, increased sampling errors due to small set size result in lower reliability of the estimate.

5.3 Implementation: the AVANA tool

The Antigenic Variability Analyzer (AVANA) tool, which supports a variety of entropy-based analyses of multiple sequence alignments, is the software engine that was developed to implement the information theory- based analysis methods described in this chapter, and which was used to perform all sequence diversity studies described in this thesis. Developed in the Java language, AVANA is a standalone tool that can run on personal computers running any Java-compatible operating systems. Because of the statistical methods used, AVANA performs well on large sequence alignments: we have found that a current standard-configuration computer (3.2 GHz Pentium 4 CPU, with 1 gigabyte of RAM, running Windows XP operating system) analyzes alignments of over 3,000 sequences in real time, with excellent speed (typically under 10 seconds for entropy analysis without statistical corrections). AVANA is published under an open-source license, and is freely available for download from http://avana.sourceforge.net/.

The AVANA tool accepted multiple sequences alignments as input, either in the standard FASTA format, or as *tab-separated alignment* files. Both formats allow multiple sequences to be included in a text file. FASTA formats use a single line description starting with angle bracket character (“>”), followed by one or more lines of sequence data. In tab-separated alignment files, each line contains an identifier following the aligned sequences, separated by a tab. Alignments can be prepared by any multiple sequence alignment tool, such as ClustalW (Chenna *et al*. 2004), MUSCLE (Edgar 2004) or T-Coffee (Notredame *et al*. 2000) and/or manually edited using a sequence editor, such as BioEdit (Hall 1999). All sequences in the alignment file must be of the same length- protein fragments must therefore be appropriately padded, as described in Section 5.1.2.

A key feature of AVANA is its supports of descriptive metadata (annotations) to accompany the loaded sequences. The tool is able to load arbitrary metadata fields, encoded in a CSV (comma-separated values) text format which can be constructed by hand, or generated from a common desktop tool such as Microsoft Excel. The first line in the CSV file

is treated as a list of metadata field headers, and subsequence lines specify the sequence id, followed by the values for the metadata fields. All values are interpreted as string values, and the user can choose what fields to include. The use of metadata is optional, since it is loaded separately from the master alignment. Metadata is used to construct subset alignments from the master alignment, by selecting values for the sequences to be included in the subset. For example, for an alignment of influenza sequences, all Human H5N1 sequences are grouped by selecting the value "H5N1" from the "Subtype" field, and the value "Human" from the "Host" field. Subsets constructed by metadata selection can then be separately analyzed, used in comparative analysis, or in consensus conservation analysis. Metadata capabilities allow the definition of *homologous alignments*, which are managed internally by the tool rather than analyzed separately as would be necessary with other alignment analysis tools.

Figures 5-3 and 5-4 show screenshots of the AVANA tool performing single-set and comparative diversity analyses of sequence alignments. The AVANA tool is capable of multiple functionalities, as follows:

- **Subset Selection and Management**. After the master alignment and its associated metadata are loaded, AVANA allows users to create any number of subset alignments, based on metadata selection. The selection procedure is performed through a dialog box, which displays selectable lists of metadata values. Sequences that meet all metadata selection criteria are selected to form the subset alignment. The selection criteria for subsets can subsequently be modified, and subsets can subsequently be deleted, or used as the source for more subsets.
- **Entropy Analysis**. The diversity of any alignment (master alignment or subset alignment) is be analyzed by computing the residue or peptide entropy at all alignment positions. Peptide length is user-selectable, and a peptide length of 1 produces residue entropy analysis. Entropy value extrapolation to infinite-size set (as described in Section 5.1.3) can optionally be applied, to produce entropy values that are independent of sequence count. AVANA can also compute local average entropy values, using a rectangular sliding window, whose size can be specified by the user.

The entropy analysis results are displayed graphically as a line plot, where the horizontal axis corresponds to the positions along the alignment, and the vertical axis represents the entropy at that position (see Figure 5-3). A position cursor allows the user to specify an alignment position, to display detailed position statistics: the entropy value, number of gaps, support, etc. The display highlights positions where the percentage of gaps exceeds a user-specified threshold, or that have excessively low statistical support. In addition, a sequence logo provides a qualitative display of the diversity observed at the selected position. At any selected position, the user can inspect the metadata of sequences that contain a particular variant, so that association of variants with specific sequence properties can be investigated. The position diversity statistics for the whole alignment can be exported as a tab-separated text file that can be imported in spreadsheet applications.

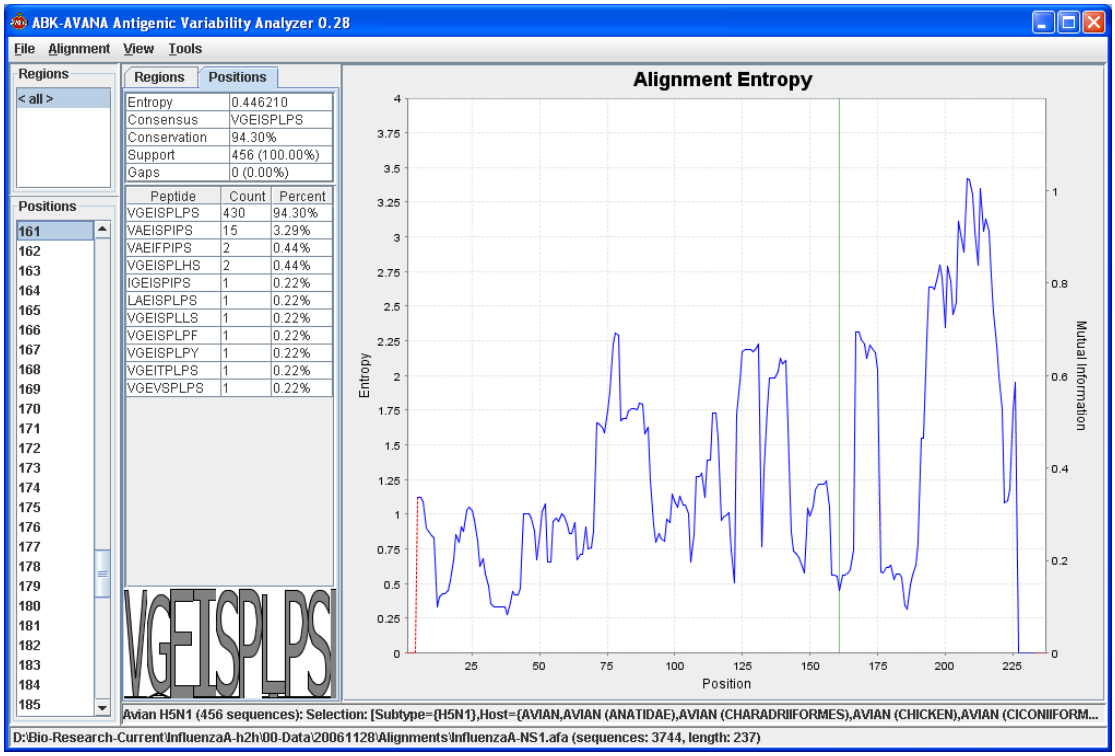


Figure 5-4: Screenshot of the Antigenic Diversity Analyzer (AVANA), showing single-set entropy analysis results.
In this example, 9-mer peptide entropy of the NS2 protein of Avian H5N1 influenza viruses was analyzed. The plot's x-axis indicates amino acid positions along the alignment, while the y-axis represents the entropy value at those positions. On the left-hand side, the position cursor was used to select a low-variability position (161), and the variants observed at that positions are listed, in decreasing order of frequency. The sequence logo in the left corner of the screen shows a clear predominance of a single 9-mer peptide at this position.

- **Region Analysis**. AVANA will perform diversity analysis on any arbitrary region of the alignment specified by the user. In this case, the peptide length used is the length of the whole regions specified. This function is useful when investigating the effect of merging multiple conserved positions.
- **Conservation Analysis**. AVANA automatically identifies conserved regions in an alignment, based on a user-specified threshold (either maximum entropy, or minimum conservation percentage). Given a user-specified minimum region length, the algorithm finds the longest possible regions that meet the threshold requirements (optionally, a maximum region length can also be specified). Conservation analysis can be performed on a single alignment, or as meta-analysis of conservation in multiple subset alignments. Multiple-alignment conservation performs conservation analysis on all the selected alignments, and only retains region that exhibit consensus conservation. Consensus-conserved regions must contain the same peptide and meet conservation requirements in all alignment. This type of analysis is used when it is desirable to separate subsets of sequences, as conservation analysis of a combined set would be biased by the difference in set size.
- **Variant Analysis**. The AVANA tool is capable of exporting variant analysis data for all positions in the alignment. For each position, variants are presented in decreasing order of frequency, and their count, percentage conservation, and cumulative percentage is given. This output is used to determine the minimum set of sequences that would cover a given proportion of the viral diversity.
- **Comparative Analysis**. Pairs of subset alignments can be compared by mutual information (MI) analysis, as described in section 5.2. The mutual information at each position is computed, and statistical set size bias compensation (see section 5.2.2) can optionally be applied. The comparative analysis is displayed graphically, as shown in Figure 5-5. Parallel displays of diversity analysis for each of the two alignments are shown, with the addition of a MI plot overlay (green line). MI peaks

are thus easily identified, and can be inspected with the position cursor. AVANA can also apply user-defined threshold in order to identify *characteristic sites*, which meet minimum MI, and minimum variant frequency and frequency ratios. The result of characteristic site analysis is a *characteristic site pattern*, which defines the differences between the two alignments at characteristic site, and can be exported for generating sequence signatures.

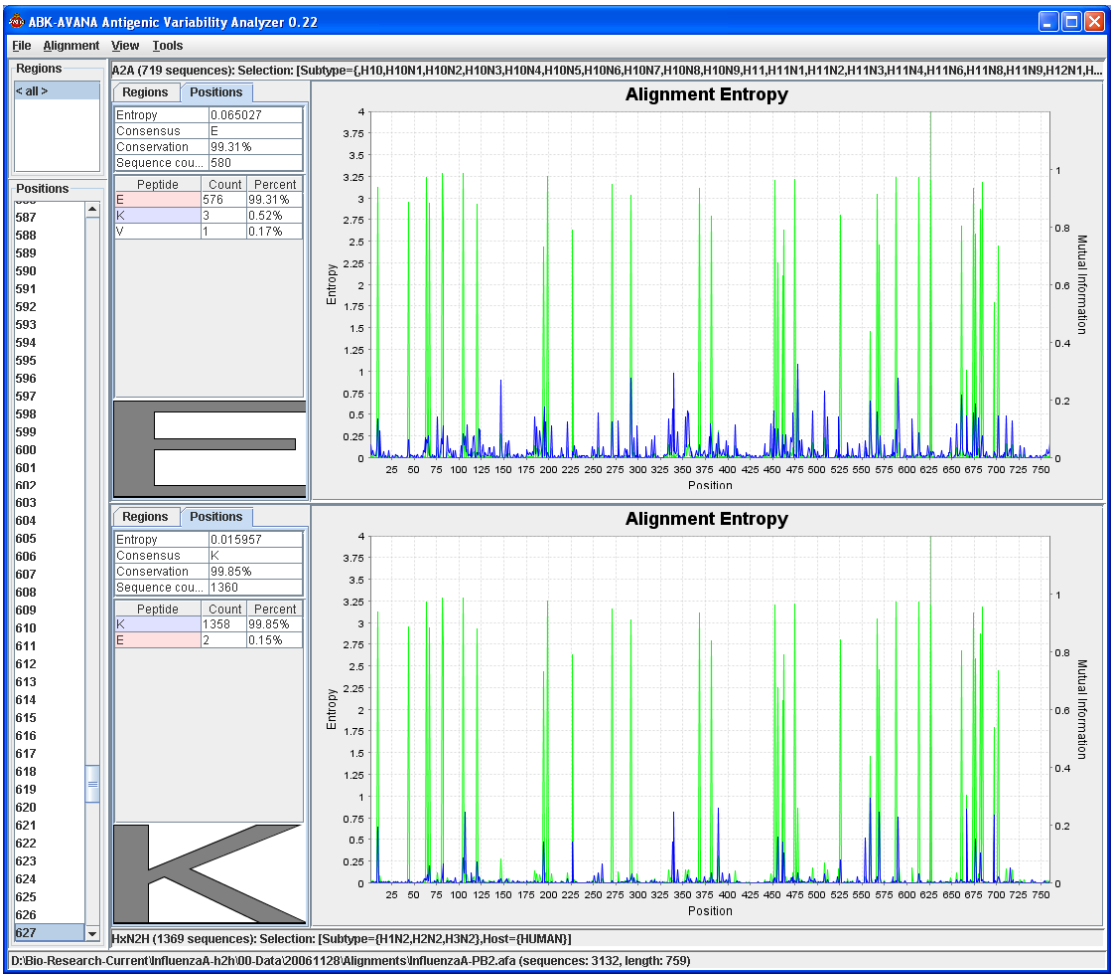


Figure 5-5: Screenshot of AVANA, showing a comparative analysis of the sequence subsets A2A (avian-to-avian transmissible strains, top) and HxN2 (human H3N2, H2N2 and H1N1, bottom) for the influenza A PB2 protein.
Single-residue entropy is plotted along the y-axis (blue line), alongside mutual information between the two subsets (green). Characteristic sites are identifiable by the presence of MI peaks. In this example, the E627K characteristic mutation is shown.

- **Sequence Signatures**. AVANA provides a tool for generating *sequence signatures* from an alignment, given a characteristic variant pattern. For each sequence in the alignment, we construct a signature by concatenating the residues present at all characteristic sites. Partial sequences may only generate partial signatures, and signatures may contain gaps. Our work on influenza internal proteins did not produce any sequence signature containing gaps. Signatures are ordered according to their metadata (for example, in chronological order), and displayed so that characteristic variants are shown on coloured backgrounds, which differ according the set characterized by the variant. For example, when displaying influenza sequence signatures, we displayed human-to-human transmissible characteristic variants on a yellow background, and avian-transmissible variants on a blue background. Variants that are not characteristic of either set were shown on a plain white background. The resulting display, which is exported as an HTML document viewable in any Web browser, provides a concise and easy to interpret view of the presence of characteristic mutations in the aligned sequences. The emergence of each characteristic mutation is shown clearly as a coloured streak within the display, so that timelines for the emergence of these mutations can be observed.

5.4 Conclusions

In this chapter, we have developed two novel methods for measuring different aspects of diversity in large sequence alignments, based on computational measures defined in information theory. These methods were implemented in the AVANA analysis tool, which is capable of analyzing alignments of thousands of sequences on current standard configuration PC systems. AVANA is knowledge-enabled, and can utilize user-defined descriptive metadata to organize alignment sequences into subsets for comparative analysis or meta-analysis. The information-theoretical methods implemented by AVANA support three novel types of analysis, defined in this chapter. First, *peptide information entropy* was defined as a

measure of variability that is particularly relevant in immunological applications. Peptide entropy accounts for the combinatorial effects of varying neighbouring residues, and therefore measures the pathogen protein diversity from an antigenic perspective, since immune responses involve short peptides. Peptide entropy was used as a step in the identification of potential peptide vaccine targets, as described in Chapter 7. Second, a *comparative analysis method* based on mutual information was developed, capable of discovering characteristic mutations associated to specific sets of sequences. In Chapter 6, we have shown that this method can reconstruct the catalogue of adaptive mutations that control human-to-human transmission of influenza A viruses. To our knowledge, this is the first application of mutual information for such purpose. Third, the catalogue of adaptive mutations can be used as descriptive metadata to produce *adaptive signatures*, which provide concise representations of the variants observed in particular viral isolates. Signatures will be used in Chapter 6 as a tool for evaluating the potential for transmissibility to humans of avian strains.

The applications detailed in Chapters 6 and 7 will demonstrate that the combination of computational scalability and metadata capabilities that characterize AVANA is a powerful enabler for new types of analysis, capable of producing important results relevant to biomedical questions.

6. CHARACTERIZATION OF INFLUENZA A VIRUS HUMAN-TO-HUMAN TRANSMISSIBILITY

In this chapter, we present a complete application of our knowledge mining approach, which demonstrates both the utility of our method and the discovery power of knowledge-enabled "second-generation" bioinformatics analysis. We conducted a complete-proteome large-scale analysis of influenza A sequences, to discover adaptive mutation sites involved in viral adaptation to humans and to evaluate the pandemic potential of avian strains. The biological knowledge mining process, shown in Figure 6-1, used knowledge aggregation to construct a large-scale dataset of influenza A sequences, as detailed in Chapter 3, Section 3.4. The AVANA tool used the aggregated metadata to partition alignment sequences into epidemiologically meaningful sets, and performed systematic comparative analysis between avian strains and the co-circulating human strains, utilizing the novel mutual information method detailed in Chapter 5, Section 5.2. The comparative analysis resulted in a catalogue of 70 amino acid sites that carried characteristic mutations associated with human-to-human (H2H) transmissibility. This is the most comprehensive such catalogue to be produced to date, demonstrating that our approach has a higher resolving power than previous method, due to the large number of sequences used, the higher sensitivity of mutual information measures, and AVANA's ability to conduct analyses based on a high-level population models. As a further demonstration of the utility of aggregating metadata of diverse types, the catalogue of adaptive mutations was used by AVANA to derive *adaptation signatures* of viral genomes, which summarize the presence of adaptive mutations in any given isolate. Adaptive signatures were used as a tool for evaluating the pandemic potential of H5N1 and other strains of avian influenza, a current question of the outmost importance for the influenza scientific community.

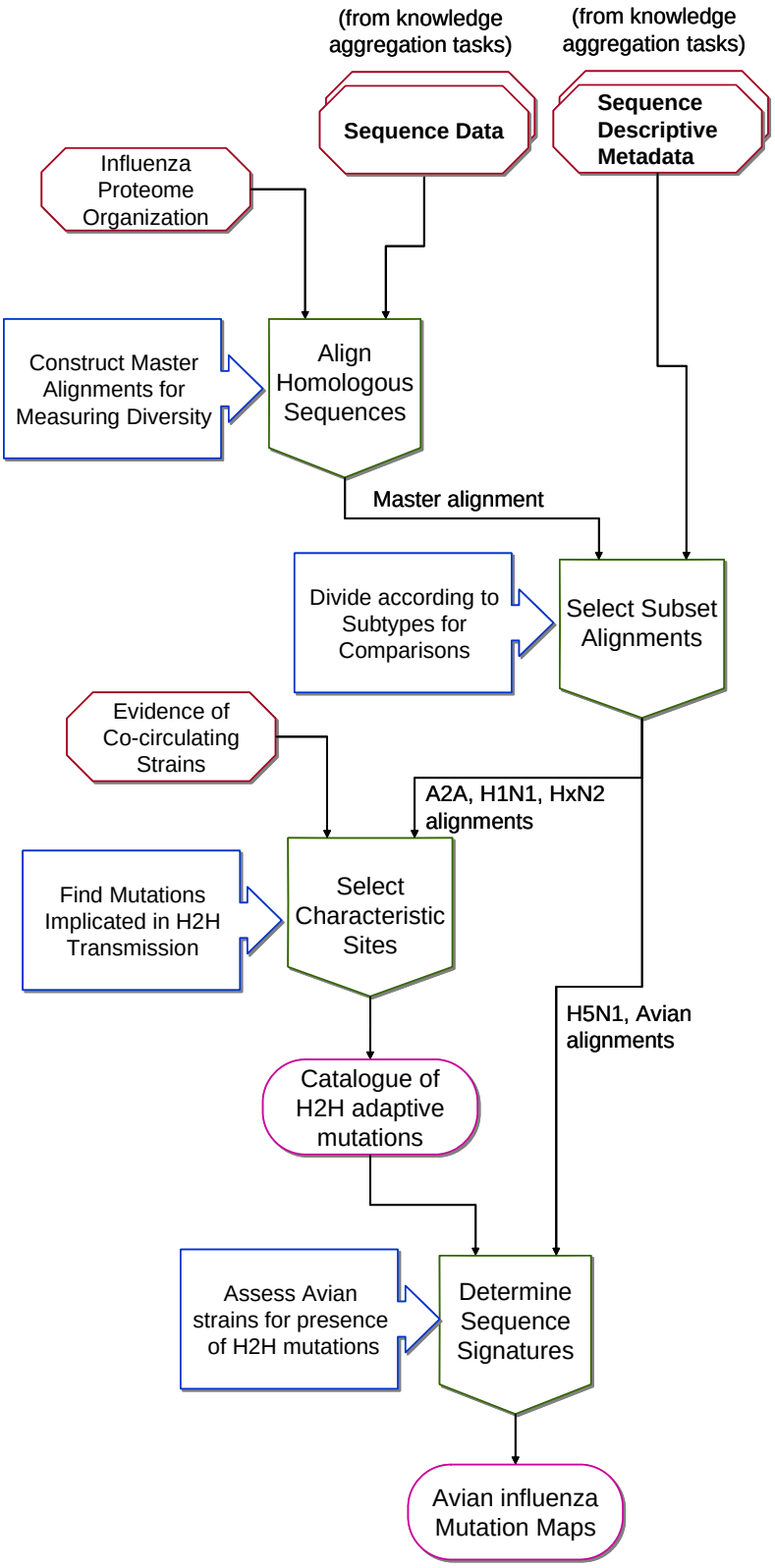


Figure 6-1: Knowledge Mining Model for the workflow of the characterization of H2H transmissibility in Influenza A viruses
Sequences aggregated using the ABK tool were organized into master alignments (one per protein). The AVANA tool used descriptive metadata to divide the master alignments into subsets, which it compared to identify characteristic sites. Finally, based on the catalogue of characteristic sites, AVANA extracted adaptive signatures for avian sequences.

6.1 Background

Influenza A viruses belong to the Orthomyxoviridae family, which circulate amongst various animal species. Aquatic wildfowl are generally accepted to be the natural reservoir of the influenza A, but these viruses routinely infect many types of domestic birds and several mammalian species. In humans, influenza A viruses cause widespread annual epidemics, and less frequent pandemics. Seasonal epidemics produce elevated economic burden and substantial mortality (Thompson *et al.* 2003). The threat of a new worldwide pandemic is a cause of the greatest concern, due to both the excessive death toll and the morbidity of pandemics. The Spanish flu of 1918/19 claimed over 40 million lives, ranking amongst the most destructive events in medical history (Potter 2001). The rapid large-scale spread of such pandemics is enabled by the introduction of novel strains, for which the human population has no immune memory. Such strains introduce new variants of at least one of the viral glycoproteins hemagglutinin (HA) or neuraminidase (NA), which are the external proteins most likely to interact with the human humoral immune system. Sixteen serologically distinct HA types, and nine NA types, are known to circulate in the avian host population; over 100 avian influenza subtypes have been catalogued to date, resulting from the combination of different HA and NA types. Of all these subtypes, only four (H1N1, H2N2, H3N2 and H1N2) are known to have circulated amongst humans in the last century. Other influenza subtypes of avian origin are known to have infected humans through avian-to-human (A2H) transmission, but without acquiring the ability to spread within the human population. The best known current source of such zoonotic infections is the highly pathogenic H5N1 virus, whose spread among poultry and wild birds has caused considerable economic damage. Over the last ten years, these H5N1 viruses have been responsible for a considerable number of human infections and deaths in Asia and Africa: according to a June 2008 WHO report, at least 385 individuals were infected since 1997, resulting in 243 fatalities (http://www.who.int/csr/disease/avian_influenza/). Although no definitive evidence of human-to-human (H2H) transmission of H5N1 has been reported, there is widespread concern

that these viruses could cause a new devastating pandemic if they acquire such capabilities. More generally, scientists and policymakers have become increasingly aware of the possibility of a pandemic caused by avian viruses, and new tools are needed for evaluating the pandemic risk posed by all avian subtypes.

The limited spread of zoonotic influenza in humans indicates that immunological naivety of the host population is not a sufficient condition for initiating a human pandemic, and additional adaptive changes in the virus are required. Studies of host range determinants suggest that such adaptations involve multiple mutations in the viral genome. Such mutations appear not to be limited to the HA and NA proteins, but rather distributed across the influenza genome, including its nine internal proteins (Neumann and Kawaoka 2006). A full reconstruction of this complex landscape of adaptive mutations is needed for the elucidation of biological mechanisms of viral adaptations to humans. In addition, detailed knowledge of adaptive mutations provides an important tool for the assessment of the pandemic potential of avian strain. Mutations of critical importance for host range, such as PB2 E627K, have been experimentally identified using laboratory mutants of the virus and primate models (Subbarao *et al*. 1993). However, such experimental approaches are prohibitively expensive if extended to genomic-scale studies. A more cost-effective approach is to conduct statistical studies, involving the comparative analysis of human-infecting and avian strains, to identify candidates for experimental studies. These *characteristic sites* are genomic positions where different residues are consistently observed in the two groups. A residue (*characteristic variant*) that is highly conserved within the human group, but rarely observed in the pool of avian strains, is likely to indicate an important adaptive mutation, whose loss would affect the virus' ability to replicate or propagate amongst human hosts. Studies based on visual inspection of small numbers of representative isolates found characteristic sites in matrix proteins (Buckler-White *et al.* 1986) and polymerases (Naffah *et al.* 2000). Wider availability of influenza sequence data from public repositories has made it possible to conduct comparative analyses of greater statistical significance. Such large-scale studies comprise hundreds or thousands of sequences, and thus require the deployment of computational

methods able to handle the resulting complexity. Large-scale computational methods are frequently based on *information entropy*, a statistic which is relative simple and fast to compute, and which summarizes multiple aspects of variability in one measure. A large-scale study used information entropy to identify characteristic sites for human transmissibility of influenza A (Chen *et al*. 2006). This analysis of 401 full viral proteomes identified characteristic sites by comparing the entropy statistics in the avian and human groups, which limited its applicability to positions highly conserved in both groups. Finkelstein *et al.* (2007) overcame this limitation by employing statistical tests that involve the comparison of frequencies of multiple residues, thus removing the requirement for residue conservation. This method processed more than 23,000 sequences, to construct a catalogue of 32 characteristic mutations in five influenza proteins.

6.2 Materials and Methods

6.2.1 Data collection and preparation

We built a dataset of all available influenza A sequences (as of September 2006) from the NCBI GenBank and GenPept databases (Wheeler *et al* 2008b), including entries mirrored from UniProt (UniProt Consortium, 2008). A total of 92,343 records were retrieved from these databases, using taxonomy-based queries; entries from different databases that referred to the same sequences were subsequently merged. Wherever sufficient information was available, sequences were annotated with descriptive metadata properties, including: isolate name, country and year of isolation, host organism, subtype, and protein name. The resulting dataset was verified by two independent curators, who discarded duplicates, laboratory strains, sequences with missing key metadata, and sequences with quality issues. The final set comprised a total of 40,169 unique sequences, including both full-length and fragment sequences, covering all influenza A proteins. The data collection and cleaning process was largely automated by the Aggregator of Biological Knowledge (ABK) tool, as described in Chapter 3.

For each of the eleven influenza proteins, a master multiple sequence alignment (MSA) was constructed using the MUSCLE 3.6 (Edgar, 2004) software, and manually inspected and corrected where required. Multiple subset alignments, to be used in comparative analyses, were extracted from the master alignments, based on metadata values. The extraction of subsets from the master set, without realignment, allowed the direct comparisons of residue statistics at each alignment site. Subsets were extracted using the metadata-enabled Antigenic Variability Analyzer (AVANA) tool (Miotto *et al.*, 2007c), developed by our team to support information-theoretical analysis tasks. AVANA was also used to conduct all comparative analysis described in this paper.

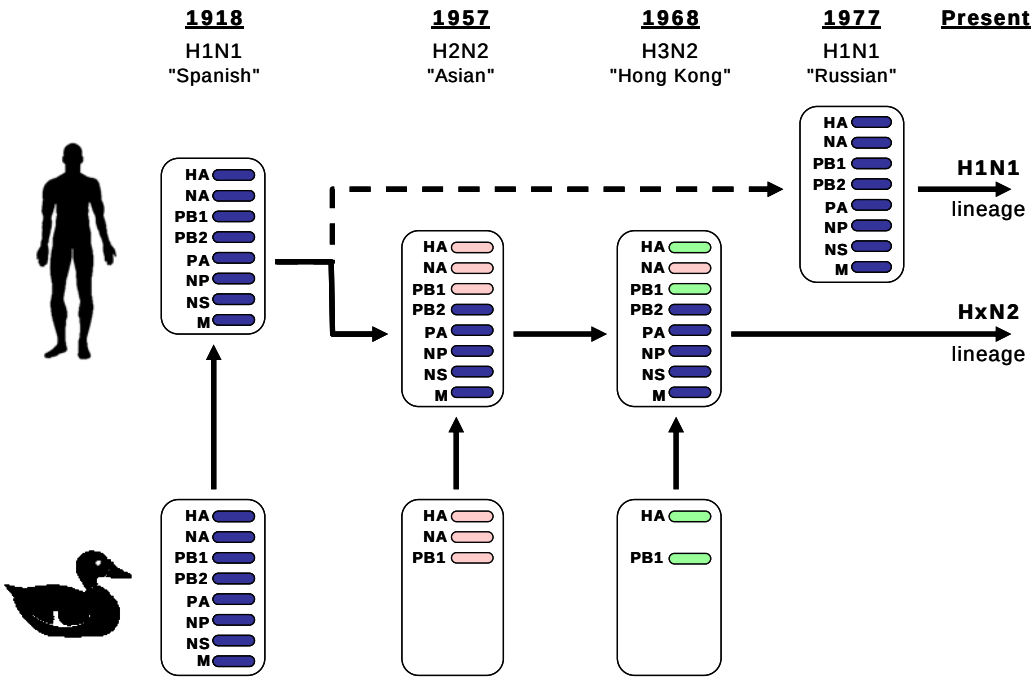


Figure 6-2: Human Influenza A reassortment events of the 20th Century.
This figure (adapted from Webster et al. (1992)) describes the reassortment events associated with human pandemics in the 20th Century. A full complement of eight gene segments of zoonotic origin causes the 1918 Spanish flu. In 1957, the H2N2 Asian flu pandemic replaced the HA, NA and PB1 segments, and in 1968 the H3N2 Hong Kong pandemic replaced the HA and PB1 segments only (Scholtissek et al. 1978). In both cases, the new subtype fully replaced the subtypes previously circulating amongst humans. The Russian pandemic of 1977 introduced an H1N1 strain almost identical to that circulating prior to 1957, and may have been caused by the release of 20-year old frozen viruses (Kendal et al. 1978). The H1N1 and HxN2 lineages have since co-circulated in the human population; recently, their reassortment has given rise to human strains of H1N2 subtype.

6.2.2 Subset Selection

The objective of our study was to identify sites where characteristic mutations are observed in the large majority of human influenza viruses. There are currently two major co-circulating lineages of human influenza: one predominantly of subtype H3N2, and another of subtype H1N1. The two lineages are thought to have emerged as a result of various pandemic events, as summarized in Figure 6-2 (Webster et al., 1992). Although both lineages have descended from the 1918 Spanish influenza strains, their internal protein constellations have evolved separately, following the disappearance of H1N1 in 1957 and it subsequent reintroduction in 1977 (Kendal *et al.* 1978). Thus, when analyzing internal proteins, we distinguish the HxN2 lineage (comprising human sequences of subtypes H2N2, H3N2 and H1N2), and the H1N1 lineage. For each of the nine internal proteins, the following three subsets were therefore extracted: **A2A** (all avian sequences, except for H1N1, H2N2, H1N2, H3N2 and H5N1 subtypes), **H1N1H** (all H1N1 human sequences) and **HxN2H** (all human sequences of subtypes H2N2, H1N2 and H3N2). Since true adaptive mutations are expected to be present in both lineages, we analyzed each lineage separately, discarding sites that are not shared by both human influenza lineages. Subtype H5N1 was removed from both avian and human subsets because of its pronounced ability to jump the species barrier, and was analyzed as a separate subset. We collected subsets of avian H5N1 (**H5N1A**) and human H5N1 (**H5N1H**) to analyze their adaptation signatures. Table 6-1 shows counts of the sequences included in each of the extracted datasets.

When grouping the HA and NA proteins by lineage, the high genetic divergence between different subtypes tends to mask adaptive mutations in statistical analyses. We therefore conducted separate comparisons of avian and human sequences for each subtype that circulates amongst humans: H1, H2 and H3 subtypes of the HA protein, and N1 and N2 subtypes of NA. Table 6-2 shows the number of sequences included in each HA subset, and Table 6-3 shows the subset sizes for the NA proteins.

	A2A	H1N1H	HxN2H	H5N1A	H5N1H	Total
M1	1047	300	1521	458	105	**3431**
M2	736	286	1517	289	95	**2923**
NP	884	316	1645	420	114	**3379**
NS1	1123	303	1448	457	95	**3426**
NS2	810	292	1419	288	81	**2890**
PA	701	279	1362	402	102	**2846**
PB1	716	303	1385	400	101	**2905**
PB2	719	281	1369	404	97	**2870**
PB1-F2	352	262	1280	-	-	**1894**
Total	**7088**	**2622**	**12946**	**3118**	**790**	**26564**

Table 6-1: Count of influenza A internal protein sequences used in the current study. Characteristic site analysis was conducted using the A2A, H1N1H and HxN2H sets. The H5N1A and H5N1H sets were used for sequence signature analysis.

	Avian	Human	Total
H1	48	768	**816**
H2	80	75	**155**
H3	115	3105	**3220**
Total	**243**	**3948**	**4191**

Table 6-2: Count of influenza A hemagglutinin protein sequences used in the current study.

	Avian	Human	Total
N1	717	360	**1077**
N2	439	1801	**2240**
Total	**1156**	**2161**	**3317**

Table 6-3: Count of influenza A neuraminidase protein sequences used in the current study.

6.2.3 Identification of characteristic sites and variants

Characteristic sites were identified using the method described in Section 5.2. All comparisons were performed using the AVANA tool, applying a statistical correction for set size bias to all comparative analyses in this study (described in Section 5.2.2), based on 200 resampling iterations. Characteristic sites and their characteristic variants (mutations) were selected based on the criteria detailed in Section 5.2.1, as follows:

- MI_{min} = 0.4. This threshold was determined by an analysis of medium-MI sites in all internal proteins of influenza, which indicated that avian and human sequences converge to the same consensus amino acids as MI falls below 0.4.
- r_{min} = 4. To determine this value, we analyzed the probability ratio $r(v,s)$ for all variants at each position in selected protein alignments, discarding variants with >99% conservation, or probabilities below 1%. For PB2, the standard deviation of $log_{10}r(v,s)$ was 0.52, corresponding to a ratio of 3.29 (log transformation was applied so that ratios could be compared on a linear scale). An identical analysis of an alignment of NS1 protein sequences produced a consistent ratio of 3.26, although NS1 is the most variable internal influenza A protein. A slightly more conservative threshold ratio of 4 was chosen for our analysis. Post-analysis verification confirmed that no H2H characteristic variant presented ratios lower than 9.65, while the highest ratio among H2H non-characteristic variants was 1.45.
- pc_{min} = 0.02. In our dataset, this translates to a minimum support of approximately 30 sequences for H2H characteristic variants. Post-analysis verification for the PB2 protein showed that the lowest support for characteristic variants was 65 sequences (residue M at site 105), indicating that no important characteristic variant was omitted by our choice of threshold.
- $pc'_{max}(S)$ = 0.052. This parameter was not applied manually; rather, we manually inspected all sites where avian variants accounted for more than 2% of sequences in at least one of the two human lineages (H1N1 and HxN2). All accepted characteristic

sites had less than 5.2% contamination from avian variants (average contamination at characteristic sites was 0.71%).

6.2.4 Reconstruction of adaptation signatures

The variants that distinguish H2H from A2A sequences at characteristic sites form a *characteristic variant pattern*, a summary of the significant differences between the two sets of sequences across the whole genome. This pattern was used to construct the *adaptation signatures* of several influenza genomes, by discarding all residues except those at characteristic sites. Residues forming the signatures were tagged as A2A-like (i.e. a characteristic variant of the A2A subset), H2H-like (an H2H characteristic variant), or as non-characteristic. The resulting signatures thus provide a succinct summary of the H2H adaptive mutations contained in the sequences represented. To facilitate the evaluation of the adaptive characteristics of multiple isolates, we developed a software program to graphically display selected signatures along a timeline, using a contrasting colour scheme to distinguish between A2A-like and H2H-like residues.

6.3 Results

This study comprised two major analysis tasks. First, we performed MI-based comparative analysis of avian and human influenza A sequences, considering each protein separately, to identify characteristic sites across the whole genome. Second, we used the full catalogue of characteristic sites to produce genomic adaptive signatures for all human and avian isolates. Human signatures were utilized to reconstruct the history of the emergence of H2H characteristic mutations, while avian signatures were assessed for the presence of human adaptive mutations.

6.3.1 Catalogue of characteristic sites

Our analysis produced a catalogue of 70 characteristic sites that met the selection criteria set in Sections 5.2.1 and 6.2.3(Table 6-4). Characteristic sites were found in eight of the nine

internal influenza proteins, suggesting that adaptation to humans requires the participation of most products encoded by the viral genome. The location of all characteristic sites found within internal proteins is shown in Figures 6-3 to 6-5, alongside the mapping of known functional domains in these proteins. As shown in Figure 6-3, the three internal proteins found to contain the highest number of characteristic sites were the polymerases PB2 (17 sites) and PA (17 sites), and the nucleoprotein NP (12 sites). These three proteins are components of the RNP complex, which encases each of the 8 RNA segments packaged within the influenza virion, and are therefore known to bind to each other and also to the viral RNA. However, the remaining RNP component, the PB1 polymerase, was found to contain only a single characteristic site. The PB1 protein is encoded by an RNA segment that was replaced by two subsequent pandemics (Figure 6-2). As a result, the two lineages of PB1 (and those of PB1-F2, which is encoded by the same RNA segment) are more substantially divergent than those of other internal proteins, because their more remote common ancestry. Therefore, adaptive mutations found in PB1 and PB1-F2 are lineage-specific, with one notable exception: a single PB1 site has produced the same adaptive mutation (V336I) in both H1N1 and HxN2 lineages independently (Figure 6-4). All remaining internal proteins were found to contain multiple characteristic sites: M1 (3 sites), M2 (9 sites), NS1 (8 sites) and NS2 (8 sites), as shown in Figure 6-5. The M2 protein contained the highest density of characteristic sites (almost 1 every 10 residues), including three sites within the M2e extracellular region, which has recently been proposed as a universal vaccine target (Tompkins *et al.* 2007). The analysis of the external HA and NA glycoproteins revealed a large number of subtype-specific adaptive mutations, as shown in Figure 6-6. A small number of HA mutations were found to occur at the same position in multiple subtypes. However, different subtypes present different residues at these sites for both avian and human isolates, making it impossible to determine whether mutations in multiple human lineages were truly equivalent adaptations. We were unable to confidently identify any adaptive mutation in the HA and NA proteins as universal in all human-transmissible strains.

(Facing page)

Table 6-4: Full catalogue of identified characteristic sites for H2H transmission of influenza A.

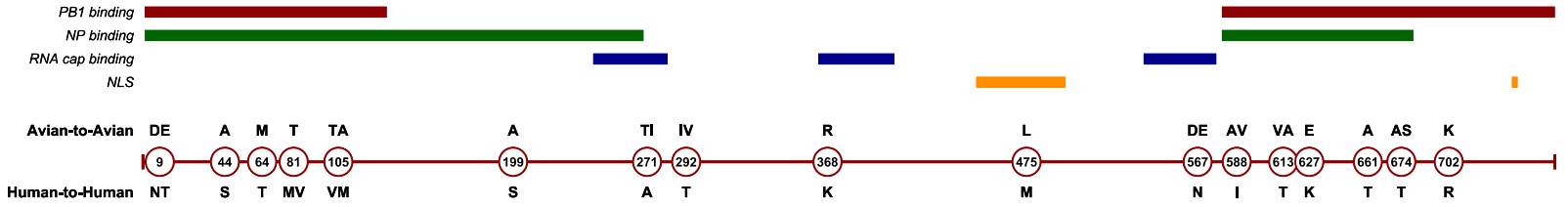
The 70 characteristic sites identified by this study are shown in this table, grouped by protein. Each row represents a site, with the columns detailing the following: the protein name; the site position within the protein sequence; the A2A characteristic variant(s) and their conservation in the A2A subset; the H2H characteristic variant(s), their conservation in the H2H subset, and the contamination with avian variants observed in the H2H subset; the characteristic variant(s) observed in the H1N1 subset alone; and the characteristic variant(s) observed in the HxN2 subset alone. .

Protein	Position	A2A		H2H			H1N1	HxN2
		CV	Cons	CV	Cons	X-pres	CV	CV
M1	115	V	99.70%	I	99.39%	0.61%	I	I
	121	T	94.94%	A	99.89%	0.11%	A	A
	137	T	99.60%	A	99.23%	0.77%	A	A
M2	11	T	97.28%	I	96.89%	3.11%	I	I
	14	G	95.99%	E	98.28%	1.72%	E	E
	20	S	97.14%	N	97.94%	2.06%	N	N
	28	I	76.36%	V	97.72%	2.11%	V	V
	54	R	98.91%	LIF	98.94%	0.61%	IL	LF
	55	L	79.18%	F	99.33%	0.67%	F	F
	57	Y	99.59%	H	97.38%	2.18%	H	H
	78	Q	99.72%	KE	99.26%	0.28%	EK	K
	86	V	99.84%	A	99.21%	0.45%	A	A
NP	16	GS	99.16%	D	99.49%	0.51%	D	D
	33	V	99.76%	I	98.97%	1.03%	I	I
	61	I	98.36%	L	99.43%	0.57%	L	L
	100	R	99.65%	VI	99.71%	0.06%	V	VI
	136	L	85.41%	MI	99.77%	0.11%	I	MI
	214	R	96.64%	K	99.32%	0.68%	K	K
	283	L	100.00%	P	99.48%	0.47%	P	P
	305	R	99.17%	K	99.33%	0.67%	K	K
	313	F	99.31%	Y	99.48%	0.52%	Y	Y
	357	Q	98.43%	KR	99.90%	0.10%	KR	K
	375	DN	96.93%	GEV	99.34%	0.56%	V	GE
	423	A	97.06%	STP	98.88%	1.00%	T	SP
NS1	22	FL	97.07%	V	98.21%	0.40%	V	V
	60	AE	97.59%	V	99.20%	0.69%	V	V
	81	I	98.66%	M	99.08%	0.69%	M	M
	84	VS	96.08%	TA	99.20%	0.80%	A	TA
	114	SG	94.84%	P	99.54%	0.46%	P	P
	171	DTA	91.79%	IN	99.20%	0.69%	N	I
	215	PSA	99.24%	T	99.37%	0.63%	T	T
	227	E	98.87%	R	99.53%	0.06%	R	R
NS2	60	S	76.77%	NH	98.89%	0.82%	H	N
	70	S	97.46%	G	99.88%	0.12%	G	G
	107	L	99.60%	F	98.77%	1.17%	F	F
PA	28	P	100.00%	L	99.14%	0.67%	L	L
	55	D	99.69%	N	99.63%	0.37%	N	N
	57	R	96.61%	Q	98.72%	0.79%	Q	Q
	65	SF	99.08%	LP	99.63%	0.37%	PL	L
	66	GS	99.69%	DE	98.84%	1.10%	ED	D
	100	V	96.15%	A	99.27%	0.37%	A	A
	225	S	98.61%	C	99.39%	0.61%	C	C
	268	L	98.84%	I	99.14%	0.73%	I	I
	321	NK	97.35%	YST	97.30%	0.74%	STY	Y
	337	AT	99.34%	S	99.75%	0.25%	S	S
	356	K	98.51%	R	99.26%	0.74%	R	R
	382	E	94.34%	D	97.37%	2.45%	D	D
	400	PSQ	89.32%	L	99.45%	0.31%	L	L
	404	A	99.48%	S	99.39%	0.55%	S	S
	409	S	91.49%	N	99.45%	0.49%	N	N
	421	S	98.91%	IV	97.79%	0.55%	I	IV
	552	T	99.81%	S	99.75%	0.12%	S	S
PB1	336	V	96.66%	I	95.98%	4.02%	I	I
PB2	9	DE	98.57%	NT	99.33%	0.49%	N	NT
	44	A	96.82%	S	99.27%	0.61%	S	S
	64	M	97.29%	T	99.58%	0.30%	T	T
	81	T	97.93%	MV	99.27%	0.30%	VM	M
	105	TA	98.41%	VM	99.45%	0.36%	VM	VM
	199	A	99.47%	S	99.76%	0.24%	S	S
	271	TI	98.59%	A	99.51%	0.37%	A	A
	292	IV	95.54%	T	99.15%	0.67%	T	T
	368	R	98.12%	K	99.33%	0.67%	K	K
	475	L	99.66%	M	99.76%	0.24%	M	M
	567	DE	98.28%	N	99.39%	0.55%	N	N
	588	AV	98.45%	I	99.63%	0.31%	I	I
	613	VA	98.28%	T	96.82%	0.61%	TI	T
	627	E	99.31%	K	99.76%	0.12%	K	K
	661	A	86.72%	T	99.39%	0.43%	T	T
	674	AS	95.69%	T	99.63%	0.18%	T	T
	702	K	89.70%	R	99.39%	0.49%	R	R

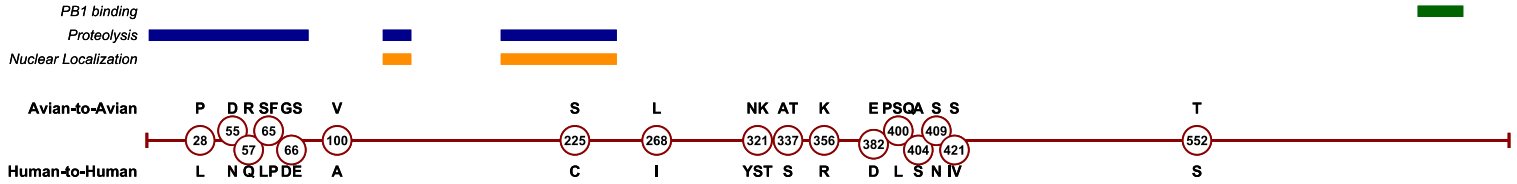
Our catalogue of characteristic sites is considerably more extensive than those produced by earlier related work. In the most comprehensive previous study, researchers at St. Jude Children's Research Hospital identified 32 of the 70 characteristic sites, also found in our study, using a large-scale dataset comparable in size to ours (Finkelstein *et al.* 2007). The greater coverage of our catalogue indicates that MI is a more sensitive measure of association than the statistical tests employed in the St. Jude study. In addition, stringent thresholds may have caused Finkelstein *et al.* to discard several characteristic sites as false negatives, as discussed by the authors themselves. Our catalogue also compares favourably with the results of a study by Chen *et al.* (2006), who identified 52 sites in ten proteins. Of these, 38 were included in our catalogue, while 12 were discarded as representative of a single lineage. Chen *et al.* listed two characteristic sites in the HA protein, which we were unable to identify.

(Facing page)

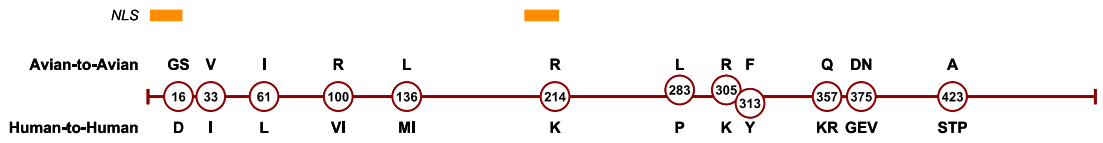
Figure 6-3: Characteristic sites identified in components of the RNP assembly of influenza A (PB2, PA, NP proteins).
Circular markers, indicating the position of characteristic sites, are placed along the sequence length of the PB2 (A), PA (B) and NP (C) proteins of influenza A. Avian (A2A) variants are indicated above each marker, while human (H2H) variants are indicated below the markers. Where multiple characteristic variants are present, they are shown in decreasing order of frequency. In the upper part of each figure, coloured lines show reported functional domains of PB2 (Mukaigawa and Nayak 1991; Poole *et al.* 2004; Honda *et al.* 1999; Fechter *et al.* 2003), PA (Nieto *et al.* 1994; Ohtsu et al. 2002; Sanz-Ezquerro *et al.* 1996) and NP (Ozawa *et al.* 2007).



A) PB2 (759 aa, 17 sites)



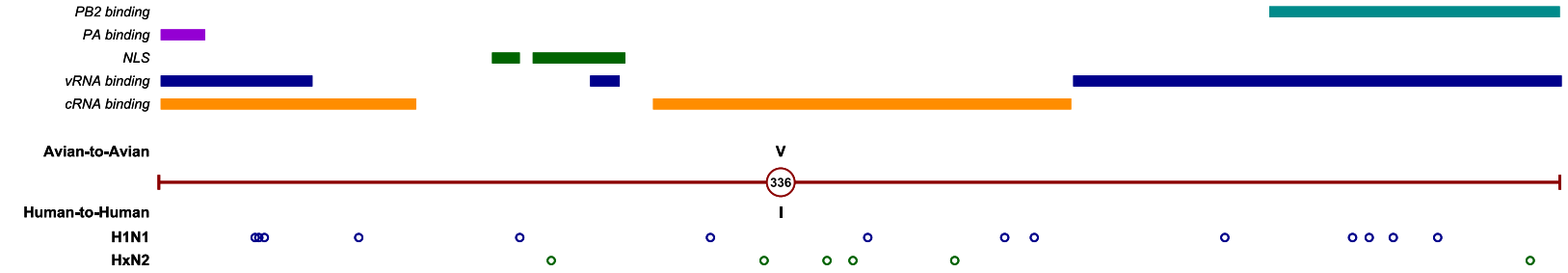
B) PA (716 aa, 17 sites)



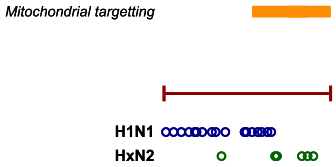
C) NP (498 aa, 12 sites)

(Facing page)

Figure 6-4: Characteristic sites identified in the PB1 (A) and PB1-F2 (B) proteins of influenza A.
RNA segment 2, which encodes both the PB1 and PB1-F2 proteins, has been replaced at the onset of the 1957 and 1968 pandemics (see Figure 6-2). As a result, the H1N1 and HxN2 lineages do not share a recent common origin for this segment. Characteristic mutations are therefore shown separately for the two lineages, in the lower part of each diagram, using blue (H1N1) and green (HxN2) circles. Known functional sites for PB1 (Ohtsu *et al.* 2002; Jones *et al.* 1986; Jung and Brownlee 2006; Gonzalez and Ortin 1999) and PB1-F2 (Yamada *et al.* 2004) are also indicated by coloured lines in the upper part of each figure.



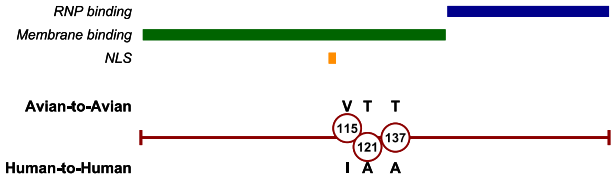
A) PB1 (757 aa, 1 site)



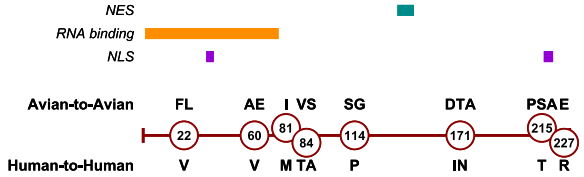
B) PB1-F2 (87 aa)

(Facing page)

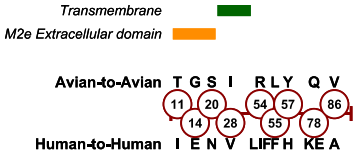
Figure 6-5: Characteristic sites identified in the matrix proteins M1 (A) and M2 (B) and non-structural proteins NS1 (C) and NS2 (D) of influenza A.
The identified characteristic sites are mapped against known functional domains of M1 (Baudin *et al.* 2001; Hui *et al*. 2003), M2 (Lamb *et al*. 1985), NS1 (Greenspan *et al.* 1988; Li *et al.* 1998; Qian *et al.* 1995) and NS2 (Iwatsuki-Horimoto *et al.* 2004; Akarsu *et al.* 2003), using the notation used in Figure 6-3.



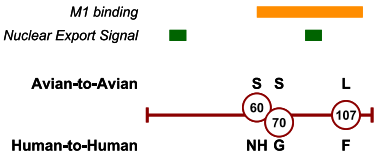
A) M1 (252 aa, 3 sites)



C) NS1 (230 aa, 8 sites)



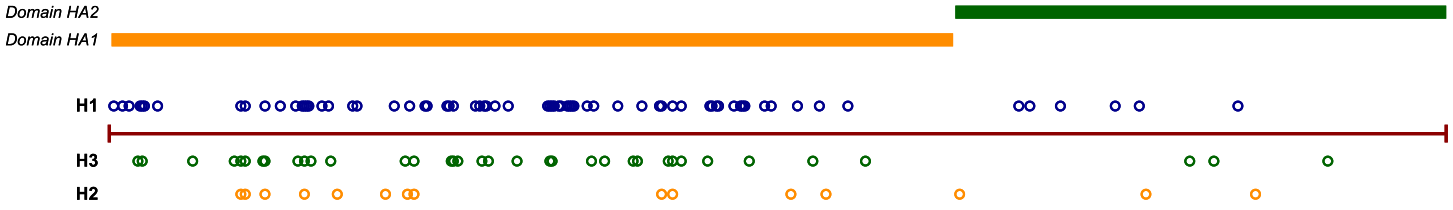
B) M2 (97 aa, 9 sites)



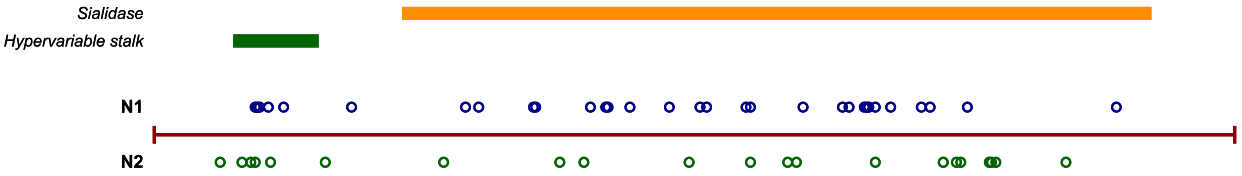
D) NS2 (12 aa, 3 sites)

(Facing page)

Figure 6-6: Characteristic sites identified in the HA (A) and NA (B) glycoproteins of influenza A.
We have shown the characteristic mutations identified for each of the subtypes present in humans: H1 (blue circles), H2 (green circles), H3 (orange circles) for HA; and N1 (blue circles), N2 (green circles) for NA. Known domains of these two proteins are indicated by coloured lines in the upper part of each figure.



A) HA (565 aa)



B) NA (454 aa)

6.3.2 Emergence of H2H adaptive mutations

To assess the stability of H2H characteristic mutations, and reconstruct the timeline of their emergence in human strains, we produced genomic adaptation signatures for all available virus genomes isolated in humans. Figure 6-7 shows the chronological display of signatures from viruses isolated between 1918 and 1972, a period spanning the three major 20th Century pandemics. In this figure, A2A and H2H characteristic residues are shown on contrasting backgrounds, making it easy to discern visually the evolutionary pattern of their emergence. A 1918 Spanish influenza pandemic isolate (A/BrevigMission/1/1918 (H1N1)), placed at the start of the timeline, is the oldest genome available. Although this strain had a primarily avian signature, it contained 23 out of 70 H2H characteristic mutations (33%), distributed in all proteins except for PB1 and NS1. This number of H2H mutations is far higher than that of other avian strains in our dataset, all of which contain no more than eight H2H mutations. The 1918 H2H mutations were conserved in later human strains, which gradually accumulated additional adaptive changes throughout the 1930s and 1940s. By 1950, the first signature without avian characteristic variants was observed (A/FW/50 (H1N1), a progenitor of the current H1N1 lineage). Both the 1957 and 1968 pandemics (indicated by red lines) left the internal protein constellation practically unchanged, except for the replacement of the PB1 segment, which removed from circulation the V336I mutation developed in the 1950s by the H1N1 strains. However, this mutation re-emerged shortly after the 1968 pandemic: by 1972, the HxN2 lineage genome possessed a full H2H signature. Five years later, a new pandemic introduced a human-adapted H1N1 strain, whose signature was identical to that of pre-1957 H1N1 strains (not shown). This new strain produced a separate lineage, with a different signature from that of HxN2. Both lineages are still co-circulating today, and their signatures have remained distinct and highly conserved throughout the intervening half-century. A comparison of all H1N1 and HxN2 signatures over this period revealed that reassortments are infrequent between the two lineages, and there is no evidence that any stable reassorted lineages has emerged over the past 90 years(data not shown). In both lineages, A2A mutations

are only observed in isolates from reported infections of zoonotic origin, usually from swine (see A/Victoria/1968 in Figure 6-7) or avian hosts (for example, human-infecting H5N1). We found no evidence that any of these sporadic events has ever established a stable H2H transmissible lineage.

6.3.3 Assessment of avian strains for H2H adaptive mutations

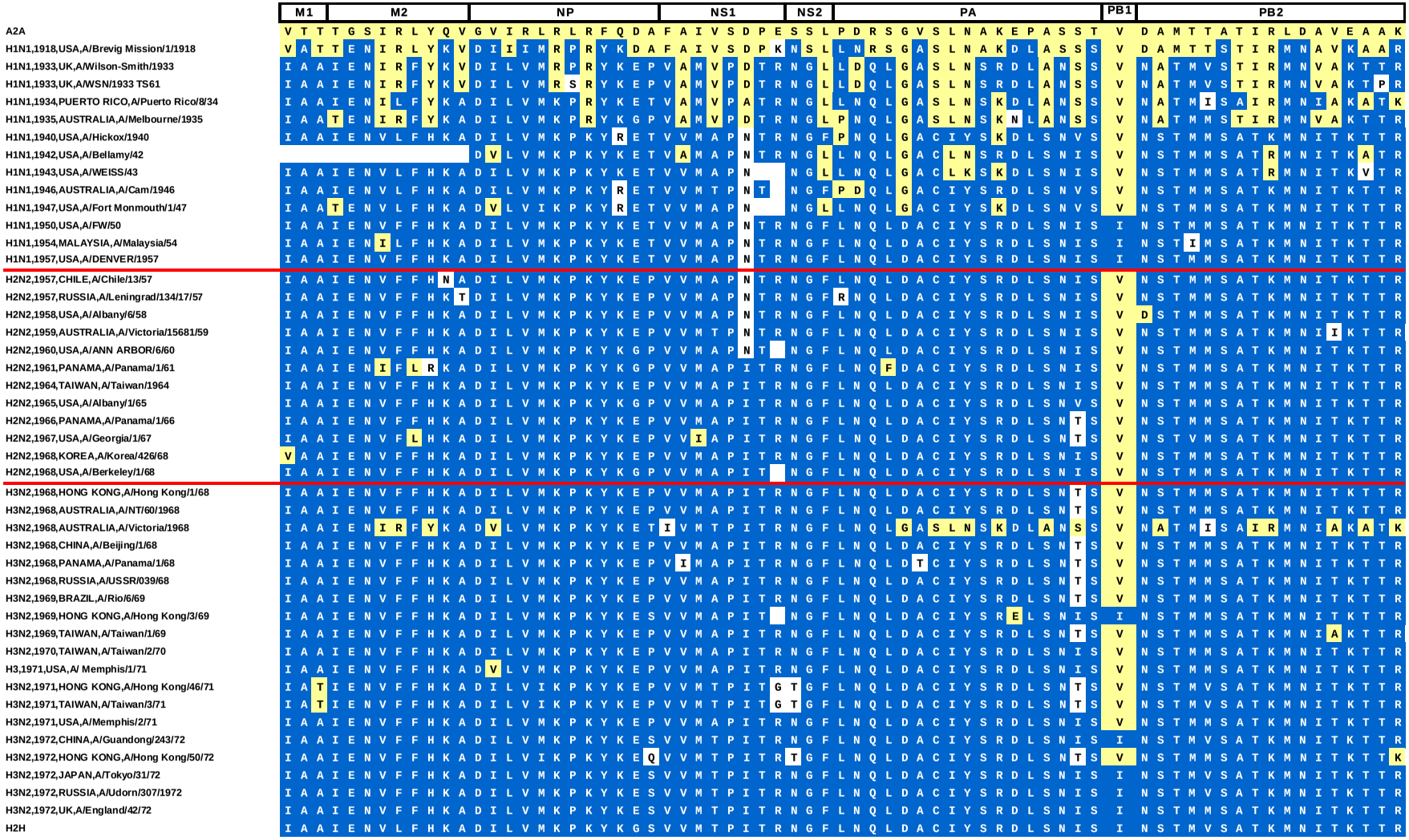
We investigated the presence of adaptive mutations in avian strains by constructing adaptation signatures for all avian sequences analyzed in this study. The vast majority of avian signatures (77%) contained no H2H mutations at all. Although this high percentage may be an overestimate (many of these signatures were incomplete due to partial sequencing of the genome), it is clear that H2H variants are very rare in the avian influenza population. In contrast, we found that human-infecting H5N1 strains had an unusually high number of H2H mutation. Figure 6-8 shows a timeline of human-infecting H5N1 sequences, arranged chronologically with a red line separating two major "waves": the 1997/98 Hong Kong cases, and the later South-East Asian infections, starting in 2003. The earlier Hong Kong isolates present up to ten H2H mutations spread over five internal proteins (see A/Hong Kong/532/97 (H5N1)), more than any other avian strains in our dataset. Later South-East Asian strains also contain a relatively high number of H2H variants (between 3 and 5), but their number is considerably lower than in the previous wave, and they are present in only three proteins. Only a single mutation is present in both H5N1 waves: Ile→Val at position 28 in the transmembrane region of the M2 protein. As expected, the signatures of avian H5N1 isolates from the 1997 Hong Kong wave presented a similar number of H2H mutations as their human-infecting counterparts (up to eight, the highest number observed in avian isolates).

In addition, we found several sequences with a high numbers of adaptive mutations, from other avian subtypes. Figure 6-9 shows the signatures of a number of isolates that contained 5 or more H2H mutations. Most of these viruses (shown above the red line) were isolated in Asia over the past decade, and belong predominantly to three subtypes (H5N1, H9N2 and H6N1). The presence of shared H2H mutations suggests that reassortments of multiple

internal proteins have occurred between these three subtypes. Isolates of other subtypes were found to contain high numbers of adaptive mutations (shown below the red line), but their mutation repertoire seem to bear little relationship to that of the Asian group of isolates. These viruses belong to subtypes that are poorly represented in public databases; they were all isolated from Charadriiformes (gulls and shorebirds), a class of birds which accounts for only 7.5% of all avian sequences in our dataset (see Table 6-5). The presence of isolates with multiple H2H adaptive mutations in this underrepresented group of hosts suggests that lineages rich in H2H mutations may be circulating in the wild bird population, practically undetected. Although it is understandable that influenza research focuses primarily on reservoir populations and economically important domestic species, it appears that by ignoring other bird populations we may fail to identify strains that are potentially transmissible to humans.

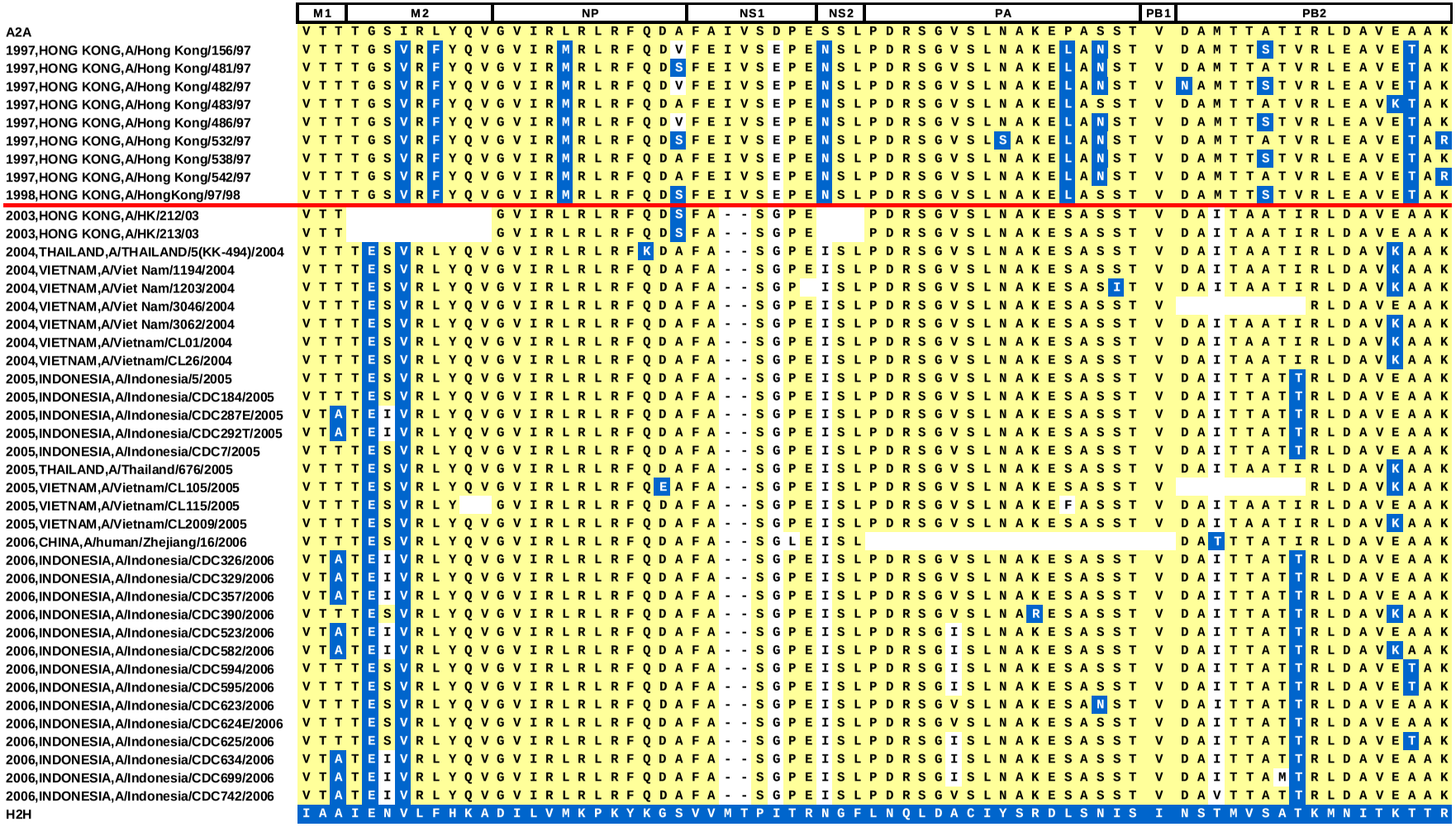
(Facing page)

Figure 6-7: Timeline of adaptation to H2H transmission for the influenza A proteome. Genomic adaptation signatures for human isolates between 1918 and 1972 are arranged in chronological order. Subtype, year and country of isolation, and isolate name are shown in the first column. The remaining columns show residues at all characteristic sites, in the same order as that given in Table 6-4. A2A characteristic mutations are shown on a dark blue background, H2H mutations on a yellow background, while all other variants are on white. Blank cells represent unknown residues in incompletely sequenced proteomes. Consensus signatures for A2A and H2H proteomes are shown in the first and last row, respectively. Red horizontal lines indicate the start of the 1957 and 1968 pandemics, which introduced the H2N2 and H3N2 subtypes respectively.


M1
M2
NP
NS1
NS2
PA
PB1
PB2
A2A
H1N1,1918,USA,A/Brevig Mission/1/1918
H1N1,1933,UK,A/Wilson-Smith/1933
H1N1,1933,UK,A/WSN/1933 TS61
H1N1,1934,PUERTO RICO,A/Puerto Rico/8/34
H1N1,1935,AUSTRALIA,A/Melbourne/1935
H1N1,1940,USA,A/Hickox/1940
H1N1,1942,USA,A/Bellamy/42
H1N1,1943,USA,A/WEISS/43
H1N1,1946,AUSTRALIA,A/Cam/1946
H1N1,1947,USA,A/Fort Monmouth/1/47
H1N1,1950,USA,A/FW/50
H1N1,1954,MALAYSIA,A/Malaysia/54
H1N1,1957,USA,A/DENVER/1957
H2N2,1957,CHILE,A/Chile/13/57
H2N2,1957,RUSSIA,A/Leningrad/134/17/57
H2N2,1958,USA,A/Albany/6/58
H2N2,1959,AUSTRALIA,A/Victoria/15681/59
H2N2,1960,USA,A/ANN ARBOR/6/60
H2N2,1961,PANAMA,A/Panama/1/61
H2N2,1964,TAIWAN,A/Taiwan/1964
H2N2,1965,USA,A/Albany/1/65
H2N2,1966,PANAMA,A/Panama/1/66
H2N2,1967,USA,A/Georgia/1/67
H2N2,1968,KOREA,A/Korea/426/68
H2N2,1968,USA,A/Berkeley/1/68
H3N2,1968,HONG KONG,A/Hong Kong/1/68
H3N2,1968,AUSTRALIA,A/NT/60/1968
H3N2,1968,AUSTRALIA,A/Victoria/1968
H3N2,1968,CHINA,A/Beijing/1/68
H3N2,1968,PANAMA,A/Panama/1/68
H3N2,1968,RUSSIA,A/USSR/039/68
H3N2,1969,BRAZIL,A/Rio/6/69
H3N2,1969,HONG KONG,A/Hong Kong/3/69
H3N2,1969,TAIWAN,A/Taiwan/1/69
H3N2,1970,TAIWAN,A/Taiwan/2/70
H3,1971,USA,A/ Memphis/1/71
H3N2,1971,HONG KONG,A/Hong Kong/46/71
H3N2,1971,TAIWAN,A/Taiwan/3/71
H3N2,1971,USA,A/Memphis/2/71
H3N2,1972,CHINA,A/Guandong/243/72
H3N2,1972,HONG KONG,A/Hong Kong/50/72
H3N2,1972,JAPAN,A/Tokyo/31/72
H3N2,1972,RUSSIA,A/Udorn/307/1972
H3N2,1972,UK,A/England/42/72
H2H

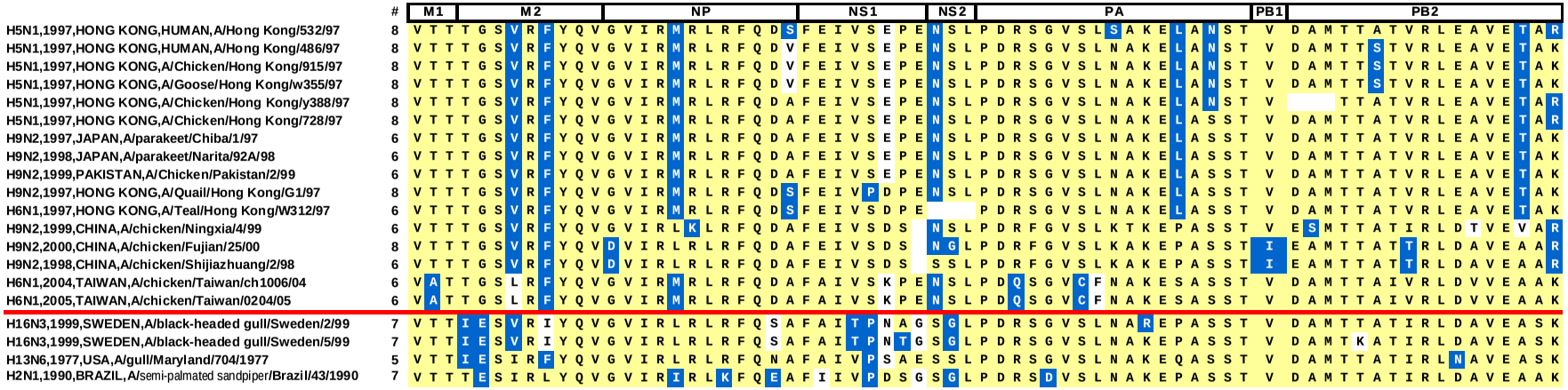
(Facing page)

Figure 6-8: Adaptation signatures of human-isolated H5N1 influenza A proteomes.
Adaptation signatures of H5N1 sequences that infected humans in the period 1997-2006 are shown. For display clarity and conciseness, a number of identical signatures were removed from this set. The same colouring scheme was used as in Figure 6-7. The red horizontal line separates the early wave of infections in Hong Kong (1997-8) from more recent South-East Asian infections (since 2003).


M1
M2
NP
NS1
NS2
PA
PB1
PB2
A2A
1997,HONG KONG,A/Hong Kong/156/97
1997,HONG KONG,A/Hong Kong/481/97
1997,HONG KONG,A/Hong Kong/482/97
1997,HONG KONG,A/Hong Kong/483/97
1997,HONG KONG,A/Hong Kong/486/97
1997,HONG KONG,A/Hong Kong/532/97
1997,HONG KONG,A/Hong Kong/538/97
1997,HONG KONG,A/Hong Kong/542/97
1998,HONG KONG,A/HongKong/97/98
2003,HONG KONG,A/HK/212/03
2003,HONG KONG,A/HK/213/03
2004,THAILAND,A/THAILAND/5(KK-494)/2004
2004,VIETNAM,A/Viet Nam/1194/2004
2004,VIETNAM,A/Viet Nam/1203/2004
2004,VIETNAM,A/Viet Nam/3046/2004
2004,VIETNAM,A/Viet Nam/3062/2004
2004,VIETNAM,A/Vietnam/CL01/2004
2004,VIETNAM,A/Vietnam/CL26/2004
2005,INDONESIA,A/Indonesia/5/2005
2005,INDONESIA,A/Indonesia/CDC184/2005
2005,INDONESIA,A/Indonesia/CDC287E/2005
2005,INDONESIA,A/Indonesia/CDC292T/2005
2005,INDONESIA,A/Indonesia/CDC7/2005
2005,THAILAND,A/Thailand/676/2005
2005,VIETNAM,A/Vietnam/CL105/2005
2005,VIETNAM,A/Vietnam/CL115/2005
2005,VIETNAM,A/Vietnam/CL2009/2005
2006,CHINA,A/human/Zhejiang/16/2006
2006,INDONESIA,A/Indonesia/CDC326/2006
2006,INDONESIA,A/Indonesia/CDC329/2006
2006,INDONESIA,A/Indonesia/CDC357/2006
2006,INDONESIA,A/Indonesia/CDC390/2006
2006,INDONESIA,A/Indonesia/CDC523/2006
2006,INDONESIA,A/Indonesia/CDC582/2006
2006,INDONESIA,A/Indonesia/CDC594/2006
2006,INDONESIA,A/Indonesia/CDC595/2006
2006,INDONESIA,A/Indonesia/CDC623/2006
2006,INDONESIA,A/Indonesia/CDC624E/2006
2006,INDONESIA,A/Indonesia/CDC625/2006
2006,INDONESIA,A/Indonesia/CDC634/2006
2006,INDONESIA,A/Indonesia/CDC699/2006
2006,INDONESIA,A/Indonesia/CDC742/2006
H2H

(Facing page)

Figure 6-9: Adaptation signatures of selected avian influenza A proteomes containing multiple H2H mutations.
In this figure, we show selected signatures of avian proteomes that were found to contain 5 or more H2H mutations. For conciseness, a number of similar signatures were removed from this set. Subtype, year and country of isolation, host (for avian viruses isolated in humans) and isolate name are shown in the first column, while the second columns shows the number of H2H mutations. The remaining columns show the signature residues, using the same colouring scheme as in Figure 6-7. Asian strains of subtypes H5N1, H9N2 and H6N1 are placed above the red horizontal line, while signatures sampled from gulls and shorebirds are shown below the red line.


#
M1
M2
NP
NS1
NS2
PA
PB1
PB2
H5N1,1997,HONG KONG,HUMAN,A/Hong Kong/532/97 8
H5N1,1997,HONG KONG,HUMAN,A/Hong Kong/486/97 8
H5N1,1997,HONG KONG,A/Chicken/Hong Kong/915/97 8
H5N1,1997,HONG KONG,A/Goose/Hong Kong/w355/97 8
H5N1,1997,HONG KONG,A/Chicken/Hong Kong/y388/97 8
H5N1,1997,HONG KONG,A/Chicken/Hong Kong/728/97 8
H9N2,1997,JAPAN,A/parakeet/Chiba/1/97 6
H9N2,1998,JAPAN,A/parakeet/Narita/92A/98 6
H9N2,1999,PAKISTAN,A/Chicken/Pakistan/2/99 6
H9N2,1997,HONG KONG,A/Quail/Hong Kong/G1/97 8
H6N1,1997,HONG KONG,A/Teal/Hong Kong/W312/97 6
H9N2,1999,CHINA,A/chicken/Ningxia/4/99 6
H9N2,2000,CHINA,A/chicken/Fujian/25/00 8
H9N2,1998,CHINA,A/chicken/Shijiazhuang/2/98 6
H6N1,2004,TAIWAN,A/chicken/Taiwan/ch1006/04 6
H6N1,2005,TAIWAN,A/chicken/Taiwan/0204/05 6
H16N3,1999,SWEDEN,A/black-headed gull/Sweden/2/99 7
H16N3,1999,SWEDEN,A/black-headed gull/Sweden/5/99 7
H13N6,1977,USA,A/gull/Maryland/704/1977 5
H2N1,1990,BRAZIL,A/semi-palmated sandpiper/Brazil/43/1990 7

Order	Count	%
Anseriformes	6734	45.7
Charadriiformes	1099	7.5
Ciconiiformes	57	0.4
Columbiformes	152	1.0
Coraciiformes	6	0.04
Falconiformes	23	0.2
Galliformes		
Chicken	4772	32.4
Turkey	716	4.9
Other	676	4.6
Passeriformes	143	1.0
Podicipediformes	43	0.3
Procellariiformes	65	0.4
Psittaciformes	46	0.3
Struthioniformes	46	0.3
Unspecified	150	1.0
Total	14728	

Table 6-5: Distribution of influenza A protein sequences among avian orders.
Sequence count and percentage is shown for every order of birds for which sequences were available in our dataset. Anseriformes sequences include domestic as well as wild waterfowl isolates.

In summary, a novel, sensitive method based on MI enabled us to construct a comprehensive catalogue of characteristic mutations, based on sequence data spanning a period of nearly a century. By applying this catalogue to the construction of adaptive signatures, we were able to reveal a number of previously unseen patterns. First, the two human lineages have stable, highly conserved constellations of internal proteins, unlikely to reassort. Second, the ability of H5N1 viruses to infect humans correlates with an unusually high level of H2H mutations. Third, there are other subtypes that circulate in avian populations with similarly high levels of adaptive mutations, which may indicate their potential to infect humans.

6.4 Discussion

6.4.1 Characteristic sites catalogue

The analysis described in Chapter 6 produced the most complete catalogue of H2H adaptive mutations published to date. This catalogue describes a complex landscape of adaptations, involving a greater number of proteins than reported by any of the previous studies. The presence of characteristic sites in eight of the nine internal influenza proteins indicates that host adaptation is highly complex and systemic in nature, requiring the participation of the whole genomic ensemble. The gradual emergence of H2H mutations in the three decades after the Spanish influenza pandemic suggests that many of these adaptive mutations are not essential individually. However, their high level of conservation over the following decades strongly implies their important role in adapting to human hosts. A plausible explanation is that the 1918 H1N1 genome contained a non-optimal set of essential components for human transmission, which has been refined over time to improve equilibrium between the virus and the host. This model does not imply that any of the 1918 mutations are individually sufficient, or even necessary, for human-to-human transmission; nor does it imply that there is only a single set of mutations capable of conferring the necessary properties to the virus. Rather, several combinations of concurrent H2H mutations may be capable of enabling sufficiently efficient infection and transmission in humans to allow the gradual refinement of the adaptive mutation repertoire. Our catalogue of characteristic sites, derived from the analysis of 90 years of refinements in human lineages, can therefore be a valuable tool for assessing the potential of zoonotic viruses to infect and circulate amongst humans.

It is unlikely that all characteristic sites identified in our catalogue play independent roles. The presence of several H2H characteristic sites in putative protein-binding domains (particularly within the RNP complex) suggests that some mutations may have co-evolved as a result of preferred structural interactions. Unfortunately, we are currently unable to identify these interactions: the structural information obtainable for the internal proteins is insufficient to map these domains accurately, and too few genomic sequences prior to 1950 are available

for the statistical identification of co-evolving residues. Even so, our data clearly indicates that (a) internal protein constellations form stable lineages in humans, and (b) these lineages do not readily reassort. Such lack of reassortments is remarkable in view of the genetic similarity and overlapping geographical spread of the two lineages, and suggests a very strong interdependency between the specific elements of the constellation. Internal proteins participate in various cellular processes, such as nuclear transport, replication and virion assembly, each of which may require adaptation to the host organism. The location of characteristic sites within putative nuclear localization signals (NLS) of various SNP components supports this proposition. These findings indicate that more attention must be paid to the host adaptation of cellular mechanisms involving internal influenza proteins. Although the external HA and NA proteins are known to contain important host range determinants, such as domains that conform to human cellular receptors (Chandrasekaran *et al.* 2008), it has been shown that efficient human-to-human transmission cannot be achieved through mutations in the glycoproteins alone (Maines *et al.* 2006). Our results support the hypothesis that concurrent mutations in the internal protein constellation are a requirement for transmissibility amongst humans, although their role in host range adaptation is still poorly understood. In this context, the reassortments of the PB1 segment in multiple pandemic events may indicate that stable PB1 mutations are not required for human host adaptation. Recent research has proposed a critical role of the PB1 gene in the high virulence of the 1918 pandemic (Pappas *et al.* 2008), and it appear likely possible that the replacement of this segment plays a role in the acquisition of a novel hemagglutinin type. In spite of repeated segment replacements, the PB1 V336I mutation has emerged in both human-transmissible lineages, and has subsequently been highly conserved, suggesting the possibility that this mutation plays an important adaptive role that should be further investigated.

The unusually high density of characteristic sites in the M2 protein may be explained by its physical arrangement in the virion assembly: M2 is a transmembrane protein, thought to interact both with the internal proteins and with the host immune system. Much attention has been focused on the M2e extracellular region of this protein, which was observed to be

conserved in humans, and thus proposed as a candidate vaccine (Neirynck *et al.* 1999). Recently, further studies have claimed that M2e-based vaccines may confer immune protection against zoonotic strains (Tompkins *et al.* 2007). Our results suggest that the M2e domain, and possibly the whole of M2, is prone to developing adaptive mutations, and its conservation in the two human lineages cannot be used as an indicator of its conservation in avian viruses. In view of our incomplete knowledge of avian influenza diversity, claims of universal protection against avian strains should be regarded with caution, especially because of the ease with which reassortments occur in these viruses.

6.4.2 Assessment of avian influenza viruses

In our analysis of avian influenza A sequences, the signatures from H5N1 isolates stood out as the richest in H2H mutations. This result was not expected, and it strongly supports the utility of our characteristic site catalogue as a useful assessment tool. The comparison of 1997 Hong Kong H5N1 signatures against those of contemporary H9N2 and H6N1 isolates from the same geographical region reveals a dynamic interplay between these three subtypes, in which viral segments appear to have been transferred through reassortments (shown in Figure 6-9). This observation is supported by results of previous studies, which have proposed that the 1997 Hong Kong H5N1 epidemic followed the reassortment of H5N1 and H9N2 viruses (Guan *et al.* 1999), and that H6N1 viruses were also involved (Hoffmann *et al.* 2000). In addition, our results suggest that these reassortments may have been instrumental for the build-up of the H5N1 adaptive mutation repertoire, a hypothesis that is further corroborated by the signature of an earlier H5N1 isolate (A/duck/Minnesota/1525/1981) which contained no H2H mutations at all. Such highly dynamic composition of the avian influenza genome puts into question the validity of labeling influenza isolates exclusively by their HA and NA subtypes. H5N1 isolates of 1981, 1997 and 2004 clearly present distinct internal protein constellations, and grouping them into a homogeneous set reveals little about their ability to adapt to humans. In addition, an excessive focus on the HA/NA subtype deviates attention from the analysis of co-circulating strains with a potential for reassortment, impairing

effective surveillance of the potential for human infectivity and transmissibility. This does not mean, of course, that we should disregard the important roles played by the glycoproteins in adapting to human receptors. On the contrary, they should be considered important components of a much larger systemic ensemble of adaptations, some of which can only be modeled by new approaches that transcend current subtype definitions.

The second wave of human-infecting H5N1 viruses presents a strikingly lower number of H2H mutations than that of the 1997 Hong Kong wave, though higher than average for avian strains. Most remarkably, the two waves only share one conserved H2H mutation (M2 I28V), while all other mutations involved in the 1997 waves have been replaced by avian variants. Thus, it appears that H5N1 viruses are not only acquiring, but also losing H2H mutations through reassortments. In addition, the adaptive mutations do not appear to be particularly stable, as evidenced by the loss of the PB2 E627K mutation, implicated in replication in humans (Subbarao *et al.* 1993) and high virulence of human H5N1 infections (Hatta *et al.* 2001). Overall, there is no evidence of a trend of gradual accumulation of H2H mutations, and this may indicate that H5N1, in its current form, poses relatively a low pandemic risk. On the other hand, the abundant evidence of H5N1 reassortments raises the concern that these avian viruses may reassort with a human lineage, combining a human-adapted internal protein constellation with an immunologically novel set of glycoproteins. Such reassortants have been produced under laboratory conditions, using human H3N2 viruses, but have failed to propagate amongst mammal models (Maines *et al.* 2006). Even if reassortants acquired the ability to circulate efficiently among humans, it is impossible to predict whether they would retain the extreme pathogenicity that has characterized human H5N1 infections: like transmissibility, pathogenicity appears to be systemically determined, and may likely be affected by the replacement of internal proteins. The introduction of avian/human reassortants triggered the pandemics of 1957 and 1968, both of which had much lower mortality than the 1918 pandemic. Although advances in disease control may be partially responsible for this decrease in severity, it is plausible that the presence of an established and well-adapted internal protein constellation may mitigate the overall

pathogenicity of the pandemic strains.

While influenza researchers focus their efforts on the threats of H5N1 viruses, the avian population could be harbouring other potential threats, which may currently go undetected. Our analysis revealed four isolates with a high number of H2H mutations, which appear to be unrelated to the human-infecting Asian strains. These four isolates belonged to less common subtypes, and were all sampled from Charadriiformes, an order of birds that is relatively understudied as influenza hosts. It is likely that much of the diversity of influenza viruses in gulls and shorebirds is as yet undiscovered: the H16 subtype was only recently identified in the gull population (Fouchier *et al.* 2005). Therefore, strains containing internal protein constellations with numerous H2H mutations could be grossly underrepresented in public databases, and hence in our dataset. Although human interactions with seabirds are less extensive than those with poultry, our focus on strains that affect domestic birds may engender a skewed perception of avian influenza epidemiology. It is necessary that large-scale influenza surveillance projects sample extensively bird groups that are currently neglected, including infected individuals that appear healthy. The catalogue of characteristic sites should provide a useful tool for characterizing new lineages as they are newly sequenced.

6.5 Conclusions

The biological knowledge mining application described in this chapter constitutes an end-to-end example of "second-generation" bioinformatics process: it used a large-scale dataset, comprising over 40,000 sequences and their metadata, aggregated using the ABK approach described in Chapter 3; it employed scalable analysis methods based on mutual information, that enabled thousands of sequences to be processed on standard office computers; it integrated metadata in the analysis task, allowing comparisons to be made between groups of sequences selected rationally according to a model of co-circulation of strains; it was developed to be "biologist-friendly" in that it allowed biomedical researchers to deploy the most appropriate conceptual model to the task in hand, by organizing groups of sequences

data using the metadata of their choice. Furthermore, the AVANA tool demonstrated that the knowledge produced as a result of analysis (the catalogue of characteristic sites) can be integrated in further analysis of the dataset, generating further knowledge (the adaptive signatures in this case).

The results of this application were a resounding affirmation of the value and resolving power of "second-generation" bioinformatics. The catalogue of characteristic sites is twice as extensive as the best previous attempt, and provides virologists with valuable insights into the adaptation mechanisms of influenza A. Adaptation to transmissibility among humans is complex, systemic, and gradually acquired through an evolutionary refinement process. The high number of sites found by our analysis permitted a "high resolution" view of adaptive mutations, through the generation of adaptive signatures, which reflect the adaptive potential of specific strains and support the tracking and evaluation of emerging adaptations in zoonotic viruses. In summary, this biological knowledge mining application made significant contributions to the field of influenza virology and answered research questions of great current importance, while demonstrating the feasibility and utility of our knowledge mining approach.

7. IDENTIFICATION OF TARGETS FOR EPITOPE-BASED VACCINES

In Chapter 6, we demonstrated that the combination of the large metadata-enriched datasets produced by knowledge aggregation, combined with a scalable analysis method based on mutual information, can lead to the significant biomedical discoveries. In this chapter, we show that an equivalent biological knowledge mining pipeline can be applied to a different type of problem, producing comparable benefits. The application described in this chapter analysed large-scale sequence datasets, such as the one created in Chapter 3, Section 3.4, using a method based on peptide entropy analysis (see Chapter 5, Section 5.1), to identify peptides that are conserved in multiple populations. This conservation analysis was conducted as the first stage in the identification of potentially immunogenic conserved peptides, an essential step in the rational identification of components of epitope-based vaccines (Sette *et al.* 2001; Brusic and August 2004).

The work described in this chapter was a collaboration between several researchers. The author of this thesis designed the knowledge aggregation and conservation analysis tasks, constructed the tools for supporting these tasks (incorporated in the ABK and AVANA tools), and contributed to the definition of the overall analysis pipeline (Khan *et al.* 2006, see Appendix A). The process was applied to the study of influenza A virus by AT Heiny (Heiny *et al.* 2007, see Appendix B), and of dengue virus by AM Khan (Khan *et al.* 2008, see Appendix C), demonstrating the generality of this application. Since this chapter describes primarily work contributed by this author, detailed results contributed by collaborators are omitted, and may be found in the relevant appendices.

7.1 Background

Epitope-based vaccines contain antigenic peptides (epitopes) capable of triggering a T cell-mediated immune response, inducing immune memory (Esser *et al.* 2003; Zinkernagel and Hengartner 2004). T cell responses may be triggered by fragments of any non-self protein,

and epitopes may therefore be selected within any protein in a pathogen's proteome. Since full-proteome experimental of epitope identification is prohibitively expensive, rational computational methods can be applied to reduce the number of potential targets (Sette *et al.* 2001). A good approach is to identify pathogen peptides that are predicted to bind to human leukocyte antigen (HLA) molecules, a prerequisite for triggering a T cell response. Because many pathogens mutate rapidly, causing the emergence of new genetic variants, it is important to choose highly conserved peptide sequences, unlikely to mutate without detriment to the pathogen. In addition, since HLA polymorphism restricts the proportion of the human population that will respond to a particular antigen (Brusic and August 2004; Ovsyannikova *et al*. 2004), promiscuous epitopes capable of binding to several HLA alleles supertypes (Sette and Sidney 1999) should be selected.

7.2 Methodology Overview

The method detailed in Khan *et al.* (2006) provides a strategy for scanning full pathogen proteomes, to identify highly conserved peptide sequences that are predicted to bind to promiscuous HLA molecules. The high conservation is an indicator of the peptide's stability over time, while promiscuous binding is likely to increases vaccine coverage within the human population. The method, modelled as a Knowledge Mining process, is shown in Figure 7-2. Its first step is a knowledge aggregation task, performed by the ABK tool described in Chapter 3, which creates a large-scale dataset of protein sequences with annotation metadata. From this master dataset, major sequence groups of interest (such as co-circulating lineages or serotypes) are selected by the AVANA tool based on metadata, and analyzed for conservation. By separating groups, and finding consensus conserved regions (Novitsky *et al.* 2002), equal weighting is given to each group, thus overcoming dataset sampling biases. The definition of “group” is dependent on the pathogen under study, and may correspond to a clade, serotype or subtype. As a result, the selection of metadata to be aggregated during data collection is also pathogen-specific, as it depends on the choice of group to be analyzed.

The conservation analysis task identifies protein stretches that meet certain conservation criteria and are as long as possible, with a minimum length of nine amino acids. We begin by selecting those 9-mer peptides that meet the conservation criteria, and then gradually extend them to neighbouring residues, until they no longer meet these criteria. Overlapping conserved regions are merged only if the resulting region meets the conservation criteria. Otherwise, each is considered separately. Conservation analysis is conducted in two stages, using two different criteria:

1. Conserved regions are identified based on an identical peptide variant being observed in a proportion $\geq p_{cmin}$ of sequences (typically, p_{cmin}=80%).
2. Low-diversity regions are identified based on peptide entropy being lower than a threshold H_{cmin} (typically, H_{cmin}=1.0 for 9-mer peptide entropy), using peptide entropy as described in Section 5.1.

Only regions that meet both conservation and low-diversity criteria are selected. The purpose of low-diversity screening is to discards regions with potential for change (and thus likely to be unstable in the future) that appear to be conserved because of uneven sampling. All conservation analysis computations were performed by the AVANA tools (Section 5.3).

The immunological potential of the conserved regions is further assessed by bioinformatics methods that are outside the scope of this work, and are detailed in Khan *et al.* (2006). The process is generic and applicable to a wide variety of pathogens, as demonstrated by the results for two important pathogens, influenza A and dengue viruses, detailed in Appendices B and C.

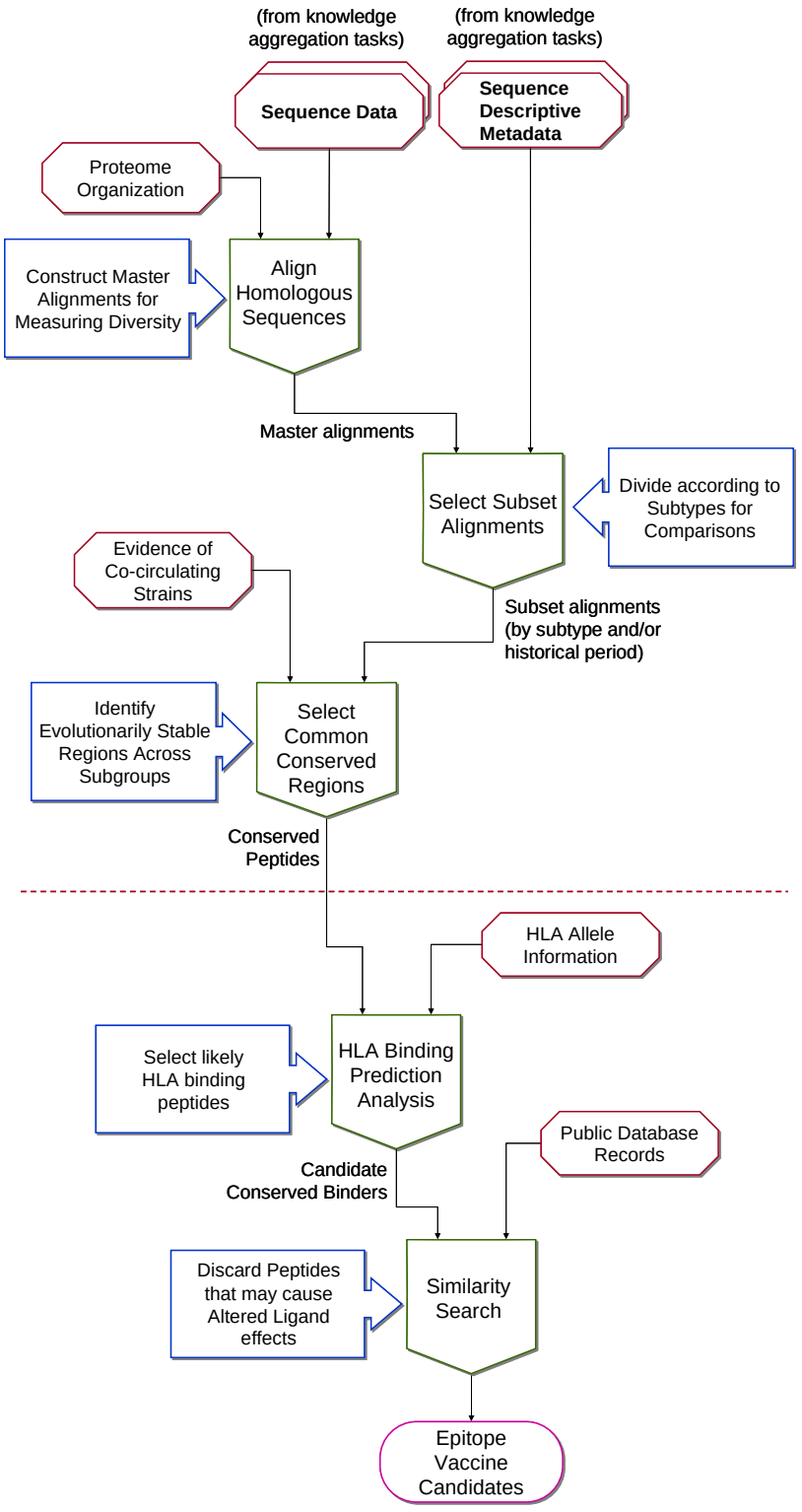


Figure 7-2: Model of the process of identification of epitope-based vaccine targets
Stages above the dotted line were performed using the entropy-based conservation analysis method described herein. The master alignments, aggregated using the ABK tool, were subdivided into co-circulating lineages by the AVANA tool, based on descriptive metadata. AVANA then identified consensus conserved regions. Downstream tasks (below the dotted line) constitute work separate from this thesis that assessed identified conserved regions for their antigenic potential and likelihood to cause altered ligands effects. See Khan (2006 and 2008) and Heiny (2007) in Appendices A, B and C for details of these downstream stages.

7.3 Applications: Influenza virus and Dengue virus

Conservation analysis was conducted by AT Heiny (2007) on the large-scale influenza A dataset (described in Chapter 3, Section 3.4), using a total of 36,343 protein human and avian sequences isolated in the last 30 years. Consensus conservation analysis was conducted on six subsets, selected based on metadata values: Human H3N2, Human H1N1, Human H1N2, Human H5N1, Avian H5N1, and other Avian. These subsets represent important currently circulating lineages. This analysis found 55 sequences of nine or more amino acids that were conserved in at least 80% in each of the subsets. Influenza polymerases (PB2, PB1, and PA), nucleoprotein (NP), and matrix (M1) proteins were particularly rich in conserved regions. Of these conserved sequences, 50 were also found to contain putative supertype HLA epitopes, and have been shortlisted for experimental validation.

In a study with similar objectives, AM Khan (2008) analyzed a dataset of 12,404 dengue virus protein sequences, aggregated using the ABK tool from GenBank and GenPept. The master alignments were subdivided into the four dengue virus serotypes (DV1 to DV4), and consensus conservation analysis was conducted using the AVANA tool. This analysis identified 44 sequences conserved in at least 80% of sequences across all serotypes. Most of the conserved peptides were found in nonstructural proteins, and a large proportion (34) exhibited extremely high (> 95%) conservation. Several conserved sequences were predicted to be immunologically relevant: 34 peptides contained predicted HLA supertype-restricted binding sequences, and are therefore candidates for further experimental studies.

7.4 Conclusion

In this chapter we have reviewed two studies, conducted by two different researchers, who applied a rational method for identifying conserved candidate epitopes in two different viruses: influenza A and dengue. Both studies were supported by a biological knowledge mining pipeline that performed knowledge aggregation methods using the ABK tool, as detailed in Chapter 3, and analyzed conservation in thousands of protein sequences, thanks to

the highly scalable peptide entropy method defined in Chapter 5, and implemented by the AVANA tool. Just as in the application described in Chapter 6, the use of aggregated metadata and of a knowledge-enabled tool was critical in the analysis of subsets according to an epidemiologically significant model. Without leveraging on metadata, the same task would require lengthy preparations of multiple datasets, introducing inflexibility in the conceptual model. For example, introducing a new epidemiologically meaningful group in the absence of metadata support would require a major rework of the datasets, while AVANA can reorganize sequence data using descriptive metadata, rapidly and efficiently.

The significant results obtained by AM Khan and AT Heiny in their studies have therefore provided independent evidence that the knowledge mining approach and information-theoretical measures defined in this thesis are applicable to a variety of analysis pipelines, and to a range of pathogens subjects. The immunological problem addressed in these studies – the identification of potential vaccine target – was significantly different from the virological questions that drove the study detailed in Chapter 6. However, both of these studies presented the hallmarks of second-generation bioinformatics: the use of large datasets to provide statistically significant results; the reliance on analysis metrics capable of handling thousands of sequences; and the use of metadata for modeling and organizing sequence data. The successful application to such different questions is clear evidence of the generality of the methods and tools described in this thesis.

8. TEXT MINING OF LITERATURE SOURCES FOR THE CURATION OF ALLERGEN DATABASES

Our final application, presented in this chapter, is profoundly different from the applications described in Chapters 6 and 7. It is an allergy application, related to the discovery of allergen cross-reactivity information for inclusion in an allergen database. However, it analyzes a completely different type of dataset, which consists of biomedical text documents rather than sequence data. This departure from the analysis of sequences is intentional: this application shows that the principles that govern "second-generation" bioinformatics are not exclusively applicable to specific types of data and analysis tasks, but rather generic and reusable. The study presented in this chapter therefore uses our biological knowledge mining approach, albeit with some differences. The upstream knowledge aggregation task is conducted by the ABK tool, which even in this case handles tens of thousands of records. As in previous applications, metadata is used to form a model, but in this case the model is constituted by text features, such as keywords, text patterns or sentences. Several sources of metadata are therefore used to annotate the records: metadata extracted from the source records (such as title, journal name); metadata generated by text analysis tasks (such as the identification of terms specified by the user); and annotation metadata, entered by the user. To enable knowledge transfer between analysis tools, these were implemented as ABK plug-ins so that metadata could be managed and augmented within the Data Management component of ABK (see Chapter 3, Section 3.3). The results of the analysis pipeline is the selection of biomedical document that are most predicted to contain the desired information. Simple user interfaces allow rapid inspection of the document, and allow the expert user to classify documents, thus augmenting the descriptive metadata. As a result, important additions to the metadata from the user occur during the analysis stage; this "knowledge feedback" – known as *active learning* – gradually improves the performance of the analysis tasks. Finally, our text mining approach is "biologist-friendly" in that every aspect of the analysis task is either controlled by the user through simple interfaces (*e.g.* the definition of text patterns, relevant vocabulary,

and search terms) or automated (*e.g.* the selection of metadata features, and machine learning classification tasks). This sets apart this from other text mining approaches, which invariably require domain knowledge to be embedded in the analysis code, and can therefore only be controlled by text mining experts.

Our text mining application has therefore demonstrated that the biological knowledge mining framework developed in this thesis can be applied to a variety of application and problem spaces. Although this study was intended as a proof-of-concept demonstration of the generality of our approaches, the results presented in this chapter strongly indicates that this text mining approach is viable, and can produce valuable savings of time and effort in real-life database curation tasks.

8.1 Background

Biological research groups around the world share their data through thousands of specialized data repositories, which focus on particular molecules, organisms or diseases. In marked contrast with large primary sequence databases such as GenBank, these "boutique" databases usually offer smaller, focused sets of richly annotated records. To ensure data content of the highest quality, these databases generally follow a manual data entry and curation (annotation) process (Fredman *et al.* 2002). Manual curation is performed by domain experts– knowledgeable scientists who are valuable and often scarce resources for their organizations. Their primary source of data is scientific literature, usually peer-reviewed journal articles. Database curators search biomedical research literature for facts of interest, and manually transfer knowledge from published papers to the database. Recently, widespread online publication of journals has dramatically improved the availability of literature (Markovitz 2000) and the automation of search operations, both of which are essential for curators. However, electronic publishing has also caused an increase in the volume of literature, which is compounded by the continuous rapid expansion of biological knowledge. As a result, the manual curation process remains a time-consuming, expensive process that is prone to omissions and inconsistencies (Rebholz-Schuhmann *et al.* 2005). This knowledge transfer

bottleneck slows down the pace of research, and therefore there is considerable interest in technological solutions that minimize the curators' involvement, or replace them altogether. In particular, text mining techniques enable various degrees of automation of the analysis of scientific literature, such as: the identification of named entities; the classification of documents; and the extraction of relevant facts (Cohen and Hersh 2005). Although they are still not capable of fully automated extraction of correct information from texts, these approaches keep improving. Yet, their adoption is very limited, for a variety of reasons. Firstly, the development of text mining tools requires technological know-how and infrastructure which are only available only to a few database curators, and not at all to average scientists. Biologists often need to work with data mining experts, who typically know little about the scientific concepts involved. Furthermore, most text mining solutions are specially designed for the task in hand, and this impair the reuse of existing software to address new tasks. This lack of reuse increases the financial and manpower cost of text mining, and delays the deployment of this technology on new problems. Finally, the usage, maintenance and customization of text mining tools are typically complex, and their performance is difficult to evaluate for researchers without data mining expertise.

8.2 Text Mining Requirements of Database Curation Processes

Curators face major challenges in all stages of the conversion from unstructured scientific literature to structured data e.g. database records. Scientific articles are highly specialized and often hard to understand even for experts in the area, making it difficult to identify all the interesting facts at the knowledge extraction stage. The variety of writing styles compounds this problem, since facts are not always clearly stated. Analyzing a paper is a lengthy exercise involving significant effort, particularly when the paper is rejected from inclusion in the database after assessment. The waste of effort can be minimized by effective selection and filtering of documents. Scientific abstracts are very valuable for evaluating the content of a paper: they provide more condensed information than the full paper text, but are information-rich, and usually summarize the main results. Although some repositories base their curation

process on scanning *all* abstracts from a range of scientific publications (Alfarano *et al.* 2005), this approach is impractical both for smaller projects and for broad research topics, so pre-selection of abstracts is highly desirable.

Recently, the use of text mining algorithms has been proposed for streamlining various aspects of the curation process. The phrase *text mining* loosely denotes the analysis of text documents by *machine learning* and *natural language processing* (NLP) algorithms. Although earlier definitions of text mining assume the automatic extraction of knowledge from text (Hearst 1999), many current implementations pragmatically aim at assisting the recovery of information from text. The text mining process has four stages (de Bruijn and Martin 2002), ordered by increasing complexity:

a) *document categorization* identifies documents relevant to given topics

b) *named entity tagging* isolates concepts and names important to the problem space

c) *fact extraction* extracts items of meaningful knowledge

d) *collection-wide extraction* discovers new knowledge by correlating facts from multiple documents.

Fact extraction systems are suitable for automating the annotation of database entries. Recent promising results include the successful annotation of genes and proteins, and extraction of biological interactions (de Bruijn and Martin 2002; Hofmann and Schomburg 2005). However, even the best state-of-the-art systems are not as accurate as human curators. Automatic maintenance of high-quality databases demands high *precision* (high proportion of true positives), which usually comes at the expense of lower *recall* (capturing a smaller portion of all published knowledge). This trade-off is evident in a study of automated annotation of enzymes (Hofmann and Schomburg 2005), which deemed 92% precision and 50% recall as "sufficient for inclusion in a high-quality database". Indeed, a high precision is necessary if the data in the repository is to be trusted, but 50% recall omits half of all available knowledge, which is an unacceptable trade-off for most human curators. These findings suggest that curators are still necessary mediators between published literature and databases.

We propose that the curatorial work can be effectively supported by *document categorization* systems that select and filter documents, reducing the workload but not the quality of results. A key factor is to leverage on the curator's tolerance of classification errors. Current classifier algorithms are capable of relatively high recall, at the cost of reduced precision; in other words, they can find a high percentage of all available knowledge, if one can accept somewhat "noisy" results. Since human curators are highly effective as quality filters, it is often acceptable to relax precision requirements to achieve higher recall. Supported by text mining systems, curators can then rapidly inspect and discard irrelevant documents, thus significantly improving annotation speed.

The most common approaches to document categorization involve *machine-learning* classifiers, trained with manually-annotated sets of documents that contain both documents of interest (positives) and other documents (negatives). The best results to date have been obtained as a result of laborious choices of algorithms and document features, to suit the specifics of a particular problem. One prize-winning system, for example, used a combination of sophisticated techniques, and non-obvious document features (figure captions), which are difficult to extract (Regev *et al.* 2003). Such powerful systems are clearly hard to reuse in different contexts, and can only be developed by highly-specialized programmers, often with NLP expertise in natural language processing. Surprisingly, very little research has addressed the need for text mining systems that can be used for a variety of diverse tasks, by curators with limited programming and linguistic expertise. Cohen and Hersh (2005) have stated that current text mining research is biased towards "evaluations based on system output independent of user needs". They have identified the major challenge in this field: bridging the gap between text mining researchers and database curators, thus "helping biomedical researchers to solve real-world problems that are inhibiting the pace of research". They highlighted the need for improvement in: a) access to full text articles rather than abstracts, b) identification of the features for analyzing text, c) measurement of true value to users, d) cooperation between end users and text mining researchers. We have identified usability and reusability of text mining tools as additional areas for improvement. Currently, even the most

accurate algorithms cannot benefit database curators, unless text mining experts are available for tools development. Of the thousands of specialized databases currently online, very few such as BIND (Alfarano *et al.*, 2005) can count on the availability of such experts. Curators need reusable, configurable and customizable tools that serve their needs, without requiring them to become skilled programmers.

Limited access and availability of full text articles are serious barriers to effective text mining. Even when the full text of a research paper is available, the need for subscription limits the application of automated data mining tools. For the time being, most biomedical discovery from text is likely to remain strongly reliant on journal abstracts, which are freely available from large repositories. PubMed contains abstracts of articles published in a large number of biomedical journals (over 17 million abstracts as of December 2007). Even when high-coverage full-text indexing becomes available, it is likely that searches on abstracts will still represent a key preliminary analysis.

8.3 Reusable Text Mining based on Active Learning

We propose to meet the needs of curators with a new class of document categorization systems, whose characteristics can be summarized as follows:

- Single, intuitive user interface, not requiring programming or linguistic abilities
- Ability to connect to major databases (e.g. PubMed) and retrieve documents transparently, from simple user-driven query mechanisms
- Simple user interface for rapid curator annotation of positives and negatives
- Simple mechanisms for capturing user knowledge where required (e.g. dictionaries of named entities)
- Ability to learn gradually, incrementally and interactively from curator's annotations, without requiring a large initial training corpus
- Ability to automate key classification decisions, such as feature selection, classifier parameter and so on, so that technological complexity is hidden from the user

We combined these characteristics into a multi-stage knowledge mining process, shown in Figure 8-1. The source data, in the form of plain-text biomedical abstracts, is retrieved by the Knowledge Aggregation system presented in Chapter 3. To enhance the specificity of document retrieval, the system performs post-retrieval filtering of documents returned as a result of a broad query. The abstracts are broken down lexically, and then filtered based on the recognition of keywords or patterns. The systems can support complex selection criteria that are impractical to specify as search queries, such as "all documents which contain the name of an influenza protein and two different geographical location names". The user can supply and control the needed keyword vocabularies and patterns. The filtered documents are subsequently passed to a classification process, in which a trained classifier ranks them by relevance. The process of feature selection and ranking is statistical and fully automated, and may include the use of multiple classifiers, which are automatically evaluated by the system. Following ranking, the curator reads top-ranked documents to extract the desired knowledge.

The abstract-reading interface allows the curator to mark the viewed documents as positive or negatives, thus augmenting the corpus of annotated documents. This simple mechanism enables an important feature not seen in previous applications: a feedback loop that injects user-generated knowledge (the document annotations) into the classification process. This feedback process, known as *active learning* (Cohn *et al.* 1994), enables prediction accuracy to improve with time, as the user annotates more documents. The active learning approach has been successfully applied to text classification tasks (Tong and Koller 2001). The corpus of annotated documents is used to perform two functions: to identify which document features (such as words, phrases etc) are to be used for classification, and to train the classifier for subsequent re-ranking of the remaining documents.

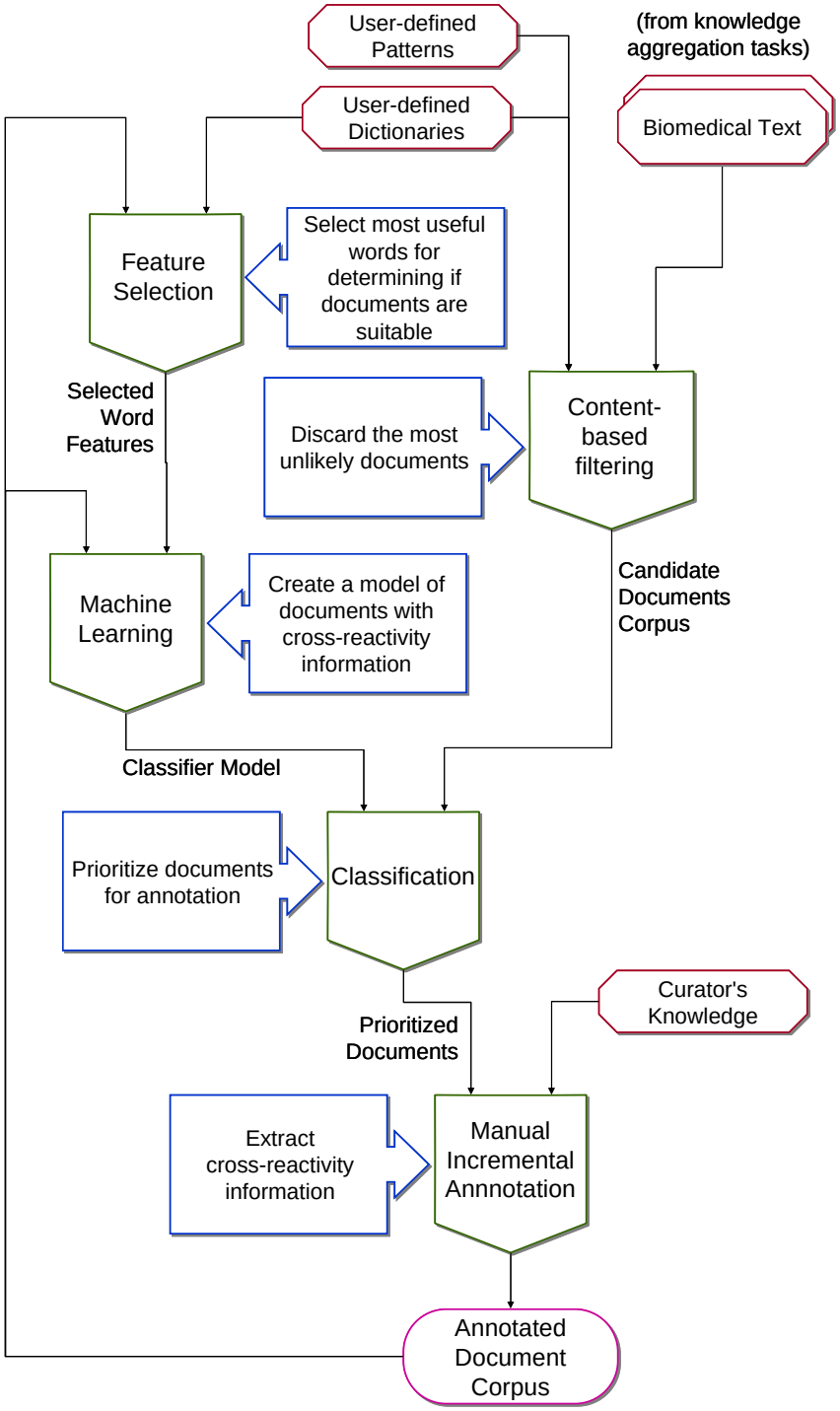


Figure 8-1: Knowledge Mining Model for the Reusable Text Mining Workflow

8.4 Materials and Methods

We developed a proof-of-concept system, based on the ABK system, to demonstrate the utility of our method on a specific real-world curation task, and measure its performance. In

particular, we aimed to demonstrate that standard off-the-shelf algorithm, combined with a choice of document features that is generic and reusable, could deliver an appropriate level of classification performance to support the manual curation task. We did not aim to produce a complete working system, and did not address the active learning process, which will be the subject of further research.

In this study, we compared two commonly used generic machine learning classifiers. Documents were pre-filtered by user queries, and ranked using statistically selected features. We compared the use of two types of features, and four scoring functions, to investigate which combination therein offered classification performance advantages.

8.4.1 Curation Task Overview

As a case study for our method, we addressed the curation needs of the ALLERDB database (Zhang *et al.*, 2006). ALLERDB contains records of human allergen proteins, extracted from literature and enriched with annotations on the biochemical properties of these allergens. We focused on the specific task of identifying information on allergen cross-reactivity. Cross-reactive allergens share structural similarities at molecular level, causing the immune system of certain individuals to react to multiple allergens (Brusic, Petrovsky *et al*. 2003). ALLERDB stores cross-reactivity information, used for allergen avoidance in patients with severe allergies.

Our document categorization task was to identify all relevant PubMed abstracts that report allergen cross-reactivity. This information generally involves two named allergens, and a statement describing cross-reactive properties. Cross-reactivity statements are not expressed consistently – some abstracts contain a clear sentence with the words “cross” and “reactivity” (or derivatives), but others imply cross-reactivity indirectly. The identification of named entities was supported by the WHO/IUIS Allergen Nomenclature (Hoffman *et al.* 1994), a naming standard for allergens. Allergen identifiers consist of a capitalized 3-letter word, followed by one lowercase letter and an integer (e.g. “Mal d 1”). The standard allows some variations (such as “Mala f 1” and “Pru av 3”). The IUIS nomenclature also provides an

"official list" of over 600 allergens, which we adapted for use as an ontology. The IUIS nomenclature was not in use before 1994 and has being gradually adopted since (Fig. 8-2).

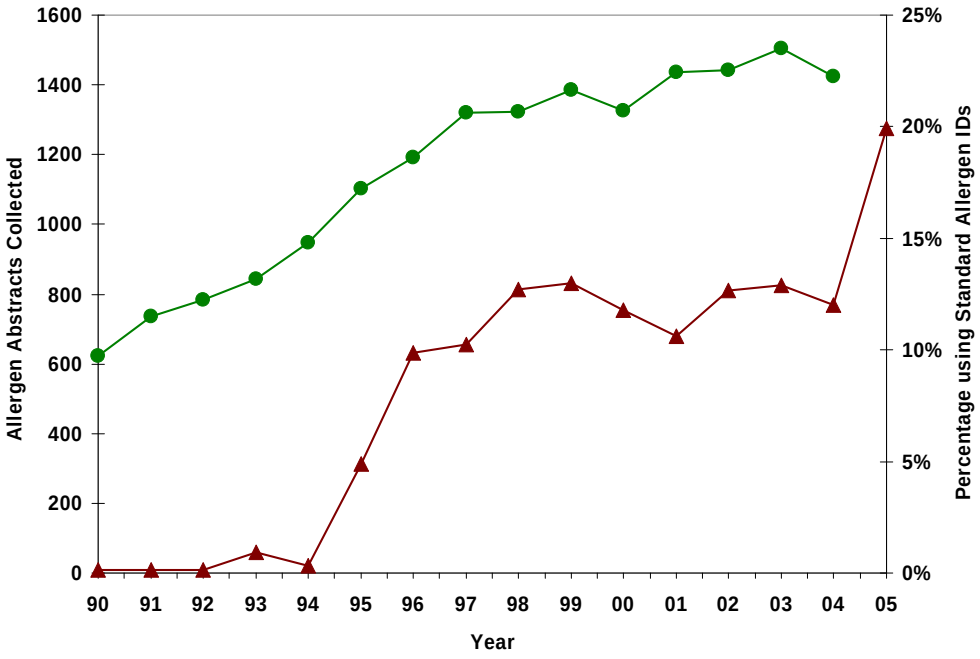


Figure 8-2: Percentage of abstracts that use IUIS allergen identifiers (triangles) and total number of abstract in the corpus (circles) for each year since 1990.
Usage of standard identifiers became widespread from 1994, and is currently around 20%. Many abstracts in the corpus have no mention of specific allergens.

8.4.2 Corpus Collection and Annotation

Our system uses the Aggregator of Biological Knowledge (ABK), detailed in Chapter 3. In this study study, the ABK mediator framework delivered user queries to PubMed, retrieving results as XML documents. We defined rules for extracting the abstract text, title, journal name and year of publication. We developed a number of reusable text analysis tools, to form a simple literature analysis workbench for conducting our study.

ABK collected 26,997 PubMed abstracts containing the word "allergen", and automatically extracted their abstract text. Named entities were identified by an ABK plug-in (the Text Analyzer Tool) which performed generic text analysis tasks, such as identifying sentences, and matching regular expressions and keywords from user-supplied lists. The tool was configured to find keywords such as "cross" and "reactive". It also found identifiers, both

from the IUIS "official list", and by matching the IUIS nomenclature pattern with a regular expression. This basic analysis uncovered 71 identifiers used in literature but not included in the IUIS official list, showing it lags behind current usage. We finally filtered documents that contained at least two different named allergen identifiers, forming a corpus of 584 abstracts.

The corpus was manually annotated by a curator, to separate positives and negatives (a positive is defined as an abstract that contains information on cross-reactivity between allergens). The annotation process was supported by the Corpus Annotator Tool, an ABK plug-in (Fig. 8-3). This tool displays the abstract text, highlighting the named entity features discovered by prior steps. Highlighting helps focus the curator's attention to key terms, and speeds up annotation.

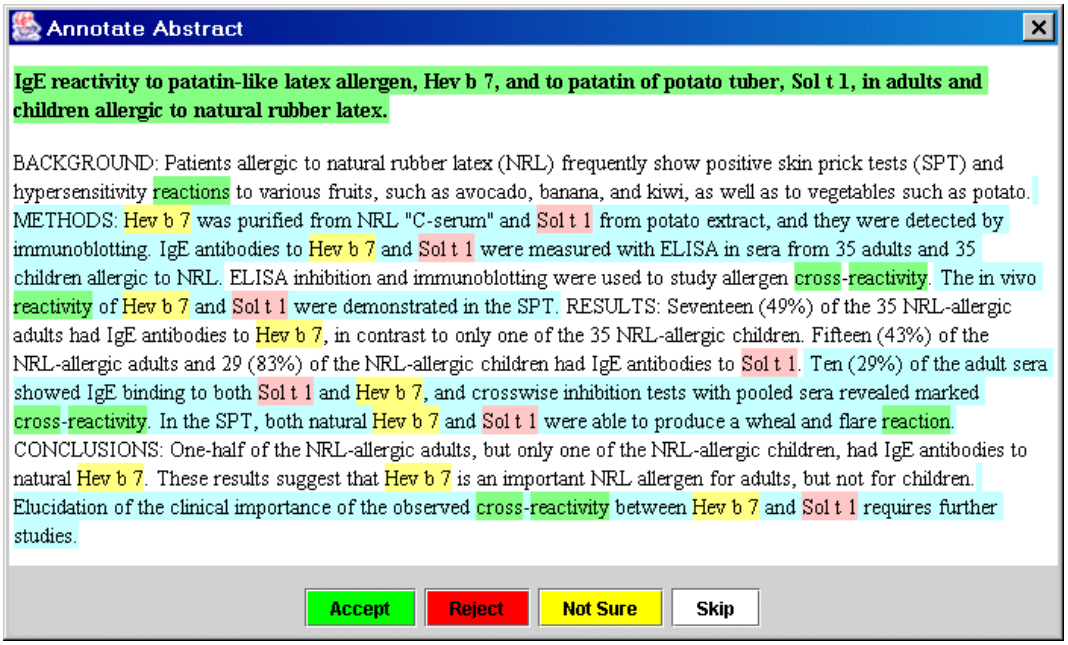


Figure 8-3: Screenshot of the ABK Corpus Annotator Tool.
Features highlighted include: IUIS allergen identifiers (yellow), other allergen identifiers (pink), cross-reactivity keywords (green), and sentences containing at least two identifiers (light blue).

Annotation of each abstract is a straightforward task: a button click determines if the abstract is a positive or negative, while a double click on the text selects key sentences– those sentences which capture the cross-reactivity information. Annotation of the full corpus by an expert identified 73 positives and 511 negatives. Only 39 positives captured cross-reactivity information in a single key sentence, while 28 required two key sentences, and the remaining

6 contained three or four key sentences. A higher number of key sentences indicate that the abstract is vaguely worded, which was sometimes hard to interpret even for the curator. Six positives did not contain the words “cross” and/or “reactive” (or their derivatives). We collected statistics on the position of key sentences within the abstracts, with the intent of investigating if this information can be used bias score features. Key sentences were significantly more likely to be found in the last third of the abstract than in the rest (Fig. 8-4).

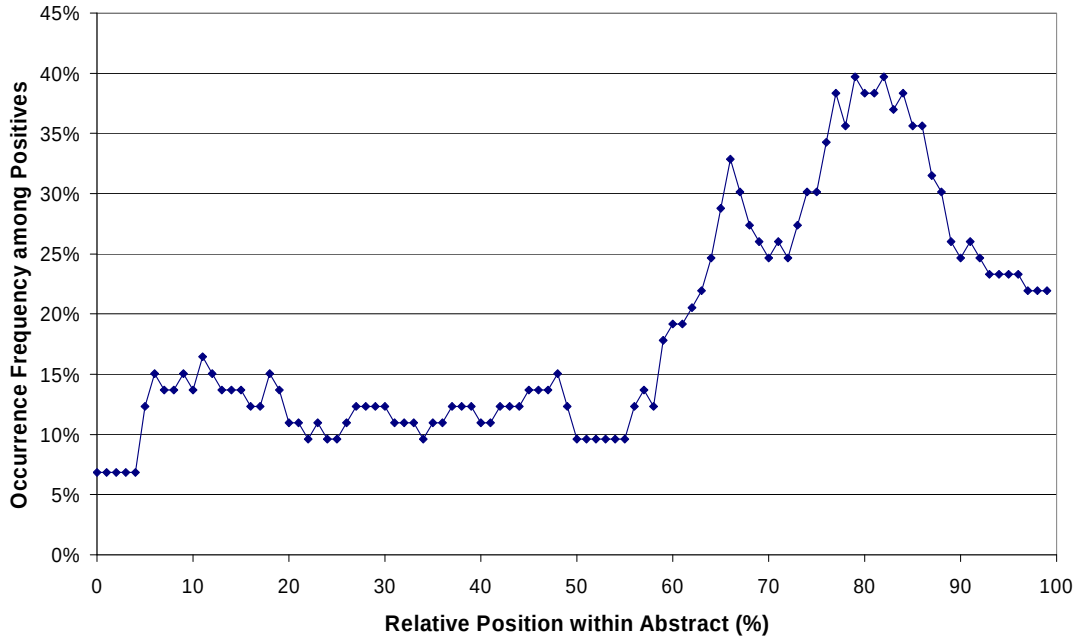


Figure 8-4: Key sentence occurrences in different parts of abstracts.
Each abstract was divided into 100 bins, and a value of 1 was assigned to each bin that overlapped with a key sentence.

8.4.3 Feature Selection and Scoring

To train classifiers, we compared two types of features: single-word and composite features, the latter consisting of group of words that co-occur frequently in sentences. This comparison aimed at testing whether the widely used bag-of-words approach (Joachims, 1998) has inherent weaknesses. The presence of a phrase such as “high blood pressure” in an abstract is clearly more informative than the presence of each constituent word. The identification of such word combinations usually demands linguistic analysis. However, commonly available statistical algorithms are able to identify sets of frequently co-occurring words, known as

frequent itemsets (Agrawal *et al.* 1993), without linguistic analysis. Frequent itemsets can be used as composite features (Deshpande and Karypis 2002).

The Text Analyzer Tool split each abstract sentence into words, discarding stop words and words beginning with digits. The remaining words were stemmed by the Porter stemmer algorithm (Porter 1980), reducing term variants (e.g. “analysis”, “analyses”, and “analyze”) to their common roots. For each abstract sentence, the Sentence Transaction Tool (an ABK plug-in) produced a transaction record, consisting of all stemmed words, without repetitions. Separate positive and negative transaction files were produced; only key sentence transactions were included for positives. The transactions were analyzed by the Apriori algorithm (Agrawal *et al.* 1993). In positive examples, 1547 itemsets (chiefly combinations of the most frequent words) had at least 5% statistical support; negatives were more heterogeneous (623 itemsets with support of 1% or above). To increase generality, the feature lists were reduced by excluding 58 words specific to major sources of human allergens (such as “dog”, “dust” and “cockroach”), based on the assumption that they were irrelevant towards cross-reactivity classification. This exclusion list is problem-specific and cannot be automated; however, it can be easily carried out by a curator, given a sufficiently intuitive user interface.

To select the most informative features, we generated feature *score vectors* using the Abstract Statistics Tool plug-in. Each vector comprised an abstract class tag (negative or positive) and a value for each frequent itemset: 1 if the itemset could be found in the abstract, 0 otherwise. The feature vectors were used to measure the information gain of each itemset and, as a result, we selected the top 64 features (following the heuristics of using approximately ten training examples for each feature). We used the same process to identify single-word features, by configuring Apriori to identify frequent itemsets of length one.

The selected features were used to produce data files for training and testing the classifiers. Each record consisted of a vector containing a score for each feature, and a class identifier. We experimented with four different score functions:

1. PRESENCE. Score is 1 if the feature is found in the abstract, 0 otherwise.
2. COUNT. Score is the number of times the feature is found in the abstract.

3. POSITION. Same as COUNT, but score is doubled for occurrences in the last 35% of the abstract (based on results shown in Fig. 3).
4. COLOCATION. Same as COUNT, but score is doubled in sentences that contain one allergen identifier, and quadrupled in those with two or more.

8.4.4 Document Classification

We used the resulting data files to train and test two types of classifiers, which are representative of highly diverse approaches to machine learning:

1. Artificial Neural Network (ANN). We chose an ANN based on a Probabilistic Neural Network (PNN) architecture, using a genetic algorithm for determining appropriate feature smoothing factors. This ANN is available in the commercial Neuroshell 2 suite (http://www.wardsystems.com/neuroshell2.asp)
2. Decision Tree (CART). We included a decision tree classifier, using the CART 5.0 package (http://www.salford-systems.com/cart.php). A cost of 4.0 was assigned to misclassified positives.

Classifier performance was assessed in terms of recall (R) and precision (P), using a test set consisting of 30% of the examples, randomly chosen by the classifier. As we previously stated, our main objective is to pre-select documents before manual curation, and the intervention of a human curator allows the precision requirements to be relaxed, privileging higher recall. We set performance targets to $R > 75\%$ and $P > 40\%$, which was deemed to be a reasonable trade-off, when accounting for the time necessary for a curator to visually discard false positives.

8.5 Results and Discussion

We observed that CART builds its decision tree almost solely on features derived from positives, while the ANN classifier also recognizes patterns in negatives. These differences account for several of the variations in classifier performances, which are shown in Table 8-1.

The most important result is that both types of classifiers exceeded our performance criteria when used with both single-word and composite features, without using any special scoring functions. Fig. 8-5 shows that ANN classifiers are considerably more precise than CART classifiers. However, lowering the precision threshold (and forcing the human curator to manually discard more false positives) permits the use of CART classifiers, which increased the recall by about 10%. The means that an additional 10% of knowledge is incorporated in the database: there is a clear trade-off between human effort and database coverage. CART's lower precision is largely due to its dependence on recognizing positives. On the other hand, ANN shows higher precision when using single-word features, many of which were derived from negatives. Although systematic reliance on negative features can actually decrease performance (e.g. when classifying diverse documents), we found that our corpus was representative of allergen-related PubMed abstracts.

The use of different scoring function showed varied impact, and in some case they impacted classification negatively. The COUNT function presented no performance advantage over PRESENCE. Interestingly, both POSITION and COLOCATION boosted the performance of CART classifiers, but brought no benefit to ANN– probably because these functions primarily boost recognition of positives. The high impact of POSITION when using single-word features with CART indicates that the presence of certain words in the last third of the abstract is a stronger indicator that the presence of positive-related phrases. Overall, the performance of ANN classifiers is appropriate to support curation tasks. Although ANN performed best without applying composite features, the highest recall figures were obtained applying COLOCATION to the CART classifier.This indicates that combinations of classifiers could yield even higher performance, a hypothesis that will be explored further.

(A) Classifier Performance using **Composite** Features

		PRESENCE	COUNT	POSITION	COLOCATION
ANN	Recall	78.9%	79.0%	63.2%	78.9%
	Precision	68.2%	68.2%	60.0%	50.0%
CART	Recall	80.0%	76.5%	64.7%	88.2%
	Precision	41.4%	50.0%	44.0%	44.1%

(B) Classifier Performance using **Single Word** Features

		PRESENCE	COUNT	POSITION	COLOCATION
ANN	Recall	78.9%	63.2%	63.2%	68.4%
	Precision	93.8%	92.3%	63.2%	81.3%
CART	Recall	76.5%	76.5%	88.2%	64.7%
	Precision	44.8%	44.8%	46.9%	44.0%

Table 8-1: Classifier performances.
Results were obtained using composite features (A), and single-word features (B). Recall and precision against a random test dataset are shown for each scoring function. The best performance figures are circled.

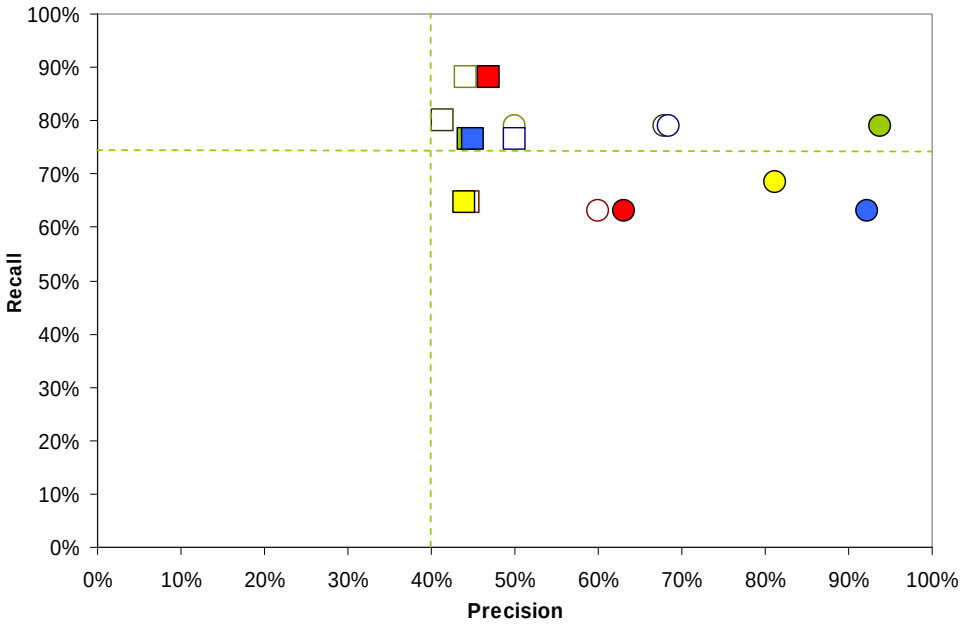


Figure 8-5: Plot comparing classifier performance figures as reported in Table 1.
ANN classifiers are represented by circles, and CART by squares. Solid markers show use of single-word features, and unfilled markers denote composite features. The dotted lines show the predetermined performance targets for our classifiers.

Finally, we investigated whether our results were biased by the selection of our pre-

filtering technique (i.e. selection based on IUIS identifiers), by applying the three top-performing ANN classifiers to the full set of 26,997 retrieved PubMed abstracts, which includes positives that do not use the IUIS standard. We then compared the classifiers' prediction against a list of abstracts that do not use the IUIS nomenclature, previously identified during manual curation of the ALLERGEN database. We found that ANN classifiers could identify 91% of these abstracts, indicating that pre-filtering bias was not a significant issue.

8.6 Conclusion

In this chapter, we have designed a biological knowledge mining pipeline for generic and reusable text mining in support of biological database curation, and implemented it as a proof-of-concept on the ABK knowledge aggregation platform. The knowledge flow through the pipeline was supported by the ABK metadata management system, which made it possible to augment the metadata with user annotations.

Our results show that our metadata-enabled approach is able to produce the desired level of performance using generic data mining algorithms, without any task-specific customization of the analysis task. This achievement sets our approach apart from most current text mining implementations, which are generally assumed to require domain-specific knowledge to be built into the analysis software, and therefore place text mining solutions out of the reach of the average biomedical research. The work presented here, on the other hand, shows that models based on user-contributed expert knowledge are able to produce substantial time and effort savings, providing further evidence that bioinformatics needs to put expert users in control of the analysis pipeline. Our knowledge mining framework represents a significant contribution in that direction.

9. CONCLUSION

At the onset of this work, we declared the intention to define a direction for "second-generation" bioinformatics, and to show that knowledge-enabled analysis pipelines can discover important new knowledge in large-scale datasets. Our aim was to provide a new class of analysis methods, suited for the analysis of the growing quantities of biological data made available by advances in biotechnology. We proposed that these methods could empower biomedical researchers if presented via biologist-friendly interfaces that hide computing complexities. In the following sections, we review the results and contributions made in this thesis, and assess them against our original aims.

9.1 Review of results

9.1.1 Biological Knowledge Mining

In Chapter 2, we have described a novel conceptual biological knowledge mining framework for describing multi-stage bioinformatics pipelines, and introduced a notation which simply but effectively captures knowledge flow through the analysis process. In this framework, knowledge is not assumed to emerge from data alone, but as a result of combining data with other knowledge (such as descriptive metadata, or analysis results from analysis tasks). Since knowledge flow and task composition are key aspects of multi-stage analysis, our framework is a contribution towards the design and formalization of large-scale bioinformatics projects.

9.1.2 Knowledge flow and Knowledge-enabled tools

The term *knowledge flow* implies that knowledge is seamlessly transferred from one task to the next. This is in stark contrast with most of today's analysis tools, which require specific input data formats, and often can only process a single type of data, making the construction of analysis pipelines a challenging proposition. In Chapter 4, we have proposed semantic technologies as a suitable platform for representing knowledge along the analysis pipeline.

We have shown that powerful reasoning and structuring tasks can be easily performed with standard tools, and produce real, quantifiable benefits when applied to our influenza A dataset. Our results make a supporting case for proposing that analysis tools should be enhanced with semantic technologies, so that they can augment knowledge with analysis results and expert user input as it flows through the analysis pipeline.

Although knowledge-enabled analysis tools are not yet available, we demonstrated their power by adding relatively simple metadata integration to the AVANA tool, presented in Chapter 5, Section 5.3. AVANA is metadata-agnostic, in that it provides a generic interface for selecting sequence subsets based on user-provided metadata, rather than demand specific data fields. This powerful approach allows users to control at will the partitioning of the data during comparative analyses and meta-analyses, leading to rapid and flexible testing of hypotheses. In contrast, current analysis tools normally demand the construction of new datasets for different analysis tasks. The addition of metadata capabilities to AVANA has produced results of considerable biomedical importance, when applied to comparative analysis of influenza A proteins (Chapter 6), and to conservation meta-analyses of viral proteins (Chapter 7). Therefore, this work has successfully pioneered a new class of tools able to leverage on knowledge in complex analysis.

9.1.3 Rule-based Biological Knowledge Aggregation

Effective aggregation of data and descriptive metadata is arguably the hardest task in large-scale analysis projects, because of system and information heterogeneities. The approach proposed in Chapter 3 this thesis, and implemented by the ABK platform (Section 3.3) combined multiple strategies that address simultaneously all important heterogeneities:

- *System heterogeneity* is handled by a mediator architecture, whose wrappers isolate users from the technical complexities of delivering queries and gathering results. Although mediator architectures are in common use, our approach is unique in that it does not demand data structural mapping to be built into the wrappers, a requirement of most mediator implementations. This key feature offers two important advantages:

it simplifies the development of new wrappers, and it makes the selection of desirable source data a choice of the end user, rather than the wrapper developer.

- *Structural heterogeneity* is addressed by user-specified structural rules for the extraction of source data. This approach is novel, since it does not require end users to possess detailed knowledge of the source data structure. Rather, they are presented with source records, and specify extraction rules by example, using point-and-click gestures. This mechanism was possible thanks to (a) the versatile XML standard, which allows structural paths to be specified independently of the actual language used by the database, and (b) an innovative user interface component for simplifying the visualization of XML documents, which is a further contribution of this work.
- *Syntactic heterogeneity* is tackled by text filters, which use regular expressions or user-defined dictionaries. User-defined dictionaries enable user control over the extraction of values, to suit the content of the source data and the requirements of task in hand. As a result, this approach is easily customizable to handle dataset-specific values, which would not be feasible if rules were predefined.
- *Semantic heterogeneity* is addressed by allowing contributions from documents extracted from multiple sources, and from multiple rules within each document, while controlling the priority of these rules. Our approach finds values, and highlights value conflicts to the user, providing facilities for their resolution.

These approaches were put to the test on a real-life large-scale aggregation task involving tens of thousands of records. The results reported in Chapter 3, Section 3.4, have shown that such an aggregation task is extremely challenging, because of high information heterogeneity. However, the biologist-friendly ABK tool provided a high degree of automation, making this task manageable with limited manpower and short timescales. The curated datasets provided the starting point of important discovery tasks, leading to a better understanding of immunological and virological aspects of the influenza A virus, whose results were published on peer-reviewed international journals.

9.1.4 Bioinformatics for applied biomedical research

Bioinformatics should serve biomedical discovery, and therefore biomedical researchers must be empowered to control bioinformatics analyses autonomously. The usability of bioinformatics analysis was a common thread throughout this work, and a key perspective from which to evaluate the contributions made. This emphasis on usability is in line with current thinking by leading bioinformatics tool developers (Kumar *et al.* 2008), and backed by evidence that biomedical researchers favour "biologist-friendly" tools (Kumar and Dudley 2007). However, no biologist-friendly tools are currently available for user-driven knowledge aggregation and metadata management. Although system and information integration issues have been intensely studied in computer science, all solutions proposed to date for integrating biological systems have required in-depth technical knowledge of the source databases, programming knowledge, specialized software and hardware infrastructure, or a combination of these. Therefore, our knowledge aggregation method is the first step towards allowing life scientists to aggregate data from multiple, heterogeneous data sources without having to perform any form of sophisticated data modelling or mapping.

The tools and methods described in the present thesis have shown that highly complex tasks, such as knowledge aggregation, meta-analyses and text mining, can indeed be controlled by users who have domain knowledge but no programming expertise. Thus, the "biologist-friendly" approaches presented in Chapters 3, 5 and 8 (namely: point-and-click specification of structural rules; user-driven rule prioritization; simplified visualization of XML documents; simple value conflict management; user-specified value dictionaries; simplified metadata value selection) have contributed a "palette" of techniques for increasing accessibility of biomedical researchers to the growing volumes of biological data.

9.1.5 Information-theoretical algorithms

In Chapter 5, we presented two novel information theoretical sequence analysis methods. Information theory was selected because of its power to transform multiple variability aspects

into simple metrics, and because of its efficiency of in terms of execution time and memory resources. When these methods were built into the AVANA tool, alignments of thousands of sequences could be processed in real time, using standard-configuration desktop systems.

In one method (Chapter 5, Section 5.2), we defined a novel use of *mutual information* (MI) for measuring the association between a mutation and a specific set of sequences. In Chapter 6, we have shown that this measure is the most sensitive used to date for identifying adaptive mutations in human influenza. The catalogue of 70 characteristic sites produced by our analysis contained approximately twice as many sites as previous leading studies. In this study, descriptive metadata was used by the AVANA tool to form and compare sequence subsets rapidly and accurately. The characteristic sites catalogue has significant impact on our understanding of influenza A biology, revealing a new picture of human host adaptation of this virus, systemically more complex than previously thought, in which constellations of mutually-adapted internal proteins play a major role. Furthermore, the catalogue of adaptive mutations was used as additional metadata to extract *adaptive signatures* of influenza strains, a novel visualization of the level of adaptation of the virus. The analysis of adaptive signatures suggested insights into the evolution of human-infecting influenza viruses, and provided a new tool for the assessment of the host-jumping potential of avian strains. All these findings constitute significant contributions to the field of influenza virology, at a time when the pandemic potential of this virus is of great concern to the scientific community. The importance of these results is a testimony of the power of knowledge-enabled bioinformatics.

The second method (Chapter 5, Section 5.1) introduced *peptide entropy*, a novel measure of the diversity of short potentially immunogenic peptides, which accounts for combinatorial complexity deriving from neighbouring residue mutations. Peptide entropy thus provides an immunologically meaningful measure of variability. In Chapter 7, it was shown to be useful for identifying conserved potential epitopes for vaccine formulation. The combination of large-scale dataset construction using ABK with metadata-enabled peptide conservation meta-analysis, detailed in Chapter 5, provided a robust pipeline for immunological conservation studies, reusable for the analysis of a variety of viruses. The results of conservation analyses

by AM Khan and AT Heiny, summarized in Chapter 7, Section 7.3, have shown that both influenza A and dengue virus proteomes contain a high number of highly conserved peptides distributed across a several viral proteins, several of which are potential T-cell epitopes. These results have illustrated the importance and utility, as well as the generality, of the knowledge aggregation method and of large-scale information theory analysis.

9.1.6 Reusable active text mining

The text mining application discussed in Chapter 8 is different from the other applications reported in this thesis, and as such it demonstrated clearly that the principles proposed in this thesis are applicable to a broad range of applications, and not limited to the viral sequence analysis. The text mining application uses the ABK platform for the query and retrieval, and allows text analysis tools to augment metadata with new knowledge about text features, such as allergen identifiers. User annotations, entered through a simple graphical tool, are also integrated with metadata, and exploited by machine learning algorithms to enhance classification performance. The text mining plug-ins described in Chapter 8 are further examples of metadata-enabled tools, capable of integrating newly derived knowledge with existing knowledge, which use ABK as a platform for managing knowledge flow.

Although biological text mining is currently an active research area, it is predominantly the domain of computer scientists and linguists. Therefore, it is a key achievement of this study to make this technology available to biomedical researchers, through an interface that they can understand and control, and without embedding domain knowledge into the software code. Our results demonstrated that generic, mainstream machine learning software can produce substantial curation effort savings, when expert knowledge is channeled into the analysis task.

9.2 Future work

Work in the field of biological knowledge mining has just begun. This thesis has identified key objectives and directions, but the technologies and methods proposed must be considered

research prototypes. It is likely that industrial-strength tools capable of supporting multi-stage knowledge-enabled bioinformatics analyses will only emerge over the next 5 to 10 years, and early implementations may not necessarily be biologist-friendly. Meanwhile, however, the tools and methods developed in the course of this thesis will continue to be improved and applied to new problems, through the many collaborations initiated during their development. In this final section, we review the opportunities for the field and for our work in particular. Knowledge flow through the analysis pipeline is arguably the most difficult issue in biological knowledge mining at present. Although semantic technologies are strong candidates for supporting knowledge flow, there are still a number of obstacles to their adoptions, which were discussed in Chapter 4. While the knowledge representation infrastructure (XML, RDF and OWL) is ready for broad adoption, much work will be needed before an agreement is reached on the set of ontologies to be used for labeling and interpreting biological knowledge. However, one advantage of RDF is that tool support can be added support without committing to specific ontologies, and further research may find generic user-friendly ways of selecting, translating or mapping RDF-encoded knowledge. The composition and orchestration of analysis tasks into complex pipelines will require analysis tools to be self-descriptive, so that their knowledge inputs and outputs can be automatically matched. In this area, progress has been made with Semantic Web Services, which are gaining some industry acceptance. However, usability remains a significant stumbling block: current solutions are too technically oriented, and likely to alienate life scientists. Innovative approaches will be needed if bioinformatics is to be controlled by those who need to benefit from it. Michalski's own vision (Kaufman and Michalski 2005) is that of intelligent agents capable of applying reasoning on the outcome of analysis tasks, making decisions on subsequent tasks, and thus creating dynamic and optimized pipelines. Although this vision is unlikely to be realized in a near future, it encourages us to address important gaps, such as the lack of languages for expressing user goals, intentions and expectations in machine-understandable forms.

There are several opportunities for extension and improvement of the methods and tools

presented in this thesis. These are some of the research areas that have been identified for the ABK system: (a) improved dictionary management, based on point-and-click mechanisms during source record inspection; (b) automated prioritization of structural rules, based on estimated accuracy after manual curation; (c) alternative mechanisms for determining the winning value from multiple rules, such as weighted consensus; (d) establishment of multiple levels of conflict, determined by the priority of conflicting rules, so that verification can be prioritized; (e) improved usability of the spreadsheet interface to facilitate curation; and (f) extraction of values from interlinked external documents. Each of these areas presents significant challenges, but has the potential to further improve the knowledge aggregation process. Similarly, important expansion opportunities have been identified for the AVANA software, including: (a) multi-subset comparative analysis; (b) detection of epitope gain/loss due to characteristic mutations; (c) combination of mutual information and genetic distance to estimate the significance of mutations; (d) comparative diversity analysis of DNA sequences and their protein products; and (e) use of adaptive signatures as a tool for genotyping sequences. While researching these improvements, we will be applying the current implementations to further work on pathogens: current project include serotype-specific conservation analysis of dengue, and host adaptation analysis of rabies virus. Finally, a project has been initiated to implement the active learning method described in Chapter 8 into standalone desktop-based tool. This tool, named Reusable Active Text Mining Annotation Tool (RATMAT), will support classification using multiple machine learning algorithms, and automated optimization of the classification process.

Nobel Laureate George Wald said: “Science goes from question to question; big questions, and little, tentative answers” (Wald 1967). In this thesis we have asked a “big” question: how must bioinformatics change to empower biomedical researchers to effectively use the growing volume of biological data for high-value discovery? Put simply, our “tentative” answer was: integrate knowledge into the analysis process. Far from removing all possible obstacles from the path towards “second-generation” bioinformatics, our contributions have focused on specific aspects that we deemed particularly important:

knowledge aggregation, knowledge enabled analysis tools, and biologist-friendly mechanisms. However, the results we obtained from applying our approaches to real-life biomedical research problems are not "tentative" at all: they present clear and unmistakable evidence that knowledge-enabled bioinformatics can produce important results that will advance biomedical research. We are therefore confident that the approaches we developed in this thesis are important first steps in a new direction, with the potential to grow into tools that will shape the biomedicine of tomorrow.

Bibliography

Achard F, Vaysseix G, Barillot E (2001) XML, bioinformatics and data integration. *Bioinformatics* 17(2), 115-125.

Ackoff RL (1989) From Data to Wisdom. *Journal of Applied Systems Analysis* 16, 3-9.

Agrawal R, Imielinski T, Swami AN (1993) Mining Association Rules between Sets of Items in Large Databases. *Proc. of the ACM Intl. Conf. on Management of Data (SIGMOD 93)*, 207-216.

Akarsu H, Burmeister WP, Petosa C, Petit I, Müller CW, Ruigrok RW, Baudin F (2003) Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2). *EMBO J.* 22, 4646-4655.

Alfarano C, Andrade CE, Anthony K, *et al*. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.* 33, D418-424.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol.* 215(3), 403-410.

Andrade J, Berglund L, Uhlén M, Odeberg J (2006) Using Grid technology for computationally intensive applied bioinformatics analyses. *In Silico Bio* 6(6), 495-504.

André M, Borgquist L, Foldevi M, Mölstad S (2002) Asking for 'rules of thumb': a way to discover tacit knowledge in general practice. *Fam Pract* 19(6), 617-622.

Attwood TK (2000) Genomics. The Babel of bioinformatics. *Science* 290(5491), 471-473.

Austin CP (2004) The impact of the completed human genome sequence on the development of novel therapeutics for human disease. *Annu Rev Med* 55, 1-13.

Bard JB, Rhee SY (2004) Ontologies in biology: design, applications and future challenges. *Nat Rev Genet.* 5(3), 213-22.

Baudin F, Petit I, Weissenhorn W, Ruigrok RWH (2001) In vitro dissection of the membrane binding and RNP binding activities of influenza virus M1 protein. *Virology* 281, 102–108.

Beckett D (ed.) (2004) RDF/XML Syntax Specification (Revised), W3C 2004. http://www.w3.org/TR/rdf-syntax-grammar/ (accessed 22 May 2007)

Bennett ST, Barnes C, Cox A, Davies L, Brown C (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics* 6(4), 373-382.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. *Nucleic Acids Res.* 36, D25-30.

Benton D (1996) Bioinformatics- principles and potential of a new multidisciplinary tool. *Trends Biotechnol* 14(8), 261-272.

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. *Sci. Am.* 279, 34-43.

Bernstein BE, Kellis M (2005) Large-scale discovery and validation of functional elements in the human genome. *Genome Biol.* 6(3), 312.

Bidartondo MI (2008) Preserving accuracy in GenBank. *Science* 319(5870), 1616.

Boisot M, Canals A (2004) Data, information and knowledge: have we got it right? Working Paper Series WP04-002, Internet Interdisciplinary Institute, Barcelona. http://www.uoc.edu/in3/dt/20388/index.html (accessed 2 June 2008)

Brazma A, Jonassen I, Eidhammer I, Gilbert D (1998) Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology* 5(2), 279-305.

Brazma A, Krestyaninova M, Sarkans U (2006) Standards for systems biology. *Nat Rev Genet* 7(8), 593-605.

Bray T, Paoli J, Sperberg-McQueen CM, Maler E, Yergeau F (2006) Extensible Markup Language (XML) 1.0 (Fourth Edition), W3C. http://www.w3.org/TR/xml/ (accessed 17 May 2008).

Bretherton FP, Singley PT (1994) Metadata: A User's View. Proc. 7th Intl. Conf. on Scientific and Statistical Database Management, Charlottesville, USA, 166-174

Brusic V, August JT (2004) The changing field of vaccine development in the genomics era. *Pharmacogenomics* 5, 597–600.

Brusic V, Koh JLY (2004) Genetic databases. In: Ruvinsky A, Graves J, Mammalian Genomics, 411-427. Wallingford, CAB International.

Brusic V, Millot M, Petrovsky N, Gendel SM, Gigonzac O, Stelman SJ (2003) Allergen databases. *Allergy* 58(11), 1093-1100.

Brusic V, Petrovsky N, Gendel SM, Millot M, Gigonzac O, Stelman SJ (2003) Computational tools for the study of allergens. *Allergy* 58(11), 1083-1092.

Brusic V, Zeleznikow J (1999) Knowledge Discovery and Data Mining in Biological Databases. *Knowledge Engineering Review* 14(3), 257-277.

Brusic V, Wilkins JS, Stanyon CA, Zeleznikow J (1998) Data learning: understanding biological data. In Merrill G and Pathak DK (Eds.). Knowledge Sharing Across Biological and Medical Knowledge Based Systems. AAAI Technical Report WS–98–04, AAAI Press, 12–19.

Buckler-White AJ, Naeve CW, Murphy BR (1986) Characterization of a gene coding for M proteins which is involved in host range restriction of an avian influenza A virus in monkeys. *J Virol.* 57(2), 697-700.

Carvalho PC, Glória RV, de Miranda AB, Degrave WM (2005) Squid - a simple bioinformatics grid. *BMC Bioinformatics* 6, 197.

Castano S, De Antonellis V (1999) A schema analysis and reconciliation tool environment for heterogeneous databases. Proc. of Intl. Database Engineering and Applications Symposium (IDEAS '99), 53-62.

Chandrasekaran A, Srinivasan A, Raman R, Viswanathan K, Raguram S, Tumpey TM, Sasisekharan V, Sasisekharan R (2008) Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin. *Nat Biotechnol.* 26, 107-113.

Chen GW, Chang SC, Mok CK *et al.* (2006) Genomic signatures of human versus avian influenza A viruses. *Emerg Infect Dis*. 12(9), 1353-1360.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31(13), 3497-3500.

Chun WC, Detlor B, Turnbull B (2000). Web Work: Information Seeking and Knowledge Work on the World Wide Web. Kluwer Academic Publishers, Dordrecht. ISBN: 0792364600

Chung SY, Wong L (1999) Kleisli: a new tool for data integration in biology. *Trends Biotechnol* 17, 351-355

Clark J (ed.) (1999). XSL Transformations (XSLT) Version 1.0. W3C Recommendation. W3C. http://www.w3.org/TR/xslt (accessed 17 May 2008)

Clark J, DeRose S (eds.) (1999) XML Path Language (XPath). W3C 1999. http://www.w3.org/TR/XPath (accessed 22 May 2007)

Cohen AM, Hersh WA (2005) A survey of current work in biomedical text mining. *Briefings in Bioinformatics* 6(1), 57-71.

Cohn D, Atlas L, Ladner R (1994). Improving generalization with active learning. *Machine Learning* 15 (2), 201-221.

Cowell JK, Hawthorn L (2007) The application of microarray technology to the analysis of the cancer genome. *Curr Mol Med* 7(1), 103-120.

Crick F (1966) Of molecules and men. Washington University Press, Seattle.

Davis R, Shrobe HE, Szolovits P (1993) What Is a Knowledge Representation? *AI Magazine* 14(1), 17-33.

Dawy Z, Goebel B, Hagenauer J, Andreoli C, Meitinger T, Mueller JC (2006) Gene mapping and marker clustering using Shannon's mutual information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 3(1), 47-56.

Debouck C, Metcalf B (2000) The impact of genomics on drug discovery. *Annu Rev Pharmacol Toxicol* 2000, 40193-40207.

de Bruijn B, Martin J (2002) Getting to the (c)ore of knowledge: mining biomedical literature. *Int J Med Inform.* 67(1-3), 7-18.

Delamothe T, Smith R (2004) Open access publishing takes off. *BMJ* 328(7430), 1-3.

Deshpande M, Karypis G (2002) Using conjunction of attribute values for classification. *ACM Intl. Conf. on Information and Knowledge Management (CIKM 2002)*, 356-364.

Do HH, Rahm E (2002) COMA - A System for Flexible Combination of Schema Matching Approaches. *Proc. Intl. Conf. Very Large Data Bases (VLDB 2002),* 610-621, Hong Kong, China.

Doan A, Domingos P, Halevy AY (2001) Reconciling schemas of disparate data sources: a machine-learning approach. *SIGMOD Rec* 30(2), 509-520.

Doan A, Halevy AY (2005) Semantic-integration research in the database community. *AI Mag* 26(1), 83-94.

Doyle H, Gass A, Lappin D (2003) A changing landscape. *PLoS Biol.* 1(3). E89.

Dupre J (1986) Review [of Alexander Rosenberg's "The Structure of Biological Science"]. *Philosophy of Science* 53(3), 461-463.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* 32(5), 1792-1797.

Edgar RC, Batzoglou S. (2006) Multiple sequence alignment. Curr Opin Struct Biol16(3), 368-373

Erhardt RA, Schneider R, Blaschke C (2006) Status of text-mining techniques applied to biomedical text. *Drug Discov Today* 11(7-8), 315-325.

Erl T (2005) Service-oriented Architecture: Concepts, Technology, and Design. Prentice Hall PTR. ISBN: 0131858580

Esser MT, Marchese RD, Kierstead LS, Tussey LG, Wang F, Chirmule N, Washabaugh MW (2003) Memory T cells and vaccines. *Vaccine* 21(5-6), 419-430.

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. *AI Magazine*, Fall 1996, 37-54.

Fechter P, Mingay L, Sharps J *et al.* (2003): Two aromatic residues in the PB2 subunit of influenza A RNA polymerase are crucial for cap binding. *J Biol Chem.* 278, 20381-20388.

Finkelstein DB, Mukatira S, Mehta PK, Obenauer JC, Su X, Webster RG, Naeve CW (2007) Persistent host markers in pandemic and H5N1 influenza viruses. *J Virol.* 81(19), 10292-10299

Forton JT, Kwiatkowski DP (2006) Searching for the regulators of human gene expression. *Bioessays* 28(10), 968-972

Fouchier RA, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, Rimmelzwaan GF, Olsen B, Osterhaus AD (2005) Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J Virol.* 79, 2814-2822.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32, W273-9

Fredman D, Siegfried M, Yuan YP *et al*. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources, *Nucleic Acids Res.* 30, 387-391.

Galperin MY (2008) The Molecular Biology Database Collection: 2008 update. *Nucleic Acids Research* 36(Database issue), D2-4. Database summaries at http://nar.oxfordjournals.org/cgi/content/full/gkm1037/DC1 (accessed 1 June 2008)

Garcia Castro A, Thoraval S, Garcia LJ, Ragan MA (2005) Workflows in bioinformatics: meta-analysis and prototype implementation of a workflow generator. *BMC Bioinformatics* 6, 87.

Garcia-Remesal M, Maojo V, Billhardt H, Crespo J, Alonso-Calvo R, Perez-Rey D, Martin F, Sousa A (2004) ARMEDA II: supporting genomic medicine through the integration of medical and genetic databases. *Proc IEEE Sym. on Bioinformatics and Bioengineering* (BIBE 2004), 227- 234.

Garcia-Solaco M, Saltor F, Castellanos M (1996) Semantic Heterogeneity in Multidatabase Systems. In Object-Oriented Multidatabase Systems. Edited by Bukhres O, Elmagarmid A . Englewood Cliffs: Prentice Hall, 129-202.

Ghedin E, Sengamalay NA, Shumway M *et al.* (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* 437(7062), 1162-1166.

Gatenby RA, Frieden BR (2007) Information theory in living systems, methods, applications, and challenges. *Bull Math Biol.* 69(2), 635-657.

Gonzalez S, Ortin J (1999) Distinct regions of influenza virus PB1 polymerase subunit recognize vRNA and cRNA templates. *EMBO J.* 18, 3767-3775.

Gooden P, Owen M, Simon S (2002) Scientific Publishing: Knowledge is Power. Morgan Stanley, London. http://www.econ.ucsb.edu/~tedb/Journals/morganstanley.pdf (accessed 12 June 2008)

Greenspan D, Palese P, Krystal M (1988) Two nuclear location signals in the influenza virus NS1 nonstructural protein. *J Virol.* 62(8), 3020-3026.

Guan Y, Shortridge KF, Krauss S, Webster RG (1999) Molecular characterization of H9N2 influenza viruses: were they the donors of the "internal" genes of H5N1 viruses in Hong Kong? *Proc Natl Acad Sci U S A.* 96(16), 9363-9367.

Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol.* 210, 1518-1525.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41, 95-98.

Härder T, Sauter G, Thomas J (1999) The intrinsic problems of structural heterogeneity and an approach to their solution. *The VLDB Journal* 8, 25-43.

Hatta M, Gao P, Halfmann P, Kawaoka Y (2001) Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* 293(5536), 1840-1842.

Hawke S (editor) (2005) Rule Interchange Format Working Group Charter. W3C, http://www.w3.org/2005/rules/wg/charter.html (accessed 17 May 2008)

Hearst M (1999) Untangling Text Data Mining. Proc. of the 37th Annual Meeting of the *Association for Computational Linguistics*, Uni. of Maryland, USA.

Heimbigner D, McLeod D (1985) A Federated Architecture for Information Management. *ACM Transactions on Information Systems* 3, 253-278

Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, Brusic V, Tan TW, August JT (2007) Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. *PLoS ONE* 2(11), e1190.

Henschel R, Muller M (2007) I/O Induced Scalability Limits of Bioinformatics Applications. *Proc. 7th IEEE Intl. Conf. Bioinformatics and Bioengineering (BIBE 2007)* 609-613.

Hernández MA, Miller RJ, Haas LM (2001) Clio: a semi-automatic tool for schema mapping. *SIGMOD Rec* 30 (2), 607.

Hernandez T, Kambhampati S (2004) Integration of biological sources: current systems and challenges ahead. *SIGMOD Rec*, 33(3), 51-60

Hoffman D, Lowenstein H, Marsh DG, Platts-Mills TAE, Thomas W (1994) Allergen Nomenclature, *Bull. of the World Health Organization* 72(5), 796-806.

Hoffmann E, Stech J, Leneva I, Krauss S, Scholtissek C, Chin PS, Peiris M, Shortridge KF, Webster RG (2000) Characterization of the influenza A virus gene pool in avian species in southern China: was H6N1 a derivative or a precursor of H5N1? *J Virol.* 74, 6309-6315.

Hofmann O, Schomburg D (2005) Concept-based annotation of enzyme classes, *Bioinformatics* 21(9), 2059-2066.

Hoffman SL, Subramanian GM, Collins FH, Venter JC (2002) Plasmodium, human and Anopheles genomics and malaria. Nature 415(6872), 702-709.

Honda A, Mizumoto K, Ishihama A (1999) Two separate sequences of PB2 subunit constitute the RNA cap-binding site of influenza virus RNA polymerase. *Genes Cells* 4(8): 475-485.

Horrocks I, Patel-Schneider PF, Boley H *et al*. (2004) SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C. http://www.w3.org/Submission/SWRL/ (accessed 17 May 2008)

Hui EK, Barman S, Yang TY, Nayak DP (2003) Basic residues of the helix six domain of influenza virus M1 involved in nuclear translocation of M1 can be replaced by PTAP and YPDL late assembly domain motifs. *J Virol.* 77, 7078-7092.

Hulo N, Bairoch A, Bulliard V *et al.* (2008) The 20 years of PROSITE. *Nucleic Acids Res* 36(Database issue), D245-249.

Hunter L, Cohen KB (2006) Biomedical language processing: What's beyond PubMed? *Mol Cell* 21. 589–594.

Iwatsuki-Horimoto K, Horimoto T, Fujii Y, Kawaoka Y (2004) Generation of influenza A virus NS2 (NEP) mutants with an altered nuclear export signal sequence. *J Virol.* 78, 10149-10155.

Jain E (2007) UniProt-RDF Project Overview. Swiss Institute of Bioinformatics. http://dev.isb-sib.ch/projects/uniprot-rdf/ (accessed 17 May 2008).

Janée G, Frew J, Hill LL (2004) Issues in Georeferenced Digital Libraries. *D-Lib Magazine* 10(5). http://www.dlib.org/dlib/may04/janee/05janee.html (accessed 15 June 2008)

Joachims T (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *European Conf. on Machine Learning (ECML-98)*, *Lecture Notes in Computer Science* 1398, 137-142.

Jones IM, Reay PA, Philpott KL (1986) Nuclear location of all three influenza polymerase proteins and a nuclear signal in polymerase PB2. *EMBO J*. 5(9), 2371-2376.

Jung TE, Brownlee GG (2006) A new promoter-binding site in the PB1 subunit of the influenza A virus polymerase. *J Gen Virol.* 87, 679-688.

Kanehisa M, Bork P (2003) Bioinformatics in the post-sequence era. *Nature Genet.* 33 Suppl, 305-310.

Karasavvas KA, Baldock R, Burger A (2004) Bioinformatics integration and agent technology. *J Biomed Inform*. 37(3), 205-219.

Karp PD, Paley S, Zhu J (2001) Database verification studies of SWISS-PROT and GenBank. *Bioinformatics* 17, 526-532.

Kaufman K, Michalski RS (2003) The Development of the Inductive Database System VINLEN: A Review of Current Research. *Proc. Intl. Intelligent Information Processing and Web Mining Conf. (IIPWM'03)*, Zakopane, Poland.

Kaufman K, Michalski RS (2005) From Data Mining to Knowledge Mining. In: Rao CR, Solka JL, Wegman EJ (Eds.). Handbook in Statistics, Vol. 24: Data Mining and Data Visualization, 47-75, Elsevier.

Kendal AP, Noble GR, Skehel JJ, Dowdle WR (1978) Antigenic similarity of influenza A (H1N1) viruses from epidemics in 1977-1978 to "Scandinavian" strains isolated in epidemics of 1950-1951. *Virology* 1978, 89(2), 632-636.

Khan AM, Miotto O, Heiny AT, Salmon J, Srinivasan KN, Nascimento EJM, Marques ET, Brusic V, Tan TW, August JT (2006) A systematic bioinformatics approach for selection of epitope-based vaccine targets. *Cell Immunol.* 244(2), 141-147.

Khan AM, Miotto O, Nascimento EJM, Srinivasan KN, Heiny AT, Zhang GL, Salmon J, Marques ET, Tan TW, Brusic V, August JT (2008) Conservation and Variability of Dengue Virus Proteins: Implications for Vaccine Design. *PLoS Neglected Tropical Diseases*. 2(8), e272.

Khandheria P, Garner HR (2007) Developing a modern web interface for database-driven bioinformatics tools. *IEEE Eng Med Biol Mag* 26(2), 96-98.

Kitano H (2002) Systems biology: a brief overview. *Science* 295(5560),1662-1664.

Kmietowicz Z (2001) Deal allows developing countries free access to journals. *BMJ* 323, 65.

Koh JLY, Krishnan SPT, Seah SH, Tan PT, Khan AM, Lee ML, Brusic V (2004) BioWare: A framework for bioinformatics data retrieval, annotation and publishing. ACM SIGIR Workshop on Search and Discovery in Bioinformatics (SIGIRBIO) (2004) Sheffield, UK.

Koh JLY, Lee ML, Brusic V (2005). A classification of biological data artifacts. In Workshop on Database Issues in Biological Databases (DBiBD). Edinburgh,UK, 53-57.

Korber BT, Kunstman KJ, Patterson BK *et al*. (1994) Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J Virol.* 1994, 68(11), 7467-7481.

Kumar S, Dudley J (2007) Bioinformatics software for biologists in the genomics era. *Bioinformatics* 23(14), 1713-1717.

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(4), 299-306.

Lamb RA, Zebedee SL, Richardson CD (1985) Influenza virus M2 protein is an integral membrane protein expressed on the infected-cell surface. *Cell* 40(3), 627-633.

Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DW, Tenenbaum JD, Karp PD. (2006) BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics* 7, 170.

Lee YS, Oh J, Kim YU, Kim N, Yang S, Hwang UW (2008) Mitome: dynamic and interactive database for comparative mitochondrial genomics in metazoan animals. *Nucleic Acids Res* 36(Database issue), D938-942.

Li X, Morie P, Roth D (2005) Semantic integration in text: from ambiguous names to identifiable entities. *AI Mag* 26(1), 45-58.

Li Y, Yamakita Y, Krug RM (1998) Regulation of a nuclear export signal by an adjacent inhibitory sequence: the effector domain of the influenza virus NS1 protein. *Proc Natl Acad Sci U S A.* 95, 4864-4869.

Long JM (1986) The POSCH data processing experience: the problem of metadata. *J Med Syst* 10(2), 173-183.

Lundy RT (1984) Metadata Management. *IEEE Database Eng Bull* 7(1), 43-48.

Luscombe NM, Greenbaum D, Gerstein M. (2001). What is bioinformatics? An introduction and overview. *IMIA Yearbook of Medical Informatics*, 83-100, Stuttgart, Schattauer.

Maines TR , Chen LM, Matsuoka Y *et al.* (2006) Lack of transmission of H5N1 avian-human reassortant influenza viruses in a ferret model. *Proc Natl Acad Sci U S A.* 103, 12121-12126.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24(3), 133-141.

Marinelli RJ, Montgomery K, Liu CL, *et al.* (2008) The Stanford Tissue Microarray Database. *Nucleic Acids Res* 36(Database issue), D871-877.

Markovitz BP (2000) Biomedicine's electronic publishing paradigm shift: copyright policy and PubMed Central, *J. American Medical Informatics Assoc.* 7, 222-229.

Markowitz VM, Chen IM, Kosky AS, Szeto E (1997) Facilities for exploring molecular biology databases on the Web: a comparative study. *Pac Symp Biocomput 1997*, 256-267.

Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21(22), 4116-4124.

Martín MA, Rey JM (2000) On the role of Shannon's entropy as a measure of heterogeneity. *Geoderma* 98, 1-3.

McBride, B (2002) Jena: a semantic Web toolkit, *IEEE Internet Computing* 06(6), 55-59.

McCarthy J (1987) Generality in artificial intelligence. *Communications of the ACM* 30 (12), 1030-1035.

McGuinness DL, van Harmelen F (2004) OWL Web Ontology Language Overview. W3C. http://www.w3.org/TR/owl-features/ (accessed 17 May 2008).

McIlraith SA, Son TC, Zeng H (2001) Semantic Web services. *IEEE Intelligent Systems* 16(2), 46- 53

McKusick VA (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* 80(4), 588-604.

Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R (2008) Microbiology in the post-genomic era. *Nat Rev Microbiol* 6, 419-430.

Michael M, Moreira JE, Shiloach D, Wisniewski RW (2007) Scale-up x Scale-out: A Case Study using Nutch/Lucene. *Proc. IEEE Intl. Parallel and Distributed Processing Symposium (IPDPS 2007)*, 1-8, Long Beach, USA

Michalski RS (2003) Knowledge Mining: A Proposed New Direction. *Sanken Symposium on Data Mining and Semantic Web*, Osaka, Japan, March 10-11, 2003.

Miotto O, Tan TW, Brusic V (2005a) Extraction by Example: Induction of Structural Rules for the Analysis of Molecular Sequence Data from Heterogeneous Sources. *Lecture Notes in Computer Science* 3578, 398-405.

Miotto O, Tan TW, Brusic V (2005b) Supporting the curation of biological databases with reusable text mining. *Genome Informatics* 16(2), 32-44.

Miotto O, Heiny AT, Tan TW, August JT, Brusic V (2008a) Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis. *BMC Bioinformatics* 9, Suppl 1, S18.

Miotto O, Tan TW, Brusic V (2008b) Rule-based Knowledge Aggregation for Large-Scale Protein Sequence Analysis of Influenza A Viruses. *BMC Bioinformatics* 9, Suppl 1, S7.

Miotto O, Heiny AT, Tan TW, August JT, Brusic V (2009a) Complete-proteome mapping of human influenza A adaptive mutations: implications for human transmissibility of avian strains. *PLOS Pathogens* (manuscript in preparation)

Miotto O, Tan TW, Brusic V (2009b) AVANA: a tool for analyzing antigenic variability in large sets of protein sequences. *Bioinformatics* (manuscript in preparation)

Morgan XC, Ni S, Miranker DP, Iyer VR (2007) Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. *BMC Bioinformatics* 8, 445.

Morris RW, Bean CA, Farber GK, *et al.* (2005) Digital biology: an emerging and promising discipline. *Trends Biotechnol*, 23(3), 113-117.

Motta S, Brusic V (2004). Mathematical Modelling of the Immune System. In Ciobanu G, Rozenberg G (eds.) Modelling in Molecular Biology, *Natural Computing Series,* Springer, 193-218

Mukaigawa J, Nayak DP (1991) Two signals mediate nuclear localization of influenza virus (A/WSN/33) polymerase basic protein 2. *J Virol.* 65(1), 245-253.

Naffakh N, Massin P, Escriou N, Crescenzo-Chaigne B, van der Werf S (2000) Genetic analysis of the compatibility between polymerase proteins from human and avian strains of influenza A viruses. *J Gen Virol.* 81, 1283-1291.

National Research Council (2007) The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. National Academies Press, Washington DC, USA.

Neerincx PBT, Leunissen JAM (2005) Evolution of web services in bioinformatics *Brief Bioinform* 6(2), 178-188.

Neirynck S, Deroo T, Saelens X, Vanlandschoot P, Jou WM, Fiers W (1999) A universal influenza A vaccine based on the extracellular domain of the M2 protein. *Nat Med.* 5(10), 1157-1163.

Neumann E (2005) A life science Semantic Web: are we there yet? *Sci STKE* 2005(283), pe22.

Neumann G, Kawaoka Y (2006) Host range restriction and pathogenicity in the context of influenza pandemic. *Emerg Infect Dis*, 12(6), 881-886.

Newsome M, Pancake C, Hanus J (1997) HyperSQL: web-based query interfaces for biological databases. *Proc. 30th Hawaii Intl. Conf. on System Sciences* 97, 329-339

Nieto A, de la Luna S, Barcena J, Portela A, Ortin J (1994): Complex structure of the nuclear translocation signal of influenza virus polymerase PA subunit. *J Gen Virol.* 75 (Pt 1),29-36.

NIH (2008) National Institutes of Health Public Access Policy. Bethesda, USA. http://publicaccess.nih.gov/ (accessed 13 June 2008)

Nilges M, Linge JP (2002) A Definition of Bioinformatics. Institut Pasteur, Paris. http://www.pasteur.fr/recherche/unites/Binfs/definition/ (accessed 1 June 2008).

Nobrega MA, Pennacchio LA (2004) Comparative genomic analysis as a tool for biological discovery. *J Physiol.* 554, 31-39.

Nonaka I, Takeuchi H (1995) The Knowledge-creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, New York.

Noor MA, Zimmerman KJ, Teeter KC (2006) Data sharing: how much doesn't get submitted to GenBank? *PLoS Biol,* 4(7), e228.

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 302(1), 205-217.

Novitsky V, Smith UR, Gilbert P *et al.* (2002) Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? *J Virol* 76, 5435-5451.

Obenauer JC, Denson J, Mehta PK *et al.* (2006) Large-scale sequence analysis of avian influenza isolates. *Science* 311, 1576-1580.

Ogbunugafor CB (2004) On reductionism in biology: pillars, leaps, and the naïve behavioral scientist. *Yale J Biol Med* 77(3-4), 101-109.

Ohtsu Y, Honda Y, Sakata Y, Kato H, Toyoda T (2002). Fine mapping of the subunit binding sites of influenza virus RNA polymerase. *Microbiol Immunol.* 46, 167-175.

Oinn T, Addis M, Ferris J *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20(17), 3045-3054.

Ouksel AM, Sheth AP (1999) Semantic interoperability in global information systems. *SIGMOD Rec* 28(1), 5-12.

Ovsyannikova IG, Jacobson RM, Poland GA (2004) Variation in vaccine response in normal populations. *Pharmacogenomics* 5, 417-427.

Ozawa M, Fujii K, Muramoto Y, Yamada S, Yamayoshi S, Takada A, Goto H, Horimoto T, Kawaoka Y (2007). Contributions of two nuclear localization signals of influenza A virus nucleoprotein to viral replication. *J Virol.* 81, 30-41.

Paninski L (2003) Estimation of entropy and mutual information. *Neural Computation* 15: 1191–1253.

Pappas C, Aguilar PV, Basler CF, Solórzano A, Zeng H, Perrone LA, Palese P, García-Sastre A, Katz JM, Tumpey TM (2008) Single gene reassortants identify a critical role for PB1, HA, and NA in the high virulence of the 1918 pandemic influenza virus. *Proc Natl Acad Sci U S A.* 2008 Feb 26;105:3064-3069.

Parker DS, Gorlick MM, Lee CJ. (2003) Evolving from bioinformatics in-the-small to bioinformatics in-the-large. *OMICS* 7(1), 37-48

Patrick K (2007) 454 life sciences: illuminating the future of genome sequencing and personalized medicine. *Yale J Biol Med*. 80(4),191-194.

Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 17(8), 700-712.

Philippi S, Köhler J (2006) Addressing the problems with life-science databases for traditional uses and systems biology. *Nat Rev Genet.* 7(6), 482-488.

Platt JC (1999) Using analytic QP and sparseness to speed training of Support Vector Machines. In: Kearns MS, Solla SA, Cohn DA (eds.) *Advances in Neural Information Processing Systems*, 11, MIT Press, Cambridge, USA.

Poole E, Elton D, Medcalf L, Digard P (2004) Functional domains of the influenza A virus PB2 protein: identification of NP- and PB1-binding sites. *Virology* 321(1), 120-133.

Potter CW (2001) A history of influenza. *J Appl Microbiol.* 91, 572-579.

Porter MF (1980) An algorithm for suffix stripping. *Program* 14(3), 130-137.

Qian XY, Chien CY, Lu Y, Montelione GT, Krug RM (1995) An amino-terminal polypeptide fragment of the influenza virus NS1 protein possesses specific RNA-binding activity and largely helical backbone structure. *RNA* 1(9), 948-956.

Quinlan JR (1992) Learning with Continuous Classes. *Proc. 5th Australian Joint Conf. on Artificial Intelligence*, Hobart, Australia, 343-348.

Rammensee HG (1995) Chemistry of peptides associated with MHC class I and class II molecules. *Curr Opin Immunol* 7, 85–96.

Rebholz-Schuhmann D, Kirsch H, Couto F (2005) Facts from text— is text mining ready to deliver? *PLoS Biology* 3(2), e65.

Regev Y, Finkelstein-Landau M, Feldman R (2003) Rule-based extraction of experimental evidence in the biomedical domain: the KDD Cup 2002. *ACM SIGKDD Explorations Newsletter* 4(2), 90–92.

Romano P, Marra D, Milanesi L (2005) Web services and workflow management for biological resources. *BMC Bioinformatics* 6 Suppl 4, S24.

Roos DS (2001) Bioinformatics- trying to swim in a sea of data. *Science* 291(5507), 1260-1261.

Ruben RJ (2003) The promotion of academic pediatric otolaryngology by journal peer review. *Int J Pediatr Otorhinolaryngol* 67 Suppl 1, S165-169.

Sanz-Ezquerro JJ, Zürcher T, de la Luna S, Ortín J, Nieto A (1996) The amino-terminal one-third of the influenza virus PA protein is responsible for the induction of proteolysis. *J Virol.* 70,1905-1911.

Scholtissek C, Rohde W, Von Hoyningen V, Rott R (1978) On the origin of the human influenza virus subtypes H2N2 and H3N2. *Virology*, 87(1), 13-20.

Schönbach C, Kowalski-Saunders P, Brusic V (2000) Data warehousing in molecular biology. *Brief Bioinform* 1(2), 190-198.

Sette A, Livingston B, McKinney D, Appella E, Fikes J, Sidney J, Newman M, Chesnut R (2001) The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation. *Biologicals* 29, 271-276.

Sette A, Sidney J (1999) Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 50, 201-212.

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27: 379-423, 623-656.

Sheth AP (1999) Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. In: Goodchild MF, Egenhofer M, Fegeas R *et al.* (eds). Interoperating Geographic Information Systems. Amsterdam, Kluwer

Sheth AP, Larson JA (1990) Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys* 22(3), 183-236.

Shirts M, Pande VS (2000) Screen Savers of the World Unite! *Science* 290(5498), 1903-1904.

Slonim N , Atwal GS , Tkacik G , Bialek W (2005) Estimating mutual information and multi-information in large networks. http://arxiv.org/abs/cs.IT/0502017 (accessed 12 June 2008)

Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 18, S231-40.

Stevens R, Bodenreider O, Lussier YA (2006) Semantic webs for life sciences. *Pac Symp Biocomput. 2006*, 112-115.

Stevens R, Zhao J, Goble C (2007) Using provenance to manage knowledge of in silico experiments. *Brief Bioinform.* 8(3), 183-194.

Stoeckert CJ Jr, Fischer S, Kissinger JC, Heiges M, Aurrecoechea C, Gajria B, Roos DS (2006) PlasmoDB v5: new looks, new genomes. *Trends Parasitol* 22(12), 543-546.

Stormo GD (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16(1), 16-23.

Strömbäck L, Hall D, Lambrix P (2007) A review of standards for data exchange within systems biology. *Proteomics* 7(6), 857-867.

Subbarao EK, London W, Murphy BR (1993) A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *J Virol.* 67(4), 1761-1764.

Swedlow JR, Lewis SE, Goldberg IG (2006) Modelling data across labs, genomes, space and time. *Nat. Cell. Biol.* 8(11), 1190-1194.

Thompson WW, Shay DK, Weintraub E, Brammer L, Cox N, Anderson LJ, Fukuda K (2003) Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA* 289:179-186.

Tompkins SM, Zhao ZS, Lo CY, Misplon JA, Liu T, Ye Z, Hogan RJ, Wu Z, Benton KA, Tumpey TM, Epstein SL (2007) Matrix protein 2 vaccination and protection against influenza viruses, including subtype H5N1. *Emerg Infect Dis.* 13(3), 426-435.

Tong S, Koller D (2001) Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research* 2, 45-66.

Tribus M (1961) Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications. Van Nostrand, New York, USA.

Trombetti GA, Merelli I, Orro A, Milanesi L (2007) BGBlast: a BLAST grid implementation with database self-updating and adaptive replication. *Stud Health Technol Inform* 126, 23-30.

Tusnády GE, Kalmár L, Simon I (2008) TOPDB: topology data bank of transmembrane proteins. *Nucleic Acids Res* 36(Database issue), D234-9.

UniProt Consortium (2008) The Universal Protein Resource (UniProt). *Nucl. Acids Res.* 36: D190-D193.

Valdar WS (2002) Scoring residue conservation. *Proteins* 48(2), 227-241.

van Vlymen J, de Lusignan S (2005) A system of metadata to control the process of query, aggregating, cleaning and analysing large datasets of primary care data. *Inform Prim Care* 13(4), 281-291.

Wald G (1967) Banquet Speech. 10 December 1967, Stockholm. http://nobelprize.org/nobel_prizes/medicine/laureates/1967/wald-speech.html (accessed 15 June 2008)

Wang X, Gorlitsky R, Almeida JS (2005) From XML to RDF: how semantic web technologies will change the design of 'omic' standards. *Nat. Biotechnol* 23(9), 1099-1103.

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y (1992): Evolution and ecology of influenza A viruses. *Microbiol Rev.* 56(1), 152-179.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-678.

Wheeler DA, Srinivasan M, Egholm M *et al.* (2008a) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452(7189), 872-876.

Wheeler DL, Barrett T, Benson DA, *et al*. (2008b) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 36, D13-21.

Widom J (1995) Research Problems in Data Warehousing. *Proc. Int. Conf. on Information and Knowledge Management (CIKM '95)* 25-30, Baltimore USA.

Wiederhold G (1992) Mediators in the Architecture of Future Information Systems. *IEEE Computer* 25, 38-49

Wiley HS, Michaels GS (2004) Should software hold data hostage? *Nat Biotechnol* 22(8), 1037-1038.

Wilkinson MD, Links M (2002) BioMOBY: an open-source biological web services proposal. *Briefings In Bioinformatics* 3(4), 331-341.

Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.

Wolstencroft K, Brass A, Horrocks I, Lord P, Sattler U, Turi D, Stevens R (2005) A Little Semantic Web Goes a Long Way in Biology. *Lecture Notes in Computer Science* 3729, 786-800.

Wong L (2000) Kleisli, a Functional Query System. *J of Functional Programming* 10(1), 19-56.

Wong L (2002) Technologies for integrating biological data. *Brief Bioinform* 3(4), 389-404.

World Health Organization (1972) A revised system of influenza virus nomenclature. A report of the WHO study group on classification. *Virology* 47(3), 854-6.

Yamada H, Chounan R, Higashi Y, Kurihara N, Kido H (2004) Mitochondrial targeting sequence of the influenza A virus PB1-F2 protein and its function in mitochondria. *FEBS Lett.* 578,1-6.

Yu U, Lee SH, Kim YJ, Kim S (2004) Bioinformatics in the post-genome era. *J Biochem Mol Biol.* 37(1), 75-82.

Yusim K, Kesmir C, Gaschen B, Addo MM, Altfeld M, Brunak S, Chigaev A, Detours V, Korber BT (2002) Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. *J Virol.* 76(17), 8757-8768.

Zagrovic B, Snow CD, Shirts MR, Pande VS (2002) Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J Mol Biol* 323(5), 927-937

Zeleny, M (1987) Management Support Systems: Towards Integrated Knowledge Management. *Human Systems Management* 7(1), 59-70.

Zeng H, Fikes R (2005). Explaining Data Incompleteness in Knowledge Aggregation. Technical Report KSL-05-04, Knowledge Systems Laboratory, Stanford University. ftp://ftp.ksl.stanford.edu/pub/KSL_Reports/KSL-05-04.pdf (accessed 15 June 2008)

Zerhouni EA (2005) US biomedical research: basic, translational, and clinical sciences. *JAMA* 294(11), 1352-1358.

Zhang ZH, Tan SC, Koh JL, Falus A, Brusic V (2006) ALLERDB database and integrated bioinformatic tools for assessment of allergenicity and allergic cross-reactivity. *Cellular Immunolology* 244(2), 90-96.

Zinkernagel RM, Hengartner H (2004) On immunity against infections and vaccines: credo 2004. *Scand J Immunol* 60(1-2), 9-13

Zuckerkandl E (2006) Intelligent design and biological complexity. *Gene* 385, 2-18.

List of Abbreviations

The following abbreviations are used in this thesis:

A2A	Avian-to-avian trasmissible (opposed to H2H).
ABK	Aggregator of Biological Knowledge, a software system developed in this thesis.
ANN	Artificial Neural Network.
AVANA	Antigenic Variability ANAlyzer, a software system developed in this thesis.
CDM	Common Data Model.
CPU	Central Processing Unit
CSV	Comma-Separated Values a computer encoding format.
DBMS	Database Management Systems.
DNA	Deoxyribonucleic acid.
DOM	Domain Object Model, a component technology of XML.
EBI	European Bioinformatics Institute.
GIS	Geographical Information Systems.
H2H	Human-to-human trasmissible.
HA	Hemagglutinin, an influenza A viral protein
HIV	Human immunodeficiency virus.
HLA	Human Leukocyte Antigen.
HTML	HyperText Markup Language, a computer encoding standard for the Web.
HTTP	HyperText Transfer Protocol, a data protocol for the Web.
IT	Information Technology
IUIS	International Union of Immunological Societies.
KDD	Knowledge Discovery in Data, a branch of computing.
M1	Matrix Protein 1, an influenza A viral protein
M2	Matrix Protein 2, an influenza A viral protein
MI	Mutual Information
MSA	Multiple Sequence Alignment.
NA	Neuraminidase, an influenza A viral protein

NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health.
NLP	Natural Language Processing, a branch of computing and linguistics.
NP	Nucleoprotein, an influenza A viral protein
NS1	Non-structural Protein 1, an influenza A viral protein
NS2	Non-structural Protein 2, an influenza A viral protein
OWL	Web Ontology Language, a component technology of the Semantic Web.
PA	Acidic RNA Polymerase, an influenza A viral protein
PB1	Basic RNA Polymerase 1, an influenza A viral protein
PB1-F2	PB1 Frame 2, an influenza A viral protein
PB2	Basic RNA Polymerase 2, an influenza A viral protein
RDBMS	Relational Database Management Systems.
RDF	Resource Description Framework, a component of the Semantic Web.
RNA	Ribonucleic acid.
RNP	Ribonucleoprotein, an influenza A viral protein complex
SOA	Service Oriented Architecture, a computing model.
SQL	Structured Query Language, a language for RDBMS systems.
SVM	Support Vector Machine, a computing algorithm.
UDDI	Universal Description, Discovery and Integration, a Web Sevices standard.
URI	Uniform Resource Identifier
WHO	World Health Organization.
WSDL	Web Services Description Language, a standard for Web Services.
XML	eXtended Markup Language, a computer encoding standard for the Web.
XSLT	XML Style Language for Transformations.

Appendix A – Reprint of Khan *et al.* (2006)

Khan AM, Miotto O, Heiny AT, Salmon J, Srinivasan KN, Nascimento EJM, Marques ET, Brusic V, Tan TW, August JT (2006)

A systematic bioinformatics approach for selection of epitope-based vaccine targets.

Cellular Immunology. 244(2), 141-147.

Available online at www.sciencedirect.com
ScienceDirect
Cellular Immunology 244 (2006) 141–147
Cellular Immunology
www.elsevier.com/locate/ycimm

A systematic bioinformatics approach for selection of epitope-based vaccine targets

Asif M. Khan [a,b], Olivo Miotto [b,c], A.T. Heiny [b], Jerome Salmon [d], K.N. Srinivasan [d,e], Eduardo J.M. Nascimento [f], Ernesto T.A. Marques Jr. [d,f,g], Vladimir Brusic [a,h], Tin Wee Tan [b], J. Thomas August [d,*]

[a] *Department of Microbiology, Yong Loo Lin School of Medicine, National University of Singapore, 5 Science Drive 2, Singapore 117597, Singapore*
[b] *Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597, Singapore*
[c] *Institute of Systems Science, National University of Singapore, 25 Heng Mui Keng Terrace, Singapore 119615, Singapore*
[d] *Department of Pharmacology and Molecular Sciences, The Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, MD 21205, USA*
[e] *Product Evaluation & Registration Division, Centre for Drug Administration, Health Sciences Authority, 11 Biopolis Way, Singapore 138667, Singapore*
[f] *Department of Medicine, Division of Infectious Diseases, The Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, MD 21205, USA*
[g] *Laboratory of Vaccinology and Experimental Therapeutics, Aggeu Magalhaes Research Center, FIOCRUZ, Brazil*
[h] *School of Land and Food Sciences, and Institute for Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia*

Received 31 January 2007; accepted 6 February 2007
Available online 16 April 2007

Abstract

Epitope-based vaccines provide a new strategy for prophylactic and therapeutic application of pathogen-specific immunity. A critical requirement of this strategy is the identification and selection of T-cell epitopes that act as vaccine targets. This study describes current methodologies for the selection process, with dengue virus as a model system. A combination of publicly available bioinformatics algorithms and computational tools are used to screen and select antigen sequences as potential T-cell epitopes of supertype human leukocyte antigen (HLA) alleles. The selected sequences are tested for biological function by their activation of T-cells of HLA transgenic mice and of pathogen infected subjects. This approach provides an experimental basis for the design of pathogen specific, T-cell epitope-based vaccines that are targeted to majority of the genetic variants of the pathogen, and are effective for a broad range of differences in human leukocyte antigens among the global human population.

Keywords: T-cell epitopes; Epitope-based vaccines; Bioinformatics; Pathogens; Immune system; Information entropy; Conserved sequences; Immunological hotspots; Altered-ligand effect; Supertypes

1. Introduction

New developments in immunoinformatics and other computational methodologies, combined with the broad versatility in the design and synthesis of genetic (DNA) vaccines, underlay new strategies for the novel design of antigen-specific, epitope-based vaccines against the many pathogens that currently have proven refractive to conventional vaccine therapy [1,2]. Early studies of epitope-based vaccines for human immunodeficiency virus (HIV)[1], malaria and tuberculosis have produced promising results [3,4], supporting the protective and therapeutic uses of these vaccines. T-cell epitopes, important for cytolytic and regulatory responses to pathogens [5–7], are necessary

* Corresponding author. Fax: +1 410 502 3066.
E-mail address: taugust@jhmi.edu (J.T. August).

[1] *Abbreviations used:* HIV, human immunodeficiency virus; HLA, human leukocyte antigen.

0008-8749/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cellimm.2007.02.005

elements of these vaccines. The rational selection of protein antigen sequences that function as T-cell epitopes in vaccine formulations is therefore crucial for successful application of this vaccination strategy [2,8].

This selection of pathogen antigen sequences to be included in epitope-based vaccines must address several determinative issues. The goal is to identify relevant T-cell epitopes, both HLA class I and II, that are both effective and sufficient in vaccine protection against pathogen challenge. A major question is the degree of protection that can be achieved without the concomitant administration of neutralizing antibody epitopes. Vaccines must also protect a broad spectrum of human population against as wide a variety of pathogenic strains as possible; this presents further challenges. Many pathogens exhibit high mutation rates, with selection of new genetic variants that are resistant to an existing immune response to earlier pathogen subtypes, or may subvert the immune response by the altered peptide ligand phenomena [9–11]. It is therefore important to choose epitopes derived from conserved peptide sequences. Also, the extreme polymorphism that characterizes human leukocyte antigens (HLAs) restricts the proportion of the human population that will respond to a particular antigen [8,12]. Thus, it is advantageous to select promiscuous T-cell epitopes that bind to several alleles of HLA supertypes for maximal population coverage [13]. The focus is on a bioinformatics-based approach as a means to enhance the optimal selection of potential targets of immune response that can then be validated by experiments that test the biological function of these antigen sequences in immune-system based assays.

In this report, we describe a combined immunoinformatics and molecular strategy for vaccine development. Based upon the growing number of bioinformatics tools and antigen sequences available in public databases [14] for identifying pathogen peptides, the *in silico* prediction of T-cell epitopes can greatly reduce the list of candidate epitopes. Such a shortlist is then the starting point for molecular experiments that can validate the vaccine targets based on the biological function of the selected antigen sequences.

2. Methodology and results

2.1. Data collection and preparation

Predictions about future mutations are derived from past evolutionary history. It is therefore important to collect sequences that are as representative as possible of the genetic variants of the pathogen, over extended periods of time and broad geographical ranges. Ideally, all available protein sequences pertaining to the pathogen should be collected from major public databases, such as the NCBI Entrez protein database (www.ncbi.nlm.nih.gov/entrez). Since public databases often contain errors, discrepancies and duplicate entries, a data cleaning process is needed to correct such anomalies [15]. For example, annotation errors and discrepancies in 17 dengue virus

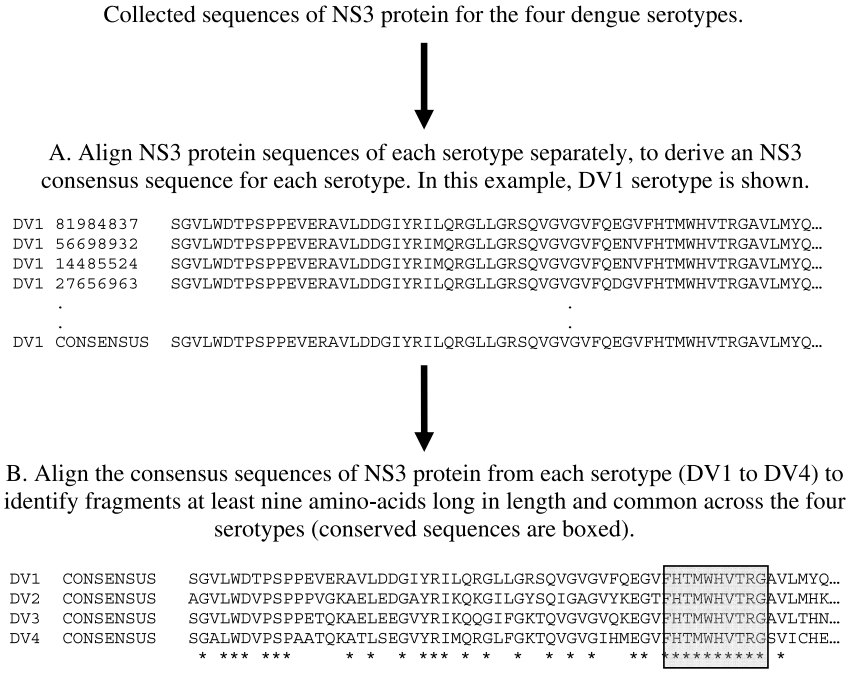


Fig. 1. Steps involved in determining sequence fragments conserved across the four serotypes for NS3 protein using a consensus-sequence-based approach. (A) The consensus sequence for NS3 protein is derived for each serotype (DV1–4) from their respective multiple sequence alignment. Each residue in the consensus sequence represents the predominant residue at that position in the corresponding multiple sequence alignment. (B) The four consensus sequences of NS3 protein (one from each serotype) are aligned to reveal sequence fragments that are at least nine amino acids long and identical across the four consensus sequences.

records were identified and corrected prior to analysis [16]. While several methods are available, we found the ABK structural rule-based approach [17] well suited to this type of task, allowing fully annotated sets of over 40,000 influenza protein sequences to be cleaned and independently verified in two weeks.

2.2. Identification of conserved sequences

The identification of conserved sequences is an initial step to overcome pathogen genomic variation that in some cases is extensive, such as HIV, influenza A viruses and dengue viruses. Multiple sequence alignments of pathogen proteins are examined by a consensus-sequence based approach [18] for the selection of sequences conserved in the large majority of variants. For pathogens with multiple groups (clades, serotypes or subtypes), pan-group consensus sequences are obtained by aligning consensus sequences derived from each of the different groups (Fig. 1), rather than by analyzing pan-group alignments that combine sequences from all groups. This prevents over-represented groups from biasing the derived consensus sequence. Identification of conserved alignment sites is based on the representation (frequency) of the consensus residue among all sequences in the alignment. Depending on the variability exhibited by different pathogen groups, the cut-off intra-group representation for conserved sequences may be set between 50% and 100%. For example, in our dengue virus analysis we only selected conserved sites common across the four serotypes, exhibiting at least 80% representation in each of the four serotypes (Fig. 2). For immunological applications, a minimum conserved sequence length of nine amino acids is required because this represents the typical length of peptides that bind to HLA molecules [19].

2.3. Entropy-based analysis of conserved sequence variability

Consensus-based methods consider each alignment site independently. However, vaccine targets are short peptides, typically 9-mers, whose combinatorial composition can produce great diversity even when adjacent sites have highly conserved residues. A more robust method based on information entropy [20] can measure the degree of variability of peptides of any length, and support inferences on their evolutionary stability. Entropy, H, representing the variability of nonamer peptides (9-mers) centered at any given alignment site, is computed from the probability, p_a of each nonamer peptide a occurring at that site:

$$H = -\sum_{a} p_a \log_2(p_a)$$

Peptides centered at any given position partially overlap peptides centered at neighbouring positions. Low entropy characterizes stable peptides, and an entropy value of 0 indicates a 100% conserved nonamer. Entropy rises with increasing variability of a site, and is affected both by the number of variants at that site, and by their respective frequency. The ABK-AVANA antigenic variability analyzer tool (O.M. et al., manuscript in preparation) can perform peptide entropy analysis. Fig. 3 shows intra- and pan-sero-

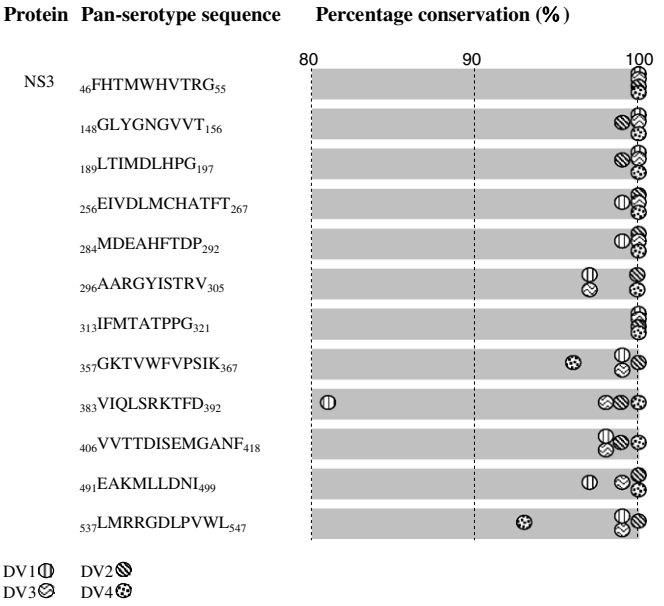


Fig. 2. Dengue pan-serotype conserved sequences of the NS3 protein and their intra-serotype percentage representation (conservation). The amino acid positions are numbered according to the aligned sequences of dengue proteins from all four serotypes.

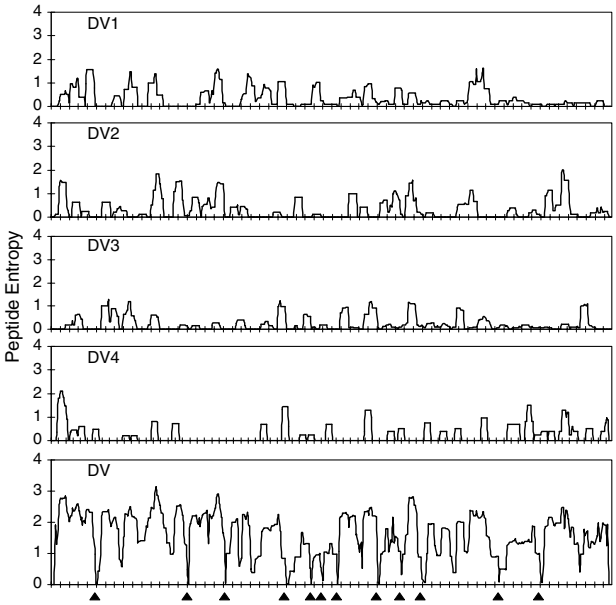


Fig. 3. Peptide entropy plots for intra- and pan-serotype alignments of dengue virus (DV) NS3 protein (intra-serotype: DV1, DV2, DV3, DV4; pan-serotype: DV). The peptide entropy value at each position is based on the frequency of nonamer peptide variants present at that position in the protein's alignment. All 12 identified pan-serotype conserved sequences of NS3 protein were found to be localized in the pan-serotype conserved antigenic regions of the protein (▲), with entropy values ranging from 0 to 0.4, indicating the high probability that these sequences will remain conserved in the future.

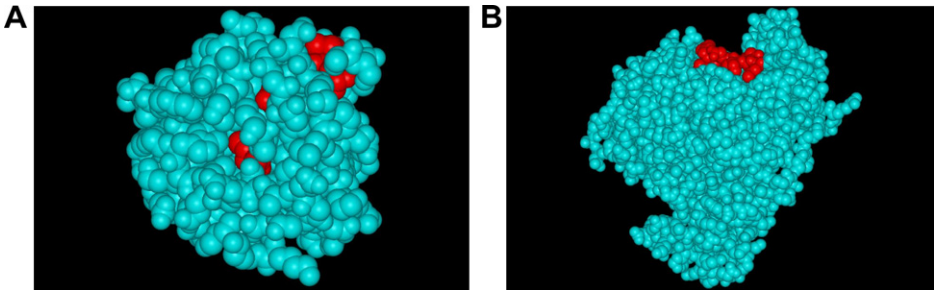


Fig. 4. Molecular location of dengue NS3 pan-serotype conserved sequences ($_{148}$GLYGNGVVT$_{156}$ and $_{189}$LTIMDLHPG$_{197}$) on the protein's 3-D structure. (A) A major portion of $_{148}$GLYGNGVVT$_{156}$ conserved sequence (in red) is localized in the buried regions of the 3-D structure. (B) Most of the $_{189}$LTIMDLHPG$_{197}$ conserved sequence (in red) is localized in the exposed region of the 3-D structure. This suggests that the conserved sequence $_{148}$GLYGNGVVT$_{156}$ is less likely to mutate compared to $_{189}$LTIMDLHPG$_{197}$, though both share identical level of intra-serotype percentage representation.

type peptide entropy plots for dengue virus NS3 protein. The data shows that each of the four serotypes has distinct patterns of highly conserved and variable regions. Thus, the pan-serotype low entropy regions were restricted to discrete short regions, which corresponded to the conserved sequences selected by consensus-sequence method.

2.4. Functional and structural correlates of the conserved sequences

It is generally recognized that conserved protein sequences represent important functional domains [21], for which mutations would be detrimental to the survival of the pathogen. The functions of conserved sequences can be elucidated by databases that comprise data on protein families, domains and functional sites, such as the Pfam database [22] (www.sanger.ac.uk/Software/Pfam). Mapping the location of a conserved sequence on the 3-D structure of the protein may also provide relevant information (Fig. 4). Many such 3-D structures are available in the PDB database [23] (www.pdb.org).

2.5. Distribution of conserved sequences in nature

Potential vaccine targets should be analyzed for specificity to the target pathogen. In vaccine design, epitopes common to other pathogens could either be useful by inducing cross-protection, or detrimental by inducing altered-ligand effect [9–11]. Identified conserved sequences should therefore be submitted to a BLAST search against all protein sequences at NCBI, excluding the target pathogen. If the sequences are found in other pathogens, the extent of their representation should be analyzed. For example, many dengue virus conserved sequences are found widely present in other *Flaviviruses*.

2.6. Characterization of candidate promiscuous T-cell epitopes

2.6.1. Algorithms for prediction of HLA binding peptides

Dedicated algorithms based on distinct prediction models are used to locate putative promiscuous T-cell epitopes for HLA class I or II supertypes within conserved sequences. Computational epitope prediction systems, such as NetCTL [24] (www.cbs.dtu.dk/services/NetCTL), MULTIPRED [25] (research.i2r.a-star.edu.sg/multipred) and TEPITOPE [26] have been proven to be effective in accurately mapping T-cell epitopes. When selecting peptides for experimental validation, putative epitopes predicted by multiple models are chosen, since consensus predictions from a combination of models have been shown to be more accurate than individual model predictions [24,27].

In addition to being promiscuous with respect to multiple alleles of an HLA supertype, some putative T-cell epitopes exhibit multiple-supertype promiscuity. This additional form of promiscuity has been observed in several viruses, such as dengue [28] and HIV [3]. T-cell epitopes specific to multiple HLA supertypes are advantageous for vaccine design because they effectively increase the numbers of epitopes to which an individual can respond, and provide much more extensive coverage of the population [3].

2.6.2. Immunological hotspots

Putative promiscuous T-cell epitopes may be localized in clusters, as reported in studies of HIV-1 [29–32] and the outer membrane of *Chlamydia trachomatis* [33], among others [34,35]. The clusters are also ideal for developing epitope-based vaccines because they contain multiple promiscuous epitopes. MULTIPRED [25] can be used to predict immunological hotspots.

2.6.3. HLA distribution analysis

The percentage of individuals in the population predicted to respond to the putative conserved promiscuous T-cell epitopes is predicted by the population coverage analysis tool of the Immune Epitope Database [36] (www.immuneepitope.org/tools/population). The tool provides allele frequencies for 78 populations grouped into 11 different geographical areas.

2.7. Probability of altered-ligand effect

The genotypic differences between primary and secondary pathogens, or between the vaccine and challenge infec-

tion, constitute a critical consideration for protective and, in some cases, pathologic immunity [11]. Because of intra- and inter-group sequence variability, most T-cell epitope sequences may contain single or multiple amino acid differences within and between the groups. Variants of the putative promiscuous T-cell epitopes are identified among the reported sequences in the pathogen groups, and their representation within the group and across groups is observed. Variants of a putative epitope at a given alignment position comprise all nonamers at that site that possess at least one amino acid difference. Putative epitopes with no or low variant representation (~100% conserved) are potentially advantageous in avoiding altered peptide ligands.

2.8. Experimental validation

2.8.1. Survey of reported human T-cell epitopes in the conserved sequences

Predictions of T-cell epitopes of the conserved sequences can in many cases be conformed (commonly without identification of the specific allele, however) by reports of experimentally confirmed T-cell epitopes. Therefore, search against both extant literature and the Immune Epitope Database (www.immuneepitope.org) is performed for reported human T-cell epitopes (both class I and II) that fully or partially overlap with identified conserved sequences. For example, eight reported human NS3 T-cell epitopes of dengue virus corresponded to the predicted promiscuous T-cell epitopes in the NS3 conserved sequences (Table 1).

Table 1
Reported human T-cell epitopes in dengue virus NS3 pan-serotype conserved sequences

Protein	Pan-serotype sequence	Reported T-cell epitopes Reference(s)
NS3	$_{46}$FHTMWHVTRG$_{55}$	[39]
	$_{148}$GLYGNGVVT$_{156}$	[39,40]
	$_{189}$LTIMDLHPG$_{197}$	[41]
	$_{256}$EIVDLMCHATFT$_{267}$	[39,42,43]
	$_{313}$IFMTATPPG$_{321}$	[39]
	$_{357}$GKTVWFVPSIK$_{367}$	[44,45]
	$_{383}$VIQLSRKTFD$_{392}$	[39]
	$_{406}$VVTTDISEMGANF$_{418}$	[39]
	$_{537}$LMRRGDLPVWL$_{547}$	[39]

The amino acid positions are numbered according to the aligned sequences of dengue proteins from all four serotypes.

2.8.2. Experimental measurements to validate predictions

Experimental measurements for validation of computational predictions are necessary for accurate interpretation of results. Such measurements currently include HLA binding assays [37], immunization of HLA transgenic mice and ELISpot assay for peptide-specific T-cell activation [38] and of pathogen infected human subjects. We performed functional assessment of the dengue virus NS1 conserved sequences: four were predicted to contain HLA-DR epitopes and three of these four were confirmed by ELISpot assay with T-cell activation peptides that closely mimic the conserved sequences (Table 2). An additional two that were also ELISpot positive were not predicted to bind to DR molecules. In summary, of seven conserved NS1 sequences, five contained HLA-DR T-cell epitopes and at least three are promiscuous for multiple HLA-DR alleles. The predictive models are helpful in selecting antigen sequences for additional study of immune responses, especially for sequences predicted by multiple algorithms.

3. Conclusion

The bioinformatics approach presented in this paper proved generic as it was successfully applied to several viruses, such as dengue virus (A.M.K. et al., manuscript in preparation), influenza (A.T.H. et al., manuscript in preparation) and HIV (K.N.S et al., manuscript in preparation). Thus, the approach can be used as a template for the analysis of other pathogens, providing a novel and generalized approach to the formulation of epitope-based vaccines that are effective against broad diversity of pathogens and applicable to the human population at large. This new methodology enables the systematic screening of pathogen data which would otherwise be impossible to carry out experimentally, due to too many pathogen sequences (high viral diversity) and variations in immune system among individuals (extensive polymorphism of HLA). It

Table 2
IFN-gamma ELISpot responses of CD8-depleted splenocytes from HLA transgenic mice immunized with peptides overlapping dengue virus NS1 pan-serotype conserved sequences

Pan-serotype sequence	Predicted DR-2, -3, -4	ELISpot positive HLA transgenic mouse	ELISpot activation peptide
$_{12}$ELKCGSGIF$_{20}$	DR-2	DR-2	$_{13}$LKCGSGIFVTNEVHT$_{27}$
$_{25}$VHTWTEQYKFQ$_{35}$	DR-4	DR-3 and -4	$_{25}$VHTWTEQYKFQADSP$_{39}$
$_{193}$AVHADMGYWIES$_{204}$	DR-2 and -3	None	$_{193}$AVHADMGYWIESQKN$_{207}$
$_{229}$HTLWSNGVLES$_{239}$	DR-3 and -4	DR-3 and -4	$_{229}$HTLWSNGVLESDMII$_{243}$
$_{266}$GPWHLGKLE$_{274}$	None	DR-3 and -4	$_{265}$AGPWHLGKLELDFNY$_{279}$
$_{294}$RGPSLRTTT$_{302}$	None	DR-4	$_{293}$TRGPSLRTTTVSGKL$_{307}$
$_{325}$GEDGCWYGMEIRP$_{337}$	None	None	$_{325}$GEDGCWYGMEIRPIS$_{339}$

The amino acid positions are numbered according to the aligned sequences of dengue proteins from all four serotypes. Prediction for DR alleles was performed by use of MULTIPRED [25], TEPITOPE [26] and ARB [46]. The ELISpot assays were performed for DR-2, DR-3 and DR-4 transgenic mice. ELISpot activation peptides are the actual peptides used to test the ELISpot.

therefore significantly reduces the efforts and cost of experimentation, while providing for systematic screening.

Acknowledgments

This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, USA, under Grant No. 5 U19 AI56541 and Contract No. HHSN2662-00400085C.

References

[1] A. Sette, J. Fikes, Epitope-based vaccines: an update on epitope identification, vaccine design and delivery, Curr. Opin. Immunol. 15 (2003) 461–470.
[2] A. Sette, B. Livingston, D. McKinney, E. Appella, J. Fikes, J. Sidney, M. Newman, R. Chesnut, The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation, Biologicals 29 (2001) 271–276.
[3] C.C. Wilson, D. McKinney, M. Anders, S. MaWhinney, J. Forster, C. Crimi, S. Southwood, A. Sette, R. Chesnut, M.J. Newman, B.D. Livingston, Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1, J. Immunol. 171 (2003) 5611–5623.
[4] H.L. Robinson, R.R. Amara, T cell vaccines for microbial infections, Nat. Med. 11 (2005) S25–S32.
[5] R.M. Zinkernagel, H. Hengartner, On immunity against infections and vaccines: credo 2004, Scand. J. Immunol. 60 (2004) 9–13.
[6] M.T. Esser, R.D. Marchese, L.S. Kierstead, L.G. Tussey, F. Wang, N. Chirmule, M.W. Washabaugh, Memory T cells and vaccines, Vaccine 21 (2003) 419–430.
[7] B. Pulendran, R. Ahmed, Translating innate immunity into immunological memory: implications for vaccine development, Cell 124 (2006) 849–863.
[8] V. Brusic, J.T. August, The changing field of vaccine development in the genomics era, Pharmacogenomics 5 (2004) 597–600.
[9] J. Sloan-Lancaster, P.M. Allen, Altered peptide ligand-induced partial T cell activation: molecular mechanisms and role in T cell biology, Annu. Rev. Immunol. 14 (1996) 1–27.
[10] B.D. Evavold, J. Sloan-Lancaster, P.M. Allen, Tickling the TCR: selective T-cell functions stimulated by altered peptide ligands, Immunol. Today 14 (1993) 602–609.
[11] A.L. Rothman, Dengue: defining protective versus pathologic immunity, J. Clin. Invest. 113 (2004) 946–951.
[12] I.G. Ovsyannikova, R.M. Jacobson, G.A. Poland, Variation in vaccine response in normal populations, Pharmacogenomics 5 (2004) 417–427.
[13] A. Sette, J. Sidney, Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism, Immunogenetics 50 (1999) 201–212.
[14] V. Brusic, V.B. Bajic, N. Petrovsky, Computational methods for prediction of T-cell epitopes—a framework for modelling, testing, and applications, Methods 34 (2004) 436–443.
[15] K.N. Srinivasan, P. Gopalakrishnakone, P.T. Tan, K.C. Chew, B. Cheng, R.M. Kini, J.L. Koh, S.H. Seah, V. Brusic, SCORPION, a molecular database of scorpion toxins, Toxicon 40 (2002) 23–31.
[16] A.M. Khan, A.T. Heiny, K.X. Lee, K.N. Srinivasan, T.W. Tan, J.T. August, Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus, BMC Bioinformatics 7 (2006) S4.
[17] O. Miotto, T.W. Tan, V. Brusic, Extraction by example: induction of structural rules for the analysis of molecular sequence data from heterogeneous sources, in: M. Gallagher, J. Hogan, F. Maire (Eds.), Lecture Notes in Computer Science 3578, Springer, Berlin, 2005, pp. 398–405.
[18] V. Novitsky, U.R. Smith, P. Gilbert, M.F. McLane, P. Chigwedere, C. Williamson, T. Ndung'u, I. Klein, S.Y. Chang, T. Peter, I. Thior, B.T. Foley, S. Gaolekwe, N. Rybak, S. Gaseitsiwe, F. Vannberg, R. Marlink, T.H. Lee, M. Essex, Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? J. Virol. 76 (2002) 5435–5451.
[19] H.G. Rammensee, Chemistry of peptides associated with MHC class I and class II molecules, Curr. Opin. Immunol. 7 (1995) 85–96.
[20] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, and 623–656.
[21] W.S. Valdar, Scoring residue conservation, Proteins 48 (2002) 227–241.
[22] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer, D.J. Studholme, C. Yeats, S.R. Eddy, The Pfam protein families database, Nucleic Acids Res. 32 (2004) D138–D141.
[23] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[24] M.V. Larsen, C. Lundegaard, K. Lamberth, S. Buus, S. Brunak, O. Lund, M. Nielsen, An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions, Eur. J. Immunol. 35 (2005) 2295–2303.
[25] G.L. Zhang, A.M. Khan, K.N. Srinivasan, J.T. August, V. Brusic, MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides, Nucleic Acids Res. 33 (2005) W172–W179.
[26] H. Bian, J. Hammer, Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE, Methods 34 (2004) 468–475.
[27] P. Donnes, O. Kohlbacher, Integrated modeling of the major events in the MHC class I antigen processing pathway, Protein Sci. 14 (2005) 2132–2140.
[28] S.J. Gagnon, W. Zeng, I. Kurane, F.A. Ennis, Identification of two epitopes on the dengue 4 virus capsid protein recognized by a serotype-specific and a panel of serotype-cross-reactive human CD4+ cytotoxic T-lymphocyte clones, J. Virol. 70 (1996) 141–147.
[29] P. Shankar, J.A. Fabry, D.M. Fong, J. Lieberman, Three regions of HIV-1 gp160 contain clusters of immunodominant CTL epitopes, Immunol. Lett. 52 (1996) 23–30.
[30] S. Surman, T.D. Lockey, K.S. Slobod, B. Jones, J.M. Riberdy, S.W. White, P.C. Doherty, J.L. Hurwitz, Localization of CD4+ T cell epitope hotspots to exposed strands of HIV envelope glycoprotein suggests structural influences on antigen processing, Proc. Natl. Acad. Sci. USA 98 (2001) 4587–4592.
[31] S.A. Brown, J. Stambas, X. Zhan, K.S. Slobod, C. Coleclough, A. Zirkel, S. Surman, S.W. White, P.C. Doherty, J.L. Hurwitz, Clustering of Th cell epitopes on exposed regions of HIV envelope despite defects in antibody activity, J. Immunol. 171 (2003) 4140–4148.
[32] J.A. Berzofsky, C.D. Pendleton, M. Clerici, J. Ahlers, D.R. Lucey, S.D. Putney, G.M. Shearer, Construction of peptides encompassing multideterminant clusters of human immunodeficiency virus envelope to induce in vitro T cell responses in mice and humans of multiple MHC types, J. Clin. Invest. 88 (1991) 876–884.
[33] S.K. Kim, R. DeMars, Epitope clusters in the major outer membrane protein of *Chlamydia trachomatis*, Curr. Opin. Immunol. 13 (2001) 429–436.
[34] V. Gupta, T.M. Tabiin, K. Sun, A. Chandrasekaran, A. Anwar, K. Yang, P. Chikhlikar, J. Salmon, V. Brusic, E.T. Marques, S.N. Kellathur, T.J. August, SARS coronavirus nucleocapsid immunodominant T-cell epitope cluster is common to both exogenous recombinant and endogenous DNA-encoded immunogens, Virology 347 (2006) 127–139.
[35] K.N. Srinivasan, G.L. Zhang, A.M. Khan, J.T. August, V. Brusic, Prediction of class I T-cell epitopes: evidence of presence of immunological hot spots inside antigens, Bioinformatics 20 (Suppl 1) (2004) I297–I302.

[36] B. Peters, J. Sidney, P. Bourne, H.H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J.V. Ponomarenko, M. Sathiamurthy, S.P. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson, A. Sette, The design and implementation of the immune epitope database and analysis resource, Immunogenetics 57 (2005) 326–336.

[37] J. Sidney, S. Southwood, C. Oseroff, M.F. del Guercio, H.M. Grey, A. Sette, Measurement of MHC/peptide interactions by gel filtration, Current Protocols in Immunology (1998) 18.13.11–18.13.19.

[38] E.F. Rosloniec, D.D. Brand, L.K. Myers, K.B. Whittington, M. Gumanovskaya, D.M. Zaller, A. Woods, D.M. Altmann, J.M. Stuart, A.H. Kang, An HLA-DR1 transgene confers susceptibility to collagen-induced arthritis elicited with human type II collagen, J. Exp. Med. 185 (1997) 1113–1122.

[39] C.P. Simmons, T. Dong, N.V. Chau, N.T. Dung, T.N. Chau, L.T.T. Thao, N.T. Dung, T.T. Hien, S. Rowland-Jones, J. Farrar, Early T-cell responses to dengue virus epitopes in Vietnamese adults with secondary dengue virus infections, J. Virol. 79 (2005) 5665–5675.

[40] I. Kurane, Y. Okamoto, L.C. Dai, L.L. Zeng, M.A. Brinton, F.A. Ennis, Flavivirus-cross-reactive, HLA-DR15-restricted epitope on NS3 recognized by human CD4+ CD8- cytotoxic T lymphocyte clones, J. Gen. Virol. 76 (Pt 9) (1995) 2243–2249.

[41] M.M. Mangada, A.L. Rothman, Altered cytokine responses of dengue-specific CD4+ T cells to heterologous serotypes, J. Immunol. 175 (2005) 2676–2683.

[42] I. Kurane, L.C. Dai, P.G. Livingston, E. Reed, F.A. Ennis, Definition of an HLA-DPw2-restricted epitope on NS3, recognized by a dengue virus serotype-cross-reactive human CD4+ CD8- cytotoxic T-cell clone, J. Virol. 67 (1993) 6285–6288.

[43] Y. Okamoto, I. Kurane, A.M. Leporati, F.A. Ennis, Definition of the region on NS3 which contains multiple epitopes recognized by dengue virus serotype-cross-reactive and flavivirus-cross-reactive, HLA-DPw2-restricted CD4+ T cell clones, J. Gen. Virol. 79 (Pt 4) (1998) 697–704.

[44] L. Zeng, I. Kurane, Y. Okamoto, F.A. Ennis, M.A. Brinton, Identification of amino acids involved in recognition by dengue virus NS3-specific, HLA-DR15-restricted cytotoxic CD4+ T-cell clones, J. Virol. 70 (1996) 3108–3117.

[45] H. Loke, D.B. Bethell, C.X. Phuong, M. Dung, J. Schneider, N.J. White, N.P. Day, J. Farrar, A.V. Hill, Strong HLA class I—restricted T cell responses in dengue hemorrhagic fever: a double-edged sword? J. Infect. Dis. 184 (2001) 1369–1373.

[46] H.H. Bui, J. Sidney, B. Peters, M. Sathiamurthy, A. Sinichi, K.A. Purton, B.R. Mothe, F.V. Chisari, D.I. Watkins, A. Sette, Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications, Immunogenetics 57 (2005) 304–314.

Appendix B – Reprint of Heiny *et al.* (2007)

Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, Brusic V, Tan TW, August JT (2007)

Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets.

PLoS ONE. 2(11), e1190.

OPEN ACCESS Freely available online

PLoS one

Evolutionarily Conserved Protein Sequences of Influenza A Viruses, Avian and Human, as Vaccine Targets

A. T. Heiny[1], Olivo Miotto[1,2], Kellathur N. Srinivasan[3,4], Asif M. Khan[1,5], G. L. Zhang[6], Vladimir Brusic[7], Tin Wee Tan[1], J. Thomas August[3]*

1 Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, **2** Institute of Systems Science, National University of Singapore, Singapore, Singapore, **3** Department of Pharmacology and Molecular Sciences, The Johns Hopkins University School of Medicine, Maryland, United States of America, **4** Product Evaluation and Registration Division, Centre for Drug Administration, Health Sciences Authority, Singapore, Singapore, **5** Department of Microbiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, **6** Institute for Infocomm Research, Singapore, Singapore, **7** Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America

Background. **Influenza A viruses generate an extreme genetic diversity through point mutation and gene segment exchange, resulting in many new strains that emerge from the animal reservoirs, among which was the recent highly pathogenic H5N1 virus. This genetic diversity also endows these viruses with a dynamic adaptability to their habitats, one result being the rapid selection of genomic variants that resist the immune responses of infected hosts. With the possibility of an influenza A pandemic, a critical need is a vaccine that will recognize and protect against any influenza A pathogen. One feasible approach is a vaccine containing conserved immunogenic protein sequences that represent the genotypic diversity of all current and future avian and human influenza viruses as an alternative to current vaccines that address only the known circulating virus strains.** ***Methodology/Principal Findings.*** **Methodologies for large-scale analysis of the evolutionary variability of the influenza A virus proteins recorded in public databases were developed and used to elucidate the amino acid sequence diversity and conservation of 36,343 sequences of the 11 viral proteins of the recorded virus isolates of the past 30 years. Technologies were also applied to identify the conserved amino acid sequences from isolates of the past decade, and to evaluate the predicted human lymphocyte antigen (HLA) supertype-restricted class I and II T-cell epitopes of the conserved sequences. Fifty-five (55) sequences of 9 or more amino acids of the polymerases (PB2, PB1, and PA), nucleoprotein (NP), and matrix 1 (M1) proteins were completely conserved in at least 80%, many in 95 to 100%, of the avian and human influenza A virus isolates despite the marked evolutionary variability of the viruses. Almost all (50) of these conserved sequences contained putative supertype HLA class I or class II epitopes as predicted by 4 peptide-HLA binding algorithms. Additionally, data of the Immune Epitope Database (IEDB) include 29 experimentally identified HLA class I and II T-cell epitopes present in 14 of the conserved sequences.** ***Conclusions/Significance.*** **This study of all reported influenza A virus protein sequences, avian and human, has identified 55 highly conserved sequences, most of which are predicted to have immune relevance as T-cell epitopes. This is a necessary first step in the design and analysis of a polyepitope, pan-influenza A vaccine. In addition to the application described herein, these technologies can be applied to other pathogens and to other therapeutic modalities designed to attack DNA, RNA, or protein sequences critical to pathogen function.**

Citation: Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, et al (2007) Evolutionarily Conserved Protein Sequences of Influenza A Viruses, Avian and Human, as Vaccine Targets. PLoS ONE 2(11): e1190. doi:10.1371/journal.pone.0001190

INTRODUCTION

One of the most important threats to human health is infection by avian influenza A viruses [1-3]. While global influenza pandemics have occurred only a few times in the past century, the H1N1 pandemic of 1918–1919 caused 20–50 million deaths and was one of the most serious disease outbreaks in recorded history. The recent evolution of the highly lethal avian H5N1 virus, while not transmissible in humans, has emphasized the continued threat of influenza viruses on a global scale. It is widely predicted, given the increased human population and density, that a new pandemic on the scale of the H1N1 infection would have a devastating effect world-wide.

The two currently approved vaccines against influenza viruses are designed specifically to mimic the most recently recognized circulating forms listed in the 2006–2007 influenza prevention and control recommendations (http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5510a1.htm). Both vaccines contain three recently isolated human strains and are subject to possible annual revision of their virus composition. The rapid mutation of the viral HA and NA proteins facilitates the selective replication of new virus strains not subject to immunity based on previous vaccination and is a serious obstacle to the effectiveness of these vaccines [4–5]. Alternative vaccine strategies that overcome the problem of rapid viral mutation, can be applied to global populations, and provide for easy production are suggested goals [6–8].

The design of a vaccine that guarantees antibody-mediated immunity to new influenza viruses is not currently feasible because the structural determinants of B-cell immunity are highly complex and there is no effective means for predicting the antibody epitope

Academic Editor: Berend Snel, Utrecht University, Netherlands

Received September 5, 2007; **Accepted** October 17, 2007; **Published** November 21, 2007

Funding: The development of the computational tools reported herein was supported by in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, USA, under Grant No. 5 U19 AI56541 and Contract No. HHSN2662-00400085C.

Competing Interests: The authors have declared that no competing interests exist.

* **To whom correspondence should be addressed.** E-mail: taugust@bs.jhmi.edu

structure of target pathogens. Cell-mediated immunity, in contrast, is based upon the binding of short sequences of antigen proteins, termed T-cell epitopes, to specialized cellular proteins, known as human leukocyte antigens (HLAs), class I (HLA I) and class II (HLA II), that facilitate the presentation of the epitopes to T-cells of the immune system [9–14]. The chemical and structural determinants of HLA-peptide binding have been defined for a number of HLA alleles [15–19]. Of particular relevance for vaccine design are supertype groupings of similar HLA alleles that display overlapping peptide-binding capacities. The supertypes cover a large fraction of the HLA diversity in the human population and antigen epitopes that bind to the supertypes are considered prime candidates for vaccine formulations [20–24]. Supertype-binding motifs and quantitative matrices have been incorporated into several computational prediction algorithms and it is now possible to identify, *in silico*, candidate HLA-restricted T-cell epitopes of protein sequences, allowing large-scale analysis of potential vaccine targets [24–27]. Moreover, increasing attention is being given to T-cell-based vaccines because they can be designed as genetic formulations to include selected regions of the viral antigens [28–31], and have the many other desirable properties associated with DNA vaccines in general. Studies have demonstrated that epitope-specific T cell responses elicited by immunization with DNA or peptide, and adoptive transfer of epitope-specific T cell clones, could mediate protective immunity, in some cases with single CTL epitopes, against various pathogens in murine experimental models [32–42]. Additionally, recent studies have shown that immunization of HLA-A2 transgenic mice against single HLA-A2-restricted T-cell epitopes conferred protection against lethal infection with influenza A virus, vaccinia virus, or LCMV [43–45]. Human clinical trials with epitope-based DNA vaccines against HIV [46] and malaria [47] were found to be safe and immunogenic for effector T-cell immune responses but in these first generation studies, failed to achieve the desired clinical goals in the vaccination of healthy volunteers.

Cellular immune responses are recognized to play a role in influenza immunity (for reviews see [48–51] and the application of T-cell epitopes has been extensively studied as an alternative to vaccines designed for humoral immunity [52–62]. Mouse immunization with DNA encoding NP elicited CTL, IFN-γ and IL-2 responses, with cross-strain protection against virus challenge, and evidence from adoptive transfer, indicated that both types of T cells act as effectors in protective immunity [54–55]. Similarly, DNA immunization with H1N1 NP or H5N1 NP or M proteins was found to protect mice against lethal challenge [56–58]. However, some influenza vaccine formulations were not successful [59–62] and there remain multiple issues for the development of a human T-cell epitope-based vaccines, including epitope selection, delivery systems, epitope processing and presentation, and undoubtedly others.

This study was focused on the large-scale analysis of all influenza A virus protein sequence data of the past 30 years that is recorded in public databases. Information entropy and consensus sequence methodologies were combined to identify sequences of 9 amino acids or longer with a history of complete conservation in 80% or more of both avian and human virus strains. These conserved sequences were further analyzed to identify targets for candidate epitope-based T-cell vaccine formulations against all current and possibly future influenza A pathogens of avian or human origin.

METHODS

Methodology overview

A general overview of the methodology is depicted in Figure 1. The details and rationale for the systematic approach adopted by this study have been previously published [63]. The primary goal was to identify viral protein sequences that have been conserved over long periods of time, and to select those sequences that have the highest potential for HLA-restricted immunogenicity in a broad spectrum of the human population. The process includes three major steps: i) extraction of all influenza A protein sequence data, processing and definition of data sets comprising relevant human and avian virus groups; ii) entropy analysis of sequence variability and identification of conserved peptide sequences of 9 amino acids or longer; and iii) prediction of supertype-restricted, HLA-binding sequences. Two complementary methods for the identification of conserved sequences were applied: a statistical entropy-based method that takes into account the combinatorial diversity of peptide epitopes and was used to elucidate the variability for different influenza subtypes, and a consensus method, which is robust against sampling biases (such as the predominance of certain influenza subtypes in the dataset), to confirm the conserved sequences. The conserved sequences were then submitted to several epitope prediction programs, whose results are combined. Sequences predicted, and in some cases demonstrated, to contain epitopes to

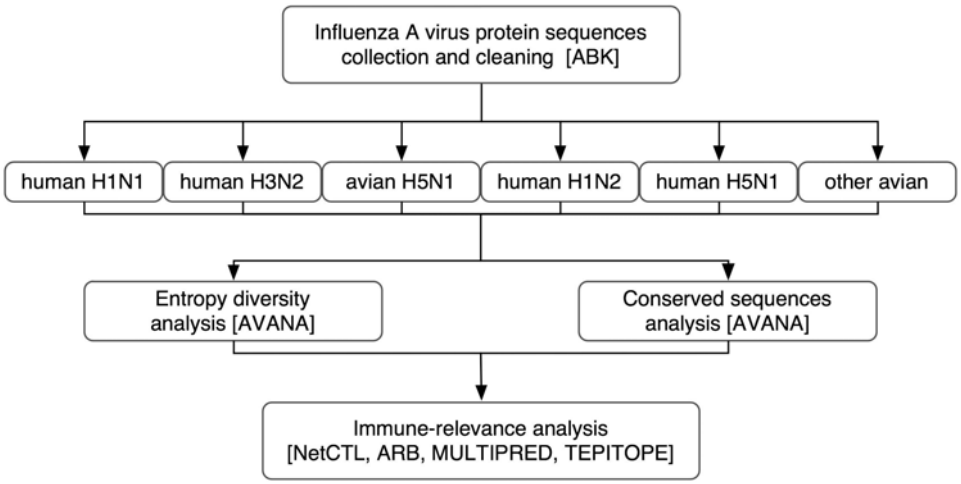


Figure 1. An overview of the methodology of this study.
doi:10.1371/journal.pone.0001190.g001

several HLA supertypes are proposed as vaccine epitope candidates because of their wide human population coverage.

Influenza A virus sequence data collection and processing

A dataset of influenza A protein sequences annotated with isolate name, country and year of isolation, host organism, subtype, and protein name, was derived from all available sequences (as of September 2006) from the NCBI GenBank and GenPept databases, including entries mirrored from the UniProt database. Collection and cleaning of a total of 85,873 records was performed the Aggregator of Biological Knowledge (ABK) [64], which applied structural and semantic rules to automate the aggregation and annotation task. The final set of 36,343 protein sequences was manually verified by two independent curators. Most human influenza subtypes were represented by more than 100 sequences of each viral protein, the count varying depending on the protein. The H1N2 subtype had a lower number of sequences (ranging from 22 to 66) because of its recent emergence [65]. Separate multiple sequence alignments of the 11 proteins were carried out with MUSCLE 3.6 [66]. Because of the great variability exhibited by the HA and NA proteins, separate alignments were obtained for each subtype (16 subtypes for HA and 9 for NA). The subtype alignments were subsequently merged using the MUSCLE tool, to obtain the final HA and NA alignments. The introduction of gaps in the resulting alignments was minimized by merging sequences based on sequence similarity between subtypes, as reported in phylogenetic studies [5,67]. The in-house developed Antigenic Variability Analyser tool (AVANA) was subsequently used to extract alignments of several subsets of the collected sequences, based on annotation values, such as viral subtype, host, and year of isolation.

Information entropy analysis

The diversity of the influenza A virus proteome was studied by creating subsets of the influenza A protein sequence alignments, comprising (1) avian sequences, subdivided into 3 decades (1977–1986, 1987–1996, and 1997–2006); (2) H5N1 viruses, subdivided into avian and human isolates; (3) circulating human subtypes, namely H1N1, H3N2, and H1N2. Assuming that each sequence represents an independent isolate, the information entropy methodology [68] was used to measure the variability of influenza A virus proteomes in the context of overlapping nine-amino acid peptides spanning the length of each influenza A protein. The rationale of this selection was the length of peptides that are bound by HLA molecules for presentation to T-cell receptors, typically from 8–20 amino acids, with nine amino acids being the predominant length of class I peptides and the core of class II peptides [69]. Applying Shannon's formula [68], the nonamer peptide entropy $H(x)$ at any given position x in the alignment is computed by

$$H(x) = -\sum_{i=1}^{n(x)} p(i,x)\log_2 p(i,x)$$

where $p(i,x)$ is the probability of a particular nonamer variant i being centered at position x. The entropy value increases with $n(x)$, the total number of variants observed at position x; it is also sensitive to the relative frequency of the variants, such that it decreases when one variant is clearly dominant (*i.e.* the position is conserved). Only sequences that contain a valid amino acid at position x were used for the entropy computation, and alignment gaps were ignored. Although gaps tend to occur in high-diversity regions, proteins that have a high fraction of gaps have reduced statistical support, yielding an artificially low entropy value; for this reason, positions where more than 50% of sequences contained a gap were discarded. Because of the statistical nature of the entropy measure, both complete protein sequences and shorter fragments were used in this computation.

In theory, nonamer entropy values can range from 0, a completely conserved nonamer sequence in all proteins analyzed, to 2^9; in practice, however, the upper bound is very much lower for alignments of closely related sequences. For finite-size sets of sequences, entropy computations are affected by the sequence count in the alignment. The effects of alignment size bias are especially noticeable for alignments containing fewer than about 100 sequences, and must be accounted for when making direct comparisons between sequence alignments of different sizes. It has been shown that, for an alignment of N sequences, alignment size bias is proportional to $1/N$ [70]. This relationship allows a correction for size bias by applying to each alignment a statistical adjustment that estimates entropy values for an infinitely-sized alignment with analogous variant distribution. To obtain such estimate, the alignment was repeatedly randomly sampled to create smaller alignments of varying size, whose entropy was measured. At each alignment position, the entropy of these subset alignments of size N was plotted against $1/N$, using a linear regression to extrapolate the entropy estimate for $N\rightarrow\infty$. The regression's coefficient of determination (r^2) was used as a goodness-of-fit of the resulting estimate, confirming the validity of our method ($r^2>0.9$ in most cases). In this study, size bias correction was applied to all entropy calculations, so that alignment sequence counts could be ignored in comparisons. All entropy values reported are therefore infinite-size set estimates, rather than the values directly computed from the alignments.

Conserved influenza A virus sequences

Collected and cleaned influenza A virus records were grouped based on (a) subtype: the circulating human subtypes (H1N1, H3N2, H1N2), H5N1, and other subtypes in avian reservoir; (b) host: human and avian; and (c) year of isolation. The method gave equal weight to all groups and obviated the problem of particular groups being over-represented (such as human H3N2). Six subgroups were derived: (1) human H1N1, (2) human H3N2, (3) human H1N2, (4) human H5N1, (5) avian H5N1, and (6) other avian subtypes. The eleven influenza A proteins of each subgroups were individually aligned using MUSCLE 3.6 [66]. The AVANA tool was used to select nonamers with conservation of $\geq$80% in each alignment. The minimum length of a conserved sequence was nine amino acids and conserved contiguous nonamers were joined as a single sequence. A consensus sequence (the most frequent sequence) for each conserved sequence in the alignments was generated for each of the 6 subgroups. Corresponding consensus sequences of the subgroups were then aligned and those sequences that were identical in each of the six subgroups and present in at least 80% of all recorded viruses were selected as the highly conserved sequences.

HLA supertype-restricted T-cell epitopes

The *in silico* prediction of HLA supertype-restricted HLA class I and class II T-cell epitope sequences in the conserved region*s* was performed through four computational systems: NetCTL MULTIPRED, ARB, and TEPITOPE. The NetCTL 1.2 algorithm [25] (http://www.cbs.dtu.dk/services/NetCTL/) predicts peptides restricted to 12 HLA class I supertypes (A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58 and B62), integrated with

predictions of HLA binding, proteasomal C-terminal cleavage and transport efficiency by the transporter associated with antigen processing (TAP) molecules. HLA binding and proteasomal cleavage predictions are performed by an artificial neural networks (ANN) method and TAP transport efficiency is predicted using a weight matrix method. The parameters used for NetCTL prediction were: 0.15 weight on C terminal cleavage (default), 0.05 weight on TAP transport efficiency (default), and 0.5 threshold for HLA supertype binding.

The Average Relative Binding (ARB) matrix binding prediction method (http://epitope.liai.org:8080/matrix/matrix_prediction.jsp) [27] is allele specific and estimates a matrix of coefficients based upon the association of each of the 20 amino acids at each possible position along the peptide sequence. In this study the data were selected for representative alleles within studied supertypes and predictions are shown for 8 HLA class I alleles of the supertypes A1 (A*0101), A2 (A*0201), A3 (A*0301), A24 (A*2402), A26 (A*2601), B7 (B*0702), and B44 (B*4402, B*4403).

MULTIPRED (http://research.i2r.a-star.edu.sg/multipred/) [26] predicts peptides that bind to HLA class I supertypes A2 (A*0201, *0202, *0203, *0204, *0205, *0206, *0207 and *0209) and A3 (A*0301, *0302, *1101, *1102, *3101, *3301 and *6801)and class II HLA-DR supertype (DRB1*0101, *0401, *1501, *0701, *0901, *1302 and DRB5*0101). Hidden Markov model (HMM) and ANN methods are the predictive engines with sum thresholds of: A2, 31.33 (ANN; SN = 0.80 and SP = 0.83) and 47.08 (HMM; SN = 0.80 and SP = 0.78); A3, 24.53 (ANN; SN = 0.90 and SP = 0.95) and 37.58 (HMM; SN = 0.80 and SP = 0.87); and DR, 23.42 (ANN; SN = 0.90 and SP = 0.92) and 51.08 (HMM; SN = 0.90 and SP = 1.00). TEPITOPE predicts 25 HLA class II (DR) alleles are HLA allele-specific; however, sequences predicted to bind to ≥5 alleles were considered supertypic.

The TEPITOPE software [24] (2000 beta version; obtained by the courtesy of J. Hammer) utilizes quantitative matrix-based motifs, obtained from experimental scanning of the binding of P1-anchored designer peptides to soluble HLA-DR molecules in *in-vitro* competition assays, to predict peptides binding to 25 common HLA-DR alleles (DRB1*0101, *0102, *0301, *0401, *0402, *0404, *0405, *0410, *0421, *0701, *0801, *0802, *0804, *0806, *1101, *1104, *1106, *1107, *1305, *1307, *1311, *1321, *1501, *1502 and DRB5*0101). The parameters for TEPITOPE predictions were: 5% quantitative threshold and putative determinants with a 10-fold inhibitory residue excluded. Predictions were performed for all 25 HLA-DR alleles and nonamer core peptides predicted to bind >5 HLA-DR alleles were selected as supertype-restricted.

Experimentally identified influenza A T-cell epitopes

T-cell epitope sequences within the conserved sequences were identified by matching the highly conserved sequences and the curated influenza epitope sequences obtained from the Immune Epitope Database and Analysis Resource (www.immuneepitope.org/) [71,72]. These epitope sequences data were derived from reported HLA binding assays (IC50≤500 nM) or T-cell assays that included ^{51}Cr release, HLA tetramer staining, and ELISPOT assays. Only epitope data from unique sequences and containing HLA restriction information were included.

RESULTS

Avian and human influenza A virus isolates

The collected and cleaned influenza A virus protein sequences were catalogued in two groups. The recently circulating (1997–2006) influenza A viruses, both avian and human comprising 25,812 sequences of the 11 influenza proteins, both full- and partial-length, from human H1N1 (2,466 sequences), H3N2 (12,199), H1N2 (405), H5N1 (1,055), avian H5N1 (4,361), and all other avian subtypes except H5N1 (5,326) (Table 1). There were over 100 sequences of each protein of every virus with exceptions for the most recent human viruses, H1N2 and H5N1, and the PB1-F2 protein. The second group comprised an additional 10,531 sequences of human H1N1 and H3N2 isolated prior to 1997, and all avian viruses isolated from 1977 to 1986, and 1987 to 1996 (Table 2). The H1N2 sequences before 1997 were excluded because the number of sequences that were available for analysis was insufficient.

Diversity of influenza A virus proteins

The diversity in the protein sequences of influenza A viruses was examined by application of the information entropy methodology to each 9 amino acid sequence of the viral proteins. Data of the

Table 1. Influenza types A virus protein sequences from the past decade (1997–2006).

Protein	Human H1N1	Human H3N2	Human H1N2	Human H5N1	Avian H5N1[a]	Other Avian[b]	Total
PB2	189	970	33	97	404	401	2,094
PB1	202	984	32	101	400	399	2,118
PB1-F2	183	955	22	47	10	74	1,291
PA	190	970	29	102	402	390	2,083
HA	517	2,032	66	106	657	976	4,354
NP	191	1,012	39	114	420	518	2,294
NA	230	1,245	49	112	577	570	2,783
M1	192	1,024	40	105	458	617	2,436
M2	192	1,045	31	95	289	335	1,987
NS1	190	984	36	95	456	662	2,423
NS2	190	978	28	81	288	384	1,949
Total	2,466	12,199	405	1,055	4,361	5,326	25,812

[a]All available sequences in the database, mainly from the past decade (1997–2006).
[b]Other avian subtypes except H5N1, from 1997 to 2006.
doi:10.1371/journal.pone.0001190.t001

Table 2. Influenza A virus protein sequences from virus isolates before 1997.

Protein	Human H1N1	Human H3N2	Other Avian[a]		Total
			1977–1986	1987–1996	
PB2	98	337	200	133	768
PB1	106	342	200	134	782
PB1-F2	81	301	181	95	658
PA	96	334	190	135	755
HA	266	1,071	326	252	1,915
NP	133	544	220	153	1,050
NA	142	443	242	145	972
M1	122	398	264	163	947
M2	106	389	258	130	883
NS1	123	373	240	198	934
NS2	111	361	222	173	867
Total	1,384	4,893	2,543	1,711	10,531

[a]Other avian subtypes of influenza A viruses except H5N1.
doi:10.1371/journal.pone.0001190.t002

past 30 years comprised 9,640 avian influenza A subtype sequences (1977–1986, 2,543 sequences; 1987–1996, 1,711; 1997–2006, 5,326 (Table 1 and 2, Figure 2). The gross patterns of protein variability of the avian viruses from each of the past 3 decades were very similar in the context of the relative diversity of the proteins. The viral surface glycoproteins, HA and NA, showed extreme sequence diversity, illustrative of the reassortment of the genome segments among the many subtypes of the avian group A viruses as well as the rapid rate of point mutation, with multiple amino acids at virtually every position (entropy >2.0) except at a single region in HA that has remained remarkably conserved despite the extreme sequence modification of every other nonamer of the protein. The PB1-F2, NS1 and NS2, and to a lesser extent M2, also showed a history of high variability. In contrast, the polymerase proteins (PB2, PB1, and PA), as well as the NP and M1, contained many historically highly conserved regions (entropy <1.0). The overall gradual increase in entropy over the three decades in many of the protein sequences, most apparent in the highly conserved sequences, is an indication of the continuing genetic evolution of the viruses as well as improved screening of sequence variants. However, these changes do not distort the overall pattern of highly conserved and highly variable sequences.

H5N1 protein entropy patterns of the 1997 to 2006 isolates from humans (1,055 sequences) and birds (4,361) were grossly similar and reflect the high mutational variability in amino acid composition of proteins of even a single subtype influenza A virus (Figure 3). Of the eleven known influenza A proteins, only the short PB1-F2 protein, the product of an alternative ORF of the PB1 RNA segment [73] showed notable differences when comparing the diversity profiles of the two groups. There is evidence that PB1-F2 is involved in the apoptosis of host immune cells, increased viral virulence in a mouse model, and destruction of alveolar macrophages [74], and the limited diversity of this protein in human isolates may have relevance to H5N1 virulence and pathogenicity. In the context of the remaining proteins, the similarity of human and avian H5N1 entropy patterns is consistent with the observations that, to date, all human H5N1 isolates represent avian to human transmission from isolated clusters of avian infection. Moreover, detailed analyses of mutations associated with human-to-human transmission have shown that all human H5N1 virus isolates have a predominant avian footprint [75]

Entropy of the protein sequences of each of the three circulating human viruses isolated between 1918 and 2006 (H1N1, 3,850 sequences; H3N2, 17,092; and H1N2, 414 [including 9 sequences before 1997]) reflect different patterns of sequence evolution (Figure 4). The complex protein sequence diversity pattern of the human H1N1 reflects its mutational evolution from its avian characteristics at the time of human transmission in 1918 to a sequence characteristic of human H1N1, with the greatest evolutionary diversity in the HA, NA, PB1-F2, NS1 and NS2, and to a lesser extent M2, similar to the diversity of the viruses in the avian host. There subsequently was further evolution of the

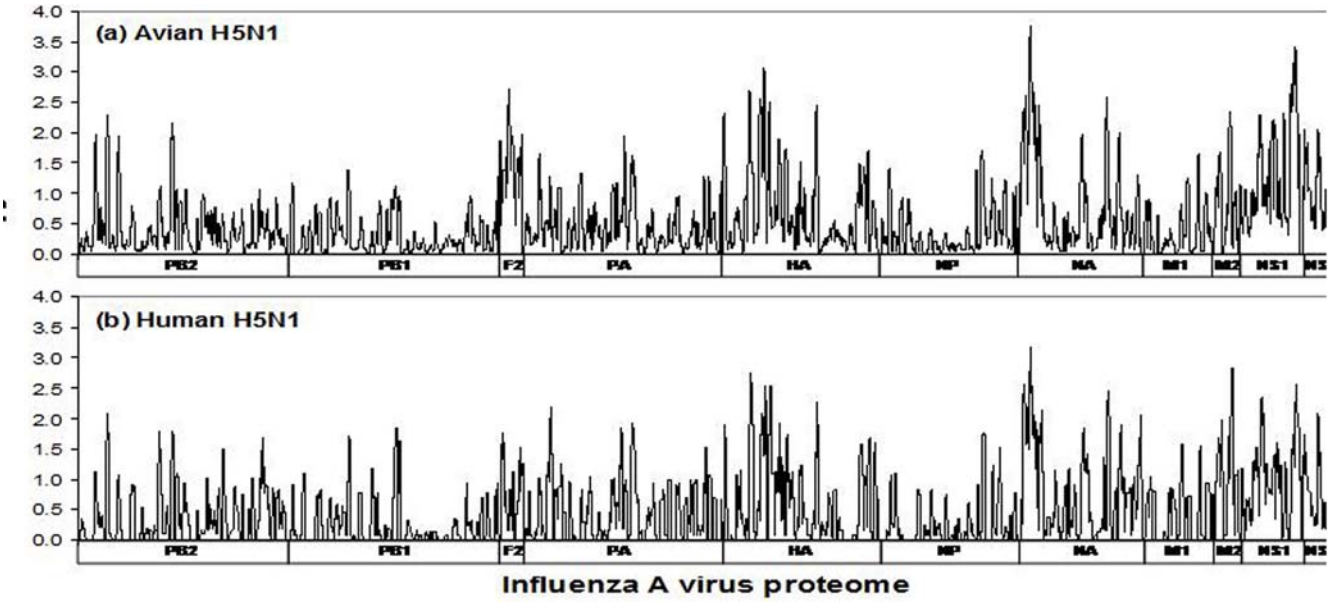


Figure 2. Entropy plots of avian influenza A viruses, excluding H5N1 subtype, for each of three decades: 1977–1986, 1987–1996, 1997–2006 (data as of September 30, 2006).
doi:10.1371/journal.pone.0001190.g002

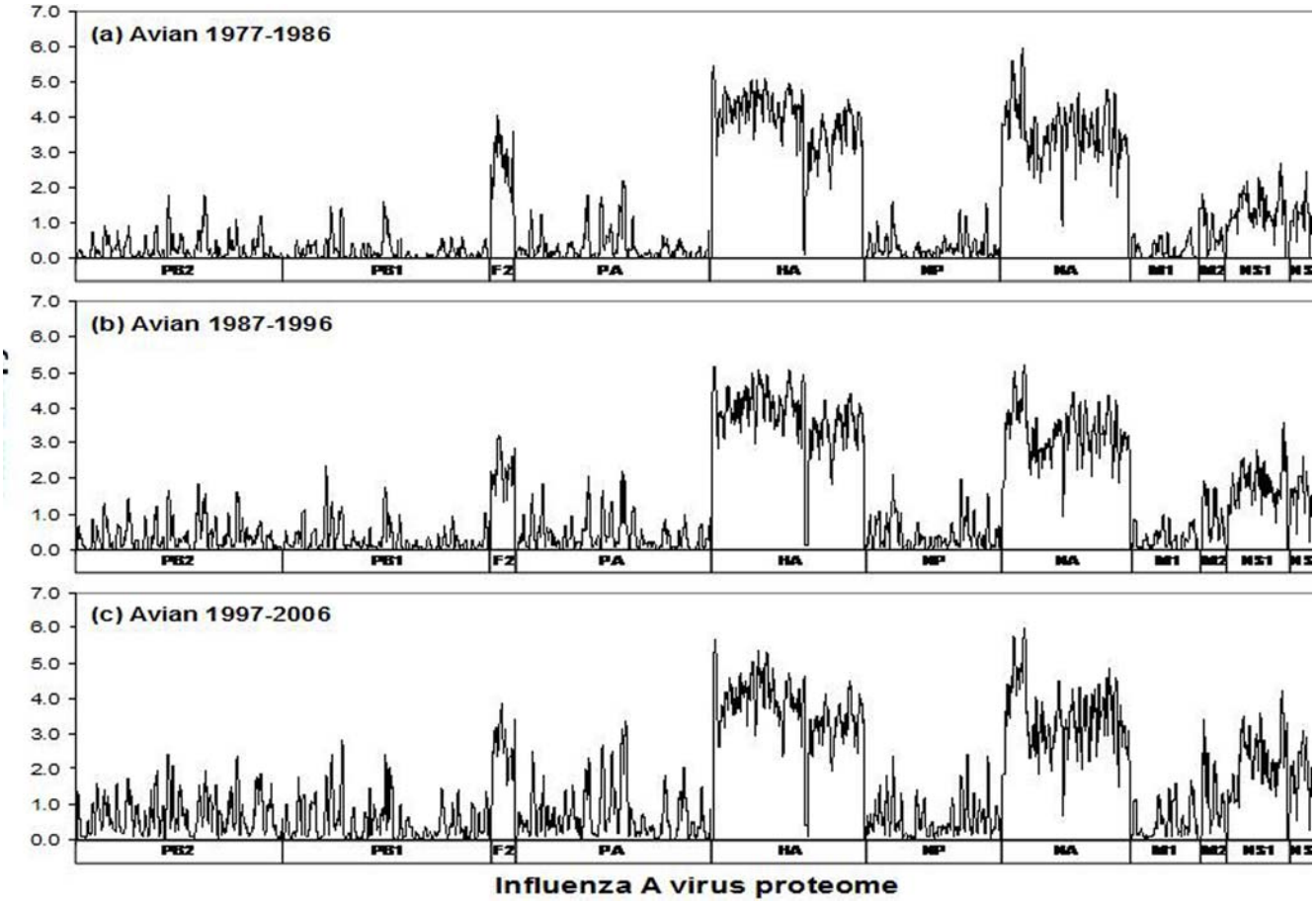


Figure 3. Entropy plots of the sequence alignments of recorded H5N1 viruses isolated from avian and human hosts (data as of September 30, 2006).
doi:10.1371/journal.pone.0001190.g003

human subtypes by gene segment exchange, resulting in H2N2 in 1957, H3N2 in 1968, and H1N2 in 1988. The continuing mutational modification of H1N2 and H3N2 have resulted in entropy patterns distinctive of the human transmitted influenza A viruses with a large number of amino acid sequence patterns that differ from those of the avian to avian counterpart. In contrast, the most recent H1N2 human subtype that appeared in 1988 (www.cdc.gov/flu/about/h1n2.htm) continues to exhibit limited evolutionary variability with many identical or highly conserved sequences regions in all of the few (22 to 66) recorded individual protein sequences (see Table 1). It is likely that the human H1N2 virus evolved from a very limited, perhaps single reassortment of the HA gene segment in the case of an individual infected with both of the human transmitted H1N1 and H3N2 viruses.

The nature of the entropy distribution of the conserved sequences is not demonstrated in these data as entropy is not a linear function but is defined both by the number of sites and frequency of variability. A given entropy value can be related to a high fraction of different amino acids at one site and limited variability at other amino acid sites, or to limited variability at a large number of amino acid sites. This absence of a direct correlation of entropy to the degree of sequence conservation is seen in the markedly diverse nonamer entropy values (~0.7 to 1.5) of the collected sequences with 80% conservation (Figure 5). A more limited range of entropy values can be associated with sequence conservation of 90–100%.

We concluded that the PB2, PB1, PA, NP, and M1 proteins of all recorded influenza A viruses, both avian and human, contain sequences of low variability and high conservation despite differences in evolutionary pathway, subtypes, and host species. These sequences with a history and predicted future of low variability are prime targets for epitope-based T-cell vaccine formulations.

Amino acid composition of the highly conserved sequences

A total of 55 peptide sequences, ranging from 9 to 58 amino acids in length, and containing a total of 965 amino acids, ~21% of the total proteome (Table 3), were completely conserved in 80% to100% of the human and avian type A viruses recorded in the past decade (Figure 6, Table S1). Twenty-six (26) were present in 90% to100% of the viruses. The majority of the conserved sequences were in the nonstructural (NS) proteins. PB2 was the most conserved with 23 sequences, comprising 50% of the protein, conserved in 80% to 100% of the documented viruses (Table 3). PB1 was also highly conserved (11 sequences, 36%) and the PA, NP, and M1 proteins contained significant fractions (16% to 27%) of conserved sequences. HA contained one sequence, FGAIAG-FIE, that was conserved in all type A viruses despite the extreme variability of all other HA amino acids (see Figure 2). There were no sequences in the PB1-F2, NA, M2, NS1 or NS2 proteins that were completely conserved in at least 80% of the viruses.

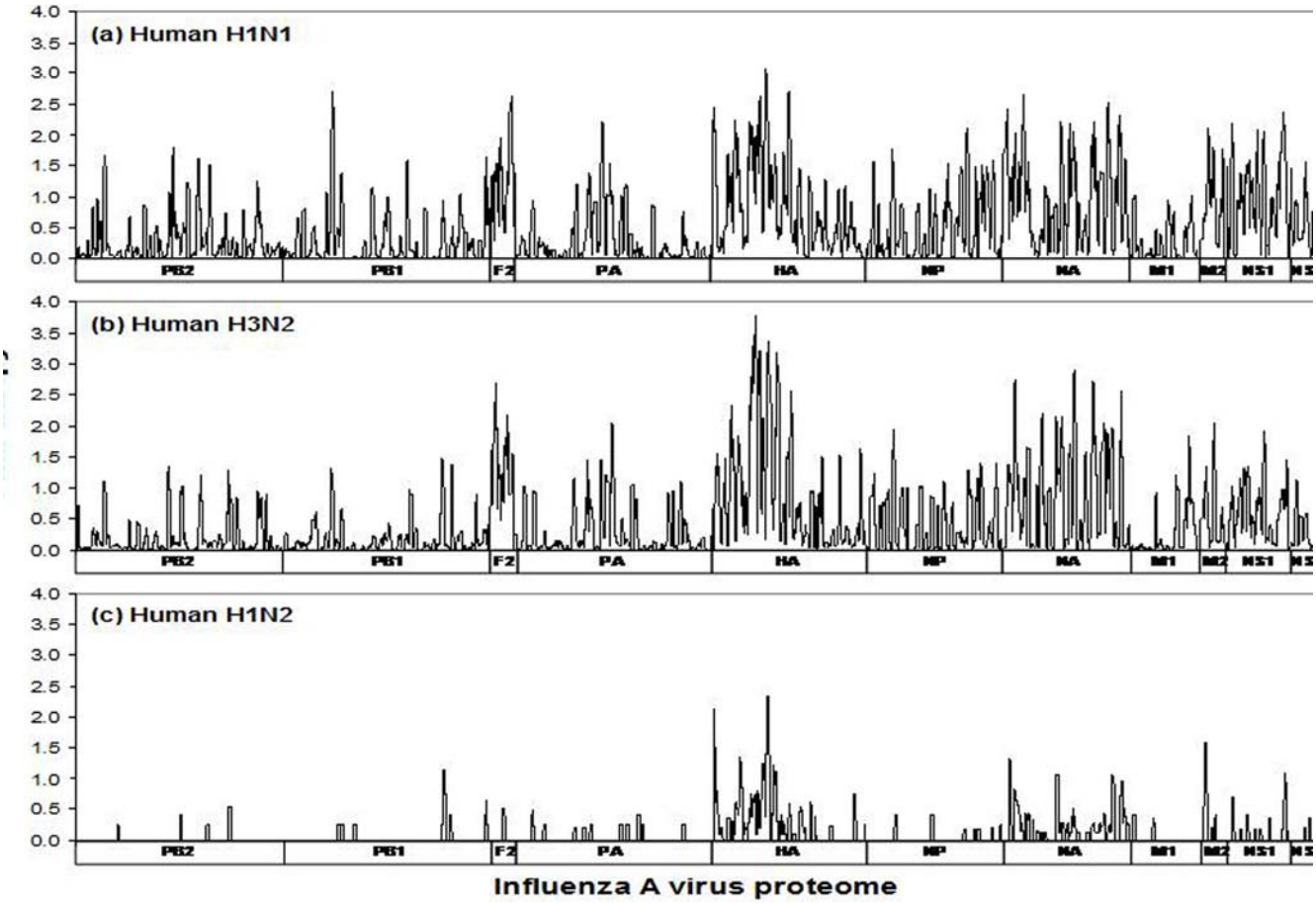


Figure 4. Entropy plots of recorded human influenza A subtypes H1N1, H3N2, and H1N2 from 1918–2006 (data as of September 30, 2006).
doi:10.1371/journal.pone.0001190.g004

The H1N1, H3N2, and H1N2 viruses circulating in humans had the highest representation of conserved sequences, with almost all of the 55 sequences present in 95% to 100% of the isolates of each virus. All but one (22 of 23) of the H1N2 PB2 conserved sequences were identical in each of the virus isolates. By comparison, only 62% to 76% of the conserved sequences of the avian and human H5N1 subgroups, respectively, and only 33% of the conserved sequences of all other avian subtypes were found in 95–100% of the isolates. The greater proportion of conserved sequences in the human isolates can be attributed to the more recent history and limited rate of evolution the influenza viruses transmitted by humans. This is especially true of the human H1N2 virus, the most recent human influenza A virus.

HLA-restricted T-cell epitopes

The association of conserved sequences and T-cell epitopes was examined by (a) *in silico* prediction of HLA-restricted binding sequences corresponding to supertype alleles by TEPITOPE [24], NetCTL [25], MULTIPRED [26] and ARB [27] algorithms; and (b) reported experimental HLA-binding and T-cell assay data. Most of the peptides representing the conserved sequences (50 of 55) were predicted to contain class I and/or class II binding sequences (Figure 7). There was no significant difference in the density of predicted epitopes in the conserved as compared to non-conserved sequences (data not shown). The detailed listing of nonamer sequences of the conserved regions and the predicted supertypes of these specific nonamers in shown as a supplement (Table S2). For example over 500 HLA class I and over 100 class II HLA binding sequences of supertype alleles were predicted, with many of the nonamer sequences predicted to bind to multiple (2 to 9) individual class I alleles. Similarly, all of the DR binding predictions were selected as supertypes on the basis of predicted binding to multiple DR-alleles (individual predictions not shown). The consistency of class I predictions by the different algorithms ranged from 31% to 66% in those supertypes (A1, A2, A3, A24, A26, B7, B44) where more than one computational system was available. The highest consistency of binding sequences cross-predicted by more than one system was observed with A2 (57%), A3 (66%), and DR (56%).

Fourteen (14) of the 55 conserved regions contained a total of 29 reported T-cell epitopes based on T-cell assay and/or HLA-binding data entered into the Immune Epitope Database and Analysis Resource (www.immuneepitope.org/) (Figure 8). These 14 experimentally derived sequences included all of the predicted HLA supertypes of the M1 protein, and 5 of the 11 predicted PB1 supertypes. The majority, 22 of the 29 reported T-cell epitopes, were present as clusters (hotspots) of 2 or more overlapping or closely associated reported epitopes; for example, PB1 518-575 contains 5 epitope sequences (9–10 amino acids) between position 537 and 574. Some of the sequences were promiscuous in their association with multiple supertype alleles, for example, the PA 29-54 sequence containing the nonamer FMYSDFHFI that was experimentally shown to bind to at least 5 class I supertype alleles (A*0201, A*0203, A*0206, A*0202, and A*6802).

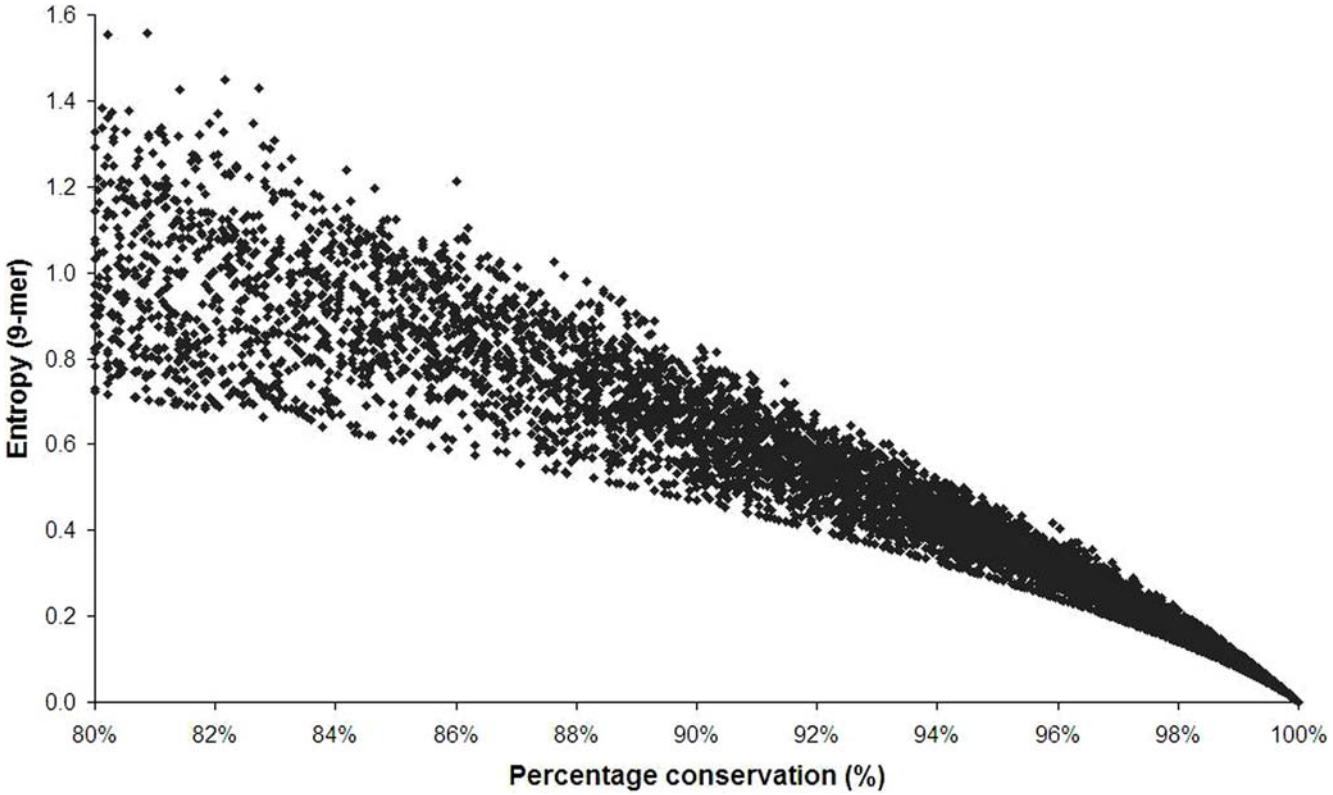


Figure 5. Entropy-sequence conservation relationship, plotted from data in this study (see Figure 2–4). The boxed region indicates area whereby conservation of ≥90% correlates to entropy of 0.8 or less.
doi:10.1371/journal.pone.0001190.g005

Table 3. The influenza A virus proteins, their length, the number of conserved sequences, and the combined length of the conserved sequences of each protein

Protein	Length (aa)[a]	Number of highly conserved sequences[b]	Total length of conserved sequences (aa)[c]
PB2	759	23	379 (50%)
PB1	757	11	271 (36%)
PB1-F2	90	0	0
PA	716	7	111 (16%)
HA	568	1	9 (2%)
NP	498	9	126 (25%)
NA	469	0	0
M1	252	4	69 (27%)
M2	97	0	0
NS1	230	0	0
NS2	121	0	0
Total	4,557	55	965 (21%)[d]

[a]Based on the complete genome sequences of A/Goose/Guangdong/1/96 (H5N1), Taxonomy ID: 93838.
[b]Number of high conserved sequences with sequence and nonamer conservation of ≥80% in influenza A virus sequences from 1997 to 2006 (human H1N1, human H3N2, human H1N2, human H5N1, avian H5N1, and other avian subtypes) in each of the 11 proteins.
[c]The sum of highly conserved sequences length in each of the 11 proteins. The numbers in parentheses indicate the percentage of highly conserved sequences length over the total protein length.
[d]The percentage of total highly conserved sequences length over total influenza A proteome length.
doi:10.1371/journal.pone.0001190.t003

All but one of these 29 unique influenza A HLA epitopes reported in the IEDB and located in the conserved sequences are class I. This HLA distribution differs markedly from the corresponding total IEDB reported influenza A epitopes representing the complete viral proteome, which show a much greater representation, almost 50%, of class II epitopes: 225 class I and 95 class II. Because the conserved sequences represent ~21% of the total proteome, if there were a random distribution of T-cell epitopes in the viral proteins, one could expect about 45 class I and 20 class II epitopes in the conserved sequences, as compared to the observed 28:1. These data are consistent with the conventional model that T-cell epitopes derived from the PB2, PB1, PA, NP, and M1 nonstructural proteins that contain the conserved sequences would be processed primarily in the cytoplasmic proteosomal class I pathway.

DISCUSSION

The marked variability of influenza A virus surface proteins, the major targets of the neutralizing antibodies, have posed a serious obstacle in the development of effective and long-lasting influenza vaccines. As a possible solution, we have identified virus protein sequences that are completely conserved in the majority of all recorded genomic variants that have evolve from avian reservoirs, both avian and human. The information entropy methodology for analysis of protein variability was modified to examine sequences of 9 amino acids or longer, instead of the more common application to single residues, as a means to relate the conserved sequences to the immune function of HLA-restricted peptides. This use of entropy methodology for the identification of highly conserved protein sequences ushers a new experimental strategy in

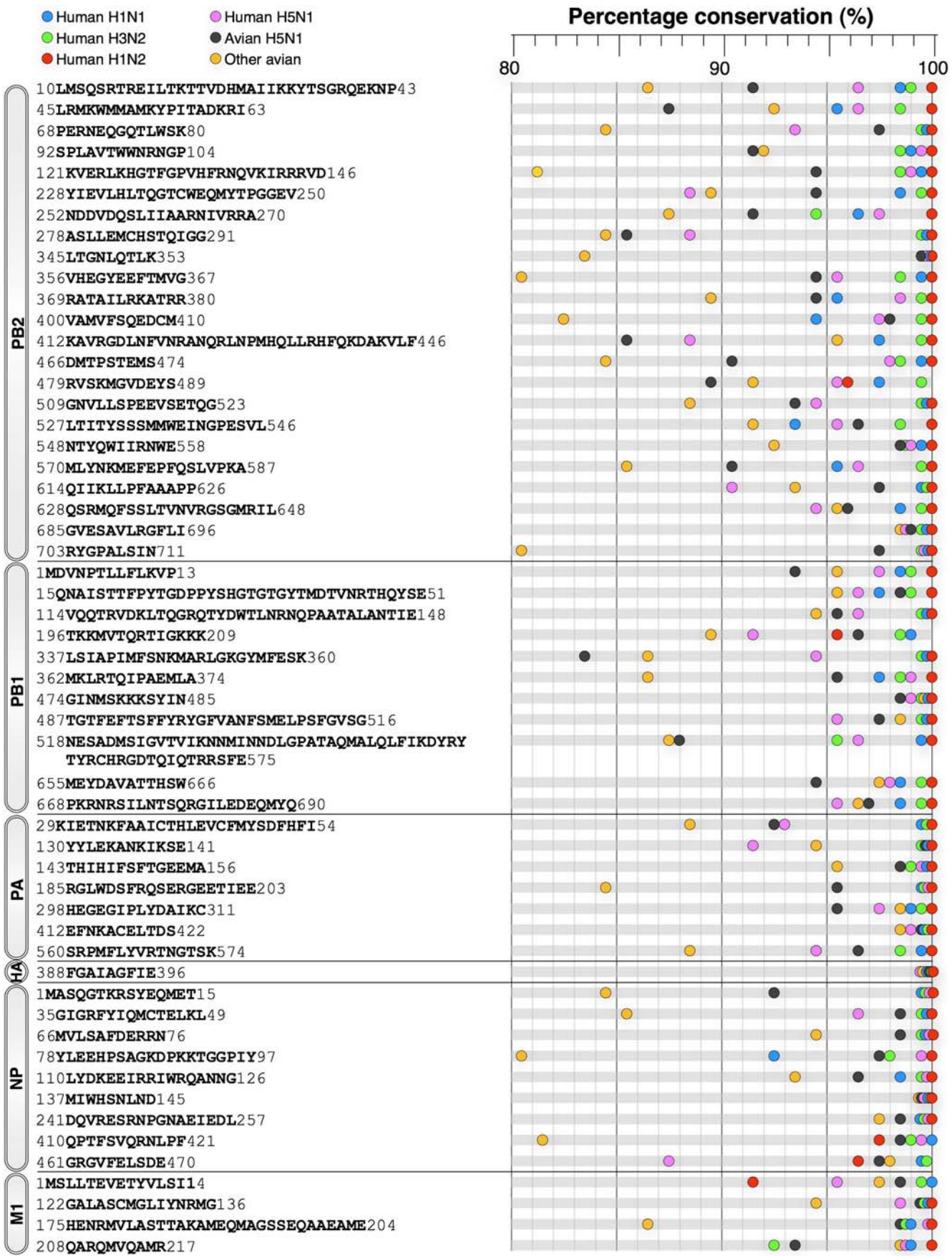


Figure 6. Highly conserved sequences of influenza A viruses in human H1N1, H3N2, H1N2, H5N1, avian H5N1, and other avian subtypes circulating between 1997 and 2006. A region in the viral proteome is considered as highly conserved when it has identical sequence conservation of at least 9 contiguous amino acids in 80% or more of the protein sequences of the analyzed dataset. The index of virus colored symbol is as shown at the top of the figure.
doi:10.1371/journal.pone.0001190.g006

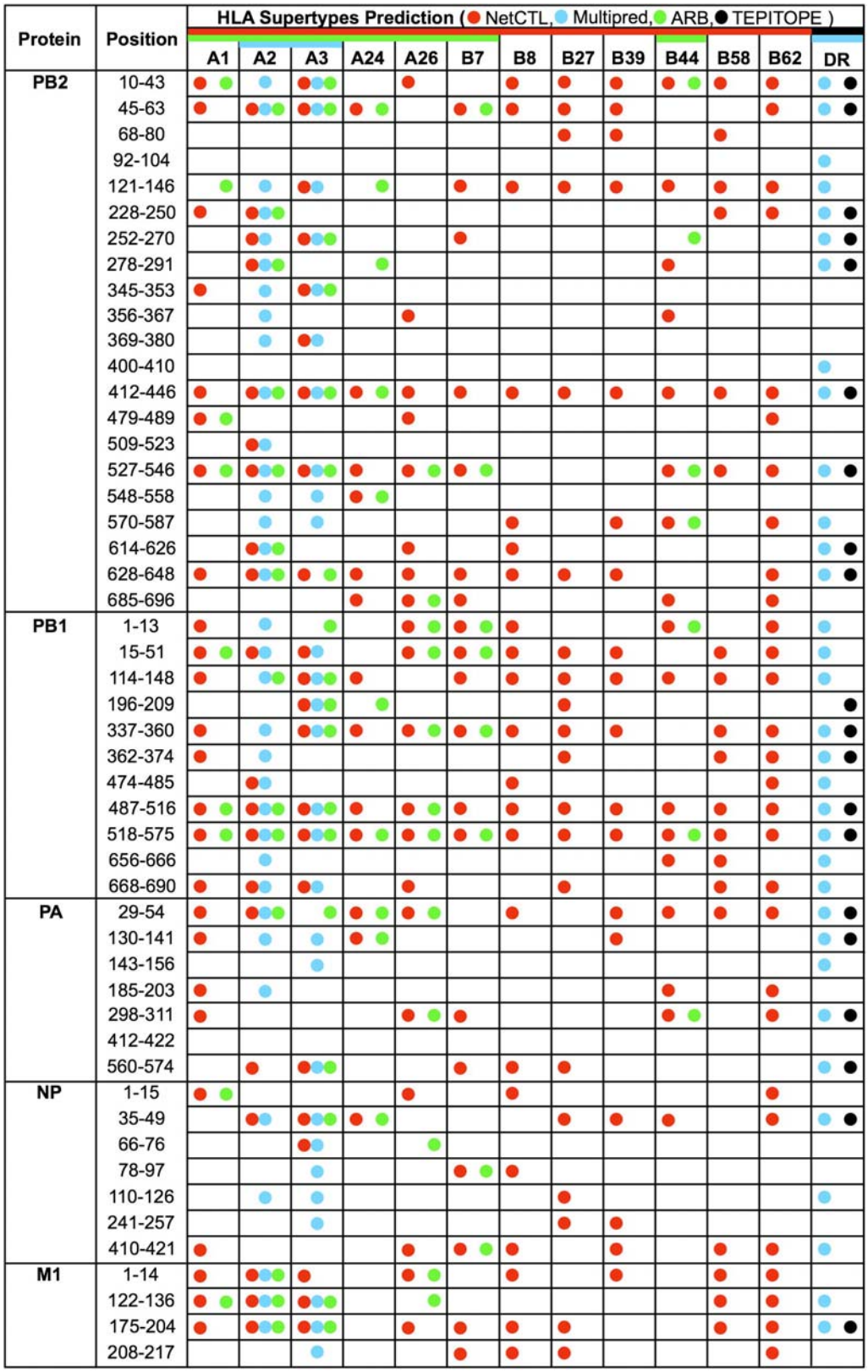


Figure 7. Highly conserved sequences of influenza A viruses and their predicted HLA class I and II supertype-restricted T-cell epitopes by NetCTL, ARB, TEPITOPE, and MULTIPRED systems. The color symbols corresponding to the prediction systems are as shown at the top of the figure. Only conserved sequences containing predicted alleles are shown. NetCTL predicts all of the listed class I supertypes; MULTIPRED predictions cover A2 and A3; and ARB predicts each of the class I except B8, B27, B39, B58, and B62. Predictions of HLA class II supertypes by MULTIPRED AND TEPITOPE is described in Materials and Methods.
doi:10.1371/journal.pone.0001190.g007

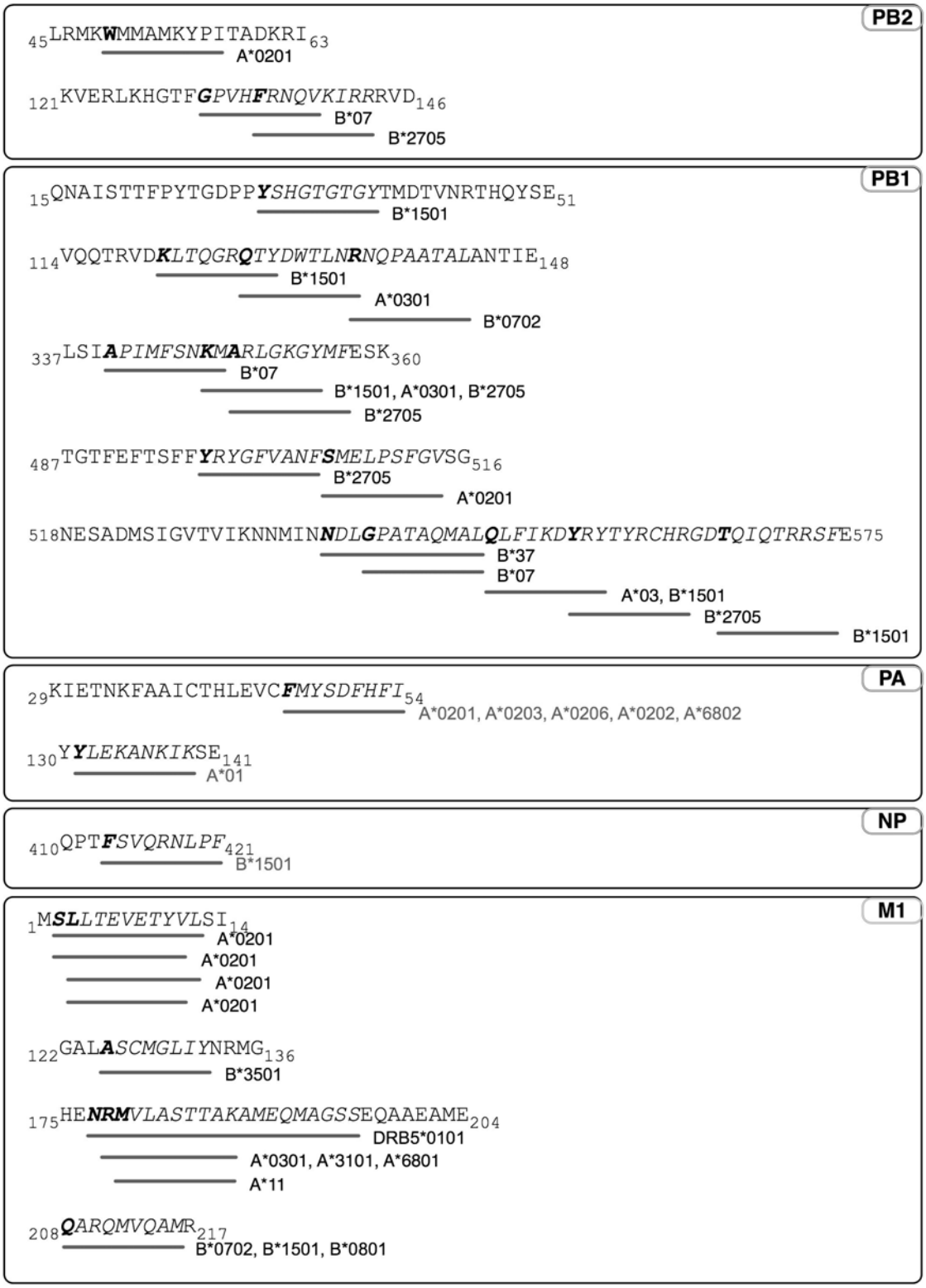


Figure 8. Highly conserved sequences of influenza A viruses and their associated HLA-restricted T-cell epitope based on data obtained from IEDB (www.immuneepitope.org/). Only sequences with identified sites are included. The first amino acid of each identified allele is shown in bold.
doi:10.1371/journal.pone.0001190.g008

the development of vaccines for pathogens with high rates of mutation. The comprehensive analysis of conserved sequences may also have other applications to pathogen diagnosis or therapy. These sequences are known or can be presumed to have critical roles in viral survival and thus are choice targets for the development of antiviral agents.

Many reports, particularly with respect to the human immunodeficiency virus type 1 (HIV-1) have described a strategic advantage in the use of computational analysis and conserved sequences for vaccine design [76–83]. Additionally, the analysis of sequence and immunology databases for the relationship between amino acid sequences and CTL epitope distributions indicate a localization of CTL epitopes in conserved regions of proteins [84]. In contrast, the highly variable regions that lacked epitopes showed evidence of past immune escape with an enrichment of amino acids that do not serve as C-terminal anchor residues and a paucity of predicted proteasome processing sites [85–86]. Likewise, the high genetic variability with continually evolving variants of influenza viruses favors sequence modifications at all sites that result in enhanced virus propagation or survival by adaptation to the host cell immune response. Therefore, a vaccine based upon sequences that are naturally highly conserved in all influenza A viruses may greatly restrict the range of possible mutants that could selectively overcome immune suppression. Such a vaccine would have significant strategic advantage provided the sequences have immune function capability, the design of the immunogen is compatible with the requirements for appropriate immune processing and presentation of the protein, and the epitopes have sufficient HLA-representation to cover the global distribution of HLA genotypes. It appears that these requirements can be satisfied given the large number of predicted supertype MHC binding sequences in the conserved regions of the influenza proteins, the experimental reports of T-cell epitopes of the conserved sequences, and our findings of T-cell responses by HLA transgenic mice to almost all conserved sequences of West Nile virus (unpublished data).

A question, however, is why influenza A differs from other pathogens that elicit immune responses to natural infection or vaccination that prevent repeated infection. It is evident that the mechanisms involved in the immune response to influenza A virus infection are in some manner more complex. A discerning report [87] addresses the ecological and immunological determinants of influenza evolution in relation to several of the characteristic features of influenza infection; i.e., the marked replacement of existing strains during a pandemic caused by antigenic shift, the short-lived viral sublineages that characterize influenza A infection and evolution, and the marked seasonality of influenza incidence. A proposed model [86] to address these characteristic features of influenza infection and evolution was that the host immune system responds in a manner that inhibits immediate re-infection but is short-lived with a time scale of weeks to months and is nonspecific to intra- and inter-subtypes. This pattern of short-lived, cross-reactive immunity points to an initial cytotoxic T-lymphocyte (CTL) response that does not persist. We attribute this to the extreme variability of the structural proteins of influenza A viruses, especially that of the HA and NA proteins. Studies of mice and model pathogens suggest that the initial response of naive $CD8^+$ T-cells to antigen requires only a brief stimulation with antigen early in the immune response, in a matter of hours, for the cells to become activated, divide, and differentiate into short lived effector cells [88–90]. This initial activation can occur in the absence of T-cell help, but without the $CD4^+$ response, the quality of the cytotoxic response to antigen challenge after priming gradually decreases and fails to respond effectively to secondary encounters with antigen. Data of several studies indicate that generation of long term $CD8^+$ T-cell immune memory requires the concurrent function of professional antigen presenting cells for class II antigen processing and presentation to $CD4^+$ helper T-cells during the initial antigen priming period [91–93]. It is likely that the major sources of T-cell epitopes, both class I and II, early after influenza infection are those proteins delivered to the immune system by the virus, including the highly variable structural proteins, HA and NA. Thus, this initial response, and the memory T-cells elicited by this response, may lack the highly conserved epitope sequences of the non-structural proteins that would be synthesized at a later stage of infection and, as cytoplasmic proteins, function primarily as endogenous class I epitopes. In this context, it is noteworthy that of the 29 reported influenza T-cell epitopes found in conserved sequences, there was only a single class II epitope, further suggesting that following natural infection, the conserved sequences elicit primarily cytotoxic T-cell responses.

We suggest that a vaccine composed of conserved influenza A virus sequences may provide a memory immunity to non-structural proteins of all viral variants as a means for augmenting the natural response to the virus structural proteins and to provide an enhanced and augmented immune response to any newly emerging avian influenza A pathogen, as well as to the persistence of mutant forms of human transmitted influenza A. This study establishes the identity of all the highly conserved sequences of both human and avian influenza proteomes as the first step in the selection of these sequences for the synthesis of a supertype, epitope-based genetic vaccine.

SUPPORTING INFORMATION

Table S1 Highly conserved sequences of influenza A viruses and their occurrence in each subgroup. [a] Highly conserved sequences refer to sequences with ⩾80% conservation in each of the six groups that were analyzed. [b] The percentage conservation (rounded down as whole numbers) was calculated as the number of sequences that are identical to the highly conserved sequences divided by the total number of sequences in the same position. The numbers in square brackets indicate the total number of unique sequences at the considered position, inclusive of the highly conserved sequences. [c] The total number of human H1N1 sequences ranged from 187 to 242. [d] The total number of human H3N2 sequences ranged from 969 to 1141. [e] The total number of human H1N2 sequences ranged from 24 to 40. [f] The total number of human H5N1 sequences ranged from 82 to 106. [g] The total number of avian H5N1 sequences ranged from 217 to 648. [h] The total number of avian influenza A subtypes sequences ranged from 210 to 633.

Found at: doi:10.1371/journal.pone.0001190.s001 (0.26 MB DOC)

Table S2 Potential HLA-restricted binding sequences in the highly conserved sequences of influenza A virus that are predicted by the NetCTL, ARB, TEPITOPE, and MULTIPRED systems. [a] Highly conserved sequences of influenza A viruses (Figure 4) and nonameric binding sequences predicted by NetCTL, ARB, TEPITOPE, and/or MULTIPRED algorithms. The numbers in parentheses indicate the number of nonameric binding sequences in a highly conserved sequence that was predicted by at least one algorithm. [b] Nonamers that bind to HLA class I were predicted using NetCTL, ARB, and MULTIPRED. NetCTL 1.2 Server predicts for T cell epitopes that bind to 12 MHC I supertypes, by integrating MHC binding, proteasomal C terminal cleavage, and TAP transport efficiency. MULTIPRED predicts for potential HLA supertype-restricted nonameric sequences that bind to two HLA class I (A2 and A3) supertypes. Only sequences that were

predicted by both artificial neural network (ANN) and hidden markov model (HMM) are included. ARB predicts for T-cell epitopes that bind to 30 MHC class I alleles and 12 class II alleles. This study focused on class I alleles that are the most common in each supertype (according to Lund et al., 2004), namely class I A*0101 in A1 supertype, A*0201 in A2 supertype, A*0301 in A3 supertype, A*2402 in A24 supertype, A*2601 in A26 supertype, B*0702 in B7 supertype, B*4402 and B*4403 in B44 supertype. Only sequences, 9aa for class I that were predicted to bind to these common alleles are listed. Nonamers that were predicted to bind in any one of the three systems are listed. [c] Nonamers that bind to HLA class II were predicted using TEPITOPE and MULTIPRED. TEPITOPE predicts for T cell epitopes that bind to 25 MHC II alleles. Only promiscuous nonameric sequences that were predicted to bind to at least 5 alleles by TEPITOPE system were listed and indicated as "DR". MULTIPRED predicts for potential HLA supertype-restricted nonameric sequences that bind to 8 HLA DRB1 alleles. Only sequences that were predicted by both artificial neural network (ANN) and hidden markov model (HMM) are included. Nonamers that were predicted to bind in any one of the two systems are listed.

Found at: doi:10.1371/journal.pone.0001190.s002 (0.55 MB DOC)

ACKNOWLEDGMENTS

The authors thank Dr. P. Nordstrom August for help with the illustrations and valuable suggestions.

Author Contributions

Conceived and designed the experiments: GZ JA AH OM KS AK VB TT. Performed the experiments: GZ AH. Analyzed the data: GZ JA AH OM AK VB TT. Contributed reagents/materials/analysis tools: JA. Wrote the paper: JA AH OM AK VB TT. Other: Designed the study: JA.

REFERENCES

1. De Jong JC, Rimmelzwaan GF, Fouchier RA, Osterhaus AD (2000) Influenza virus: a master of metamorphosis. J Infect 40: 218–228.
2. Treanor J (2004) Weathering the influenza vaccine crisis. N Engl J Med 351: 2037–2040.
3. Kilbourne ED (2006) Influenza pandemics of the 20th century. Emerg Infect Dis 12: 9–14.
4. Ghedin E, Sengamalay NA, Shumway M, Zaborsky J, Feldblyum T, et al. (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature 437: 1162–1166.
5. Obenauer JC, Denson J, Mehta PK, Su X, Mukatira S, et al. (2006) Large-scale sequence analysis of avian influenza isolates. Science 311: 1576–1580.
6. Lambkin R, Novelli P, Oxford J, Gelder C (2004) Human genetics and responses to influenza vaccination: clinical implications. Am J Pharmacogenomics 4: 293–298.
7. Subbarao K, Murphy BR, Fauci AS (2006) Development of effective vaccines against pandemic influenza. Immunity 24: 5–9.
8. Uscher-Pines L, Omer SB, Barnett DJ, Burke TA, Balicer RD (2006) Priority setting for pandemic influenza: an analysis of national preparedness plans. PLoS Med 3: e436.
9. Benacerraf B (1978) A hypothesis to relate the specificity of T lymphocytes and the activity of I region-specific Ir genes in macrophages and B lymphocytes. J Immunol 120: 1809–1812.
10. Townsend AR, Gotch FM, Davey J (1985) Cytotoxic T cells recognize fragments of the influenza nucleoprotein. Cell 42: 457–467.
11. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, et al. (1987) The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature 329: 512–518.
12. Germain RN (1994) MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. Cell 76: 287–299.
13. Cresswell P, Ackerman AL, Giodini A, Peaper DR, Wearsch PA (2005) Mechanisms of MHC class I-restricted antigen processing and cross-presentation. Immunol Rev 207: 145–157.
14. Trombetta ES, Mellman I (2005) Cell biology of antigen processing in vitro and in vivo. Annu Rev Immunol 23: 975–1028.
15. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351: 290–296.
16. van Bleek GM, Nathenson SG (1991) The structure of the antigen-binding groove of major histocompatibility complex class I molecules determines specific selection of self-peptides. Proc Natl Acad Sci U S A 88: 11032–11036.
17. Engelhard VH (1994) Structure of peptides associated with class I and class II MHC molecules. Annu Rev Immunol 12: 181–207.
18. Rammensee HG, Friede T, Stevanoviic S (1995) MHC ligands and peptide motifs: first listing. Immunogenetics 41: 178–228.
19. Hammer J, Valsasnini P, Tolba K, Bolin D, Higelin J, et al. (1993) Promiscuous and allele-specific anchors in HLA-DR-binding peptides. Cell 74: 197–203.
20. Sidney J, del Guercio MF, Southwood S, Engelhard VH, Appella E, et al. (1995) Several HLA alleles share overlapping peptide specificities. J Immunol 154: 247–259.
21. Sette A, Sidney J (1999) Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics 50: 201–212.
22. Lund O, Nielsen M, Kesmir C, Petersen AG, Lundegaard C, et al. (2004) Definition of supertypes for HLA molecules using clustering of specificity matrices. Immunogenetics 55: 797–810.
23. Doytchinova IA, Flower DR (2005) In silico identification of supertypes for class II MHCs. J Immunol 174: 7085–7095.
24. Bian H, Hammer J (2004) Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE. Methods 34: 468–475.
25. Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, et al. (2005) An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. Eur J Immunol 35: 2295–303.
26. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V (2005) MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucleic Acids Res 33(Web Server issue): W172–179.
27. Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, et al. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics 57: 304–314.
28. Wilson CC, McKinney D, Anders M, MaWhinney S, Forster J, et al. (2003) Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1. J Immunol 171: 5611–5623.
29. Sette A, Fikes J (2003) Epitope-based vaccines: an update on epitope identification, vaccine design and delivery. Curr Opin Immunol 15: 461–470.
30. Brusic V, August JT (2004) The changing field of vaccine development in the genomics era. Pharmacogenomics 5: 597–600.
31. Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, et al. (2007) Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nat Med 13: 100–106.
32. Klavinskis LS, Whitton JL, Oldstone MB (1989) Molecularly engineered vaccine which expresses an immunodominant T-cell epitope induces cytotoxic T lymphocytes that confer protection from lethal virus infection. J Virol 63: 4311–4316.
33. Castrucci MR, Hou S, Doherty PC, Kawaoka Y (1994) Protection against lethal lymphocytic choriomeningitis virus (LCMV) infection by immunization of mice with an influenza virus containing an LCMV epitope recognized by cytotoxic T lymphocytes. J Virol 68: 3486–3490.
34. Stemmer C, Quesnel A, Prevost-Blondel A, Zimmermann C, Muller S, et al. (1999) Protection against lymphocytic choriomeningitis virus infection induced by a reduced peptide bond analogue of the H-2Db-restricted CD8(+) T cell epitope GP33. J Biol Chem 274: 5550–5556.
35. Tsuji M, Bergmann CC, Takita-Sonoda Y, Murata K, Rodrigues EG, et al. (1998) Recombinant Sindbis viruses expressing a cytotoxic T-lymphocyte epitope of a malaria parasite or of influenza virus elicit protection against the corresponding pathogen in mice. J Virol 72: 6907–6910.
36. Gonzalez-Aseguinolaza G, Nakaya Y, Molano A, Dy E, Esteban M, et al. (2003) Induction of protective immunity against malaria by priming-boosting immunization with recombinant cold-adapted influenza and modified vaccinia Ankara viruses expressing a $^{CD8=}$-T-cell epitope derived from the circumsporozoite protein of Plasmodium yoelii. J Virol 77: 11859–11866.
37. Del Val M, Schlicht HJ, Volkmer H, Messerle M, Reddehase MJ, et al. (1991) Protection against lethal cytomegalovirus infection by a recombinant vaccine containing a single nonameric T-cell epitope. J Virol 65: 3641–3646.
38. La Posta VJ, Auperin DD, Kamin-Lewis R, Cole GA (1993) Cross-protection against lymphocytic choriomeningitis virus mediated by a CD4+ T-cell clone specific for an envelope glycoprotein epitope of Lassa virus. J Virol 67: 3497–3506.
39. Oukka M, Manuguerra JC, Livaditis N, Tourdot S, Riche N, et al. (1996) Protection against lethal viral infection by vaccination with nonimmunodominant peptides. J Immunol 157: 3039–3045.
40. Blaney JE Jr, Nobusawa E, Brehm MA, Bonneau RH, Mylin LM, et al. (1998) Immunization with a single major histocompatibility complex class I-restricted cytotoxic T-lymphocyte recognition epitope of herpes simplex virus type 2 confers protective immunity. J Virol 72: 9567–9574.

41. Feltkamp MC, Vreugdenhil GR, Vierboom MP, Ras E, van der Burg SH, et al. (1995) Cytotoxic T lymphocytes raised against a subdominant epitope offered as a synthetic peptide eradicate human papillomavirus type 16-induced tumors. Eur J Immunol 25: 2638–2642.
42. Harty JT, Bevan MJ (1992) $^{CD8=}$ T cells specific for a single nonamer epitope of Listeria monocytogenes are protective in vivo. J Exp Med 175: 1531–1538.
43. Plotnicky H, Cyblat-Chanal D, Aubry JP, Derouet F, Klinguer-Hamour C, et al. (2003) The immunodominant influenza matrix T cell epitope recognized in human induces influenza protection in HLA-A2/K(b) transgenic mice. Virology 309: 320–329.
44. Snyder JT, Belyakov IM, Dzutsev A, Lemonnier F, Berzofsky JA (2004) Protection against lethal vaccinia virus challenge in HLA-A2 transgenic mice by immunization with a single $^{CD8=}$ T-cell peptide epitope of vaccinia and variola viruses. J Virol 78: 7052–7060.
45. Botten J, Whitton JL, Barrowman P, Sidney J, Whitmire JK, et al. (2007) HLA-A2-restricted protection against lethal lymphocytic choriomeningitis. J Virol 81: 2307–2317.
46. Hanke T, McMichael AJ, Dorrell L (2007) Clinical experience with plasmid DNA- and modified vaccinia virus Ankara-vectored human immunodeficiency virus type 1 clade A vaccine focusing on T-cell induction. J Gen Virol 88: 1–12.
47. Moorthy VS, Imoukhuede EB, Keating S, Pinder M, Webster D, et al. (2004) Phase 1 evaluation of 3 highly immunogenic prime-boost regimens, including a 12-month reboosting vaccination, for malaria vaccination in Gambian men. J Infect Dis 189: 2213–2219.
48. McMichael AJ, Gotch FM (1989) Recognition of influenza A virus by human cytotoxic T lymphocytes. Adv Exp Med Biol 257: 109–114.
49. Townsend AR (1987) Recognition of influenza virus proteins by cytotoxic T lymphocytes. Immunol Res 6: 80–100.
50. Askonas BA, Taylor PM, Esquivel F (1988) Cytotoxic T cells in influenza infection. Ann N Y Acad Sci 532: 230–237.
51. Lamb JR, McMichael AJ, Rothbard JB (1987) T-cell recognition of influenza viral antigens. Hum Immunol 19: 79–89.
52. Swain SL, Agrewala JN, Brown DM, Jelley-Gibbs DM, Golech S, et al. (2006) CD4+ T-cell memory: generation and multi-faceted roles for CD4+ T cells in protective immunity to influenza. Immunol Rev 211: 8–22.
53. Thomas PG, Keating R, Hulse-Post DJ, Doherty PC (2006) Cell-mediated protection in influenza infection. Emerg Infect Dis 12: 48–54.
54. Ulmer JB, Donnelly JJ, Parker SE, Rhodes GH, Felgner PL, et al. (1993) Heterologous protection against influenza by injection of DNA encoding a viral protein. Science 259: 1745–1749.
55. Ulmer JB, Fu TM, Deck RR, Friedman A, Guan L, et al. (1998) Protective CD4+ and $^{CD8=}$ T cells against influenza virus induced by vaccination with nucleoprotein DNA. J Virol 72: 5648–5653.
56. Fu TM, Guan L, Friedman A, Schofield TL, Ulmer JB, et al. (1999) Dose dependence of CTL precursor frequency induced by a DNA vaccine and correlation with protective immunity against influenza virus challenge. J Immunol 162: 4163–4170.
57. Epstein SL, Tumpey TM, Misplon JA, Lo CY, Cooper LA, et al. (2002) DNA vaccine expressing conserved influenza virus proteins protective against H5N1 challenge infection in mice. Emerg Infect Dis 8: 796–801.
58. Epstein SL, Kong WP, Misplon JA, Lo CY, Tumpey TM, et al. (2005) Protection against multiple influenza A subtypes by vaccination with highly conserved nucleoprotein. Vaccine 23: 5404–5410.
59. Fomsgaard A, Nielsen HV, Kirkby N, Bryder K, Corbet S, et al. (1999) Induction of cytotoxic T-cell responses by gene gun DNA vaccination with minigenes encoding influenza A virus HA and NP CTL-epitopes. Vaccine 18: 681–691.
60. Lawson CM, Bennink JR, Restifo NP, Yewdell JW, Murphy BR (1994) Primary pulmonary cytotoxic T lymphocytes induced by immunization with a vaccinia virus recombinant expressing influenza A virus nucleoprotein peptide do not protect mice against challenge. J Virol 68: 3505–3511.
61. Moskophidis D, Kioussis D (1998) Contribution of virus-specific $^{CD8=}$ cytotoxic T cells to virus clearance or pathologic manifestations of influenza virus infection in a T cell receptor transgenic mouse model. J Exp Med 188: 223–232.
62. Crowe SR, Miller SC, Woodland DL (2006) Identification of protective and non-protective T cell epitopes in influenza. Vaccine 24: 452–456.
63. Khan AM, Miotto O, Heiny AT, Salmon J, Srinivasan KN, et al. (2007) A systematic bioinformatics approach for selection of epitope-based vaccine targets. Cell Immunol 244: 141–147.
64. Miotto O, Tan TW, Brusic V (2007) Rule-based Knowledge Aggregation for Large-Scale Protein Sequence Analysis of Influenza A Viruses. BMC Bioinformatics, 8 Suppl 10: S7.
65. Cox NJ, Neumann G, Donis RO, Kawaoka Y (2005) Orthomyxoviruses: influenza In: Mahy BH, ter Meulen V, eds. Topley and Wilson's Microbiology and Microbial Infections, 10th Edition, Virology Volume 1, Chapter 32.
66. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
67. Fouchier RA, Munster V, Wallensten A, Bestebroer TM, Herfst S, et al. (2005) Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. J Virol 79: 2814–2822.
68. Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27: 379–423 and 623–656.
69. Rammensee HG (1995) Chemistry of peptides associated with MHC class I and class II molecules. Curr Opin Immunol 7: 85–96.
70. Paninski L (2003) Estimation of entropy and mutual information. Neural Computation 15: 1191–1253.
71. Peters B, Sidney J, Bourne P, Bui HH, Buus S, et al. (2005) The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol 3: e91.
72. Peters B, Bui HH, Frankild S, Nielson M, Lundegaard C, et al. (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput Biol 2: e65.
73. Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, et al. (2001) A novel influenza A virus mitochondrial protein that induces cell death. Nat Med. 7(12): 1306–1312.
74. Coleman JR (2007) The PB1-F2 protein of Influenza A virus: increasing pathogenicity by disrupting alveolar macrophages. Virol J. 4: 9.
75. Miotto O, Heiny AT, Tan TW, August JT, Brusic V (2007) Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis. BMC Bioinformatics, 8 Suppl 10: S18.
76. Mazumder R, Hu ZZ, Vinayaka CR, Sagripanti JL, Frost SD, et al. (2007) Computational analysis and identification of amino acid sites in dengue E proteins relevant to development of diagnostics and vaccines. Virus Genes. 35: 175–186.
77. Nickle DC, Rolland M, Jensen MA, Pond SL, Deng W, et al. (2007) Coping with viral diversity in HIV vaccine design. PLoS Comput Biol. 3: e751.
78. Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, et al. (2007) Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nat. Med. 13: 100–106.
79. De Groot AS, Marcon L, Bishop EA, Rivera D, Kutzler M, et al. (2005) HIV vaccine development by computer assisted design: the GAIA vaccine. Vaccine. 23: 2136–2148.
80. Wilson CC, McKinney D, Anders M, MaWhinney S, Forster J, et al. (2003) Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1. J Immunol 171: 5611–5623.
81. Corbet S, Nielsen HV, Vinner L, Lauemoller S, Therrien D, et al. (2003) Optimization and immune recognition of multiple novel conserved HLA-A2, human immunodeficiency virus type 1-specific CTL epitopes. The Journal of general virology 84: 2409–2421.
82. Wilson CC, McKinney D, Anders M, MaWhinney S, Forster J, et al. (2003) Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1. Journal of immunology (Baltimore, Md : 1950) 171: 5611–5623.
83. Thakar MR, Bhonge LS, Lakhashe SK, Shankarkumar U, Sane SS, et al. (2005) Cytolytic T lymphocytes (CTLs) from HIV-1 subtype C-infected Indian patients recognize CTL epitopes from a conserved immunodominant region of HIV-1 Gag and Nef. The Journal of infectious diseases 192: 749–759.
84. Yusim K, Kesmir C, Gaschen B, Addo MM, Altfeld M, et al. (2002) Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. Journal of virology 76: 8757–8768.
85. Allen TM, Altfeld M, Yu XG, O'Sullivan KM, Lichterfeld M, et al. (2004) Selection, transmission, and reversion of an antigen-processing cytotoxic T-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. Journal of virology 78: 7069–7078.
86. Yokomaku Y, Miura H, Tomiyama H, Kawana-Tachikawa A, Takiguchi M, et al. (2004) Impaired processing and presentation of cytotoxic-T-lymphocyte (CTL) epitopes are major escape mechanisms from CTL immune pressure in human immunodeficiency virus type 1 infection. Journal of virology 78: 1324–1332.
87. Ferguson NM, Galvani AP, Bush RM (2003) Ecological and immunological determinants of influenza evolution. Nature 422: 428–433.
88. van Stipdonk MJ, Lemmens EE, Schoenberger SP (2001) Naive CTLs require a single brief period of antigenic stimulation for clonal expansion and differentiation. Nat Immunol 2: 423–429.
89. Sun JC, Williams MA, Bevan MJ (2004) CD4+ T cells are required for the maintenance, not programming, of memory $^{CD8=}$ T cells after acute infection. Nature immunology 5: 927–933.
90. Prlic M, Hernandez-Hoyos G, Bevan MJ (2006) Duration of the initial TCR stimulus controls the magnitude but not functionality of the $^{CD8=}$ T cell response. The Journal of experimental medicine 203: 2135–2143.
91. Sun JC, Bevan MJ (2003) Defective CD8 T cell memory following acute infection without CD4 T cell help. Science (New York, NY) 300: 339–342.
92. Janssen EM, Lemmens EE, Wolfe T, Christen U, von Herrath MG, et al. (2003) CD4+ T cells are required for secondary expansion and memory in CD8= T lymphocytes. Nature 421: 852–856.
93. Shedlock DJ, Shen H (2003) Requirement for CD4 T cell help in generating functional CD8 T cell memory. Science (New York, NY) 300: 337–339.

Appendix C – Reprint of Khan *et al.* (2008)

Khan AM, Miotto O, Nascimento EJM, Srinivasan KN, Heiny AT, Zhang GL, Salmon J, Marques ET, Tan TW, Brusic V, August JT (2008)

Conservation and Variability of Dengue Virus Proteins: Implications for Vaccine Design.

PLoS Neglected Tropical Diseases. 2(8), e272.

OPEN ACCESS Freely available online

PLoS NEGLECTED TROPICAL DISEASES

Conservation and Variability of Dengue Virus Proteins: Implications for Vaccine Design

Asif M. Khan[1], Olivo Miotto[1,2], Eduardo J. M. Nascimento[3], K. N. Srinivasan[4,5], A. T. Heiny[1], Guang Lan Zhang[6], E. T. Marques[3,4], Tin Wee Tan[1], Vladimir Brusic[6], Jerome Salmon[4], J. Thomas August[4]*

1 Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, **2** Institute of Systems Science, National University of Singapore, Singapore, **3** Department of Medicine, Division of Infectious Diseases, The Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **4** Department of Pharmacology and Molecular Sciences, The Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **5** Product Evaluation and Registration Division, Centre for Drug Administration, Health Sciences Authority, Singapore, **6** Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America

Abstract

Background: Genetic variation and rapid evolution are hallmarks of RNA viruses, the result of high mutation rates in RNA replication and selection of mutants that enhance viral adaptation, including the escape from host immune responses. Variability is uneven across the genome because mutations resulting in a deleterious effect on viral fitness are restricted. RNA viruses are thus marked by protein sites permissive to multiple mutations and sites critical to viral structure-function that are evolutionarily robust and highly conserved. Identification and characterization of the historical dynamics of the conserved sites have relevance to multiple applications, including potential targets for diagnosis, and prophylactic and therapeutic purposes.

Methodology/Principal Findings: We describe a large-scale identification and analysis of evolutionarily highly conserved amino acid sequences of the entire dengue virus (DENV) proteome, with a focus on sequences of 9 amino acids or more, and thus immune-relevant as potential T-cell determinants. DENV protein sequence data were collected from the NCBI Entrez protein database in 2005 (9,512 sequences) and again in 2007 (12,404 sequences). Forty-four (44) sequences (pan-DENV sequences), mainly those of nonstructural proteins and representing ~15% of the DENV polyprotein length, were identical in 80% or more of all recorded DENV sequences. Of these 44 sequences, 34 (~77%) were present in ≥95% of sequences of each DENV type, and 27 (~61%) were conserved in other *Flaviviruses*. The frequencies of variants of the pan-DENV sequences were low (0 to ~5%), as compared to variant frequencies of ~60 to ~85% in the non pan-DENV sequence regions. We further showed that the majority of the conserved sequences were immunologically relevant: 34 contained numerous predicted human leukocyte antigen (HLA) supertype-restricted peptide sequences, and 26 contained T-cell determinants identified by studies with HLA-transgenic mice and/or reported to be immunogenic in humans.

Conclusions/Significance: Forty-four (44) pan-DENV sequences of at least 9 amino acids were highly conserved and identical in 80% or more of all recorded DENV sequences, and the majority were found to be immune-relevant by their correspondence to known or putative HLA-restricted T-cell determinants. The conservation of these sequences through the entire recorded DENV genetic history supports their possible value for diagnosis, prophylactic and/or therapeutic applications. The combination of bioinformatics and experimental approaches applied herein provides a framework for large-scale and systematic analysis of conserved and variable sequences of other pathogens, in particular, for rapidly mutating viruses, such as influenza A virus and HIV.

Citation: Khan AM, Miotto O, Nascimento EJM, Srinivasan KN, Heiny AT, et al. (2008) Conservation and Variability of Dengue Virus Proteins: Implications for Vaccine Design. PLoS Negl Trop Dis 2(8): e272. doi:10.1371/journal.pntd.0000272

Editor: Eva Harris, University of California Berkeley, United States of America

Received February 5, 2008; **Accepted** July 10, 2008; **Published** August 13, 2008

Funding: This project has been funded in part with the Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, USA, under Grant No. 5 U19 AI56541. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: taugust@jhmi.edu

Introduction

Dengue viruses (DENVs) are mosquito-borne pathogens of the family *Flaviviridae*, genus *Flavivirus*, which are phylogenetically related to other important human pathogens, such as *Yellow fever* (YFV), *Japanese encephalitis* (JEV), and *West Nile* (WNV) viruses, among others. DENVs are enveloped, single-stranded RNA (+) viruses coding for a polyprotein precursor of approximately 3,400 amino acids, which is cleaved into three structural (capsid, C; precursor membrane and membrane, prM/M; envelope, E) and seven nonstructural proteins (NS1, 2a, 2b, 3, 4a, 4b and 5). Viral replication occurs in the cytoplasm in association with virus-induced membrane structures and involves the NS proteins. There are 4 genetically distinct DENV types, referred to as DENV-1 to -4, with multiple genotypic variants [1,2]. DENVs are transmitted to humans primarily by *Aedes aegypti* mosquitoes and cause a wide range of symptoms from an unapparent or mild dengue fever (DF) to severe dengue hemorrhagic fever (DHF)/dengue shock

Author Summary

Dengue viruses (DENVs) circulate in nature as a population of 4 distinct types, each with multiple genotypes and variants, and represent an increasing global public health issue with no prophylactic and therapeutic formulations currently available. Viral genomes contain sites that are evolutionarily stable and therefore highly conserved, presumably because changes in these sites have deleterious effects on viral fitness and survival. The identification and characterization of the historical dynamics of these sites in DENV have relevance to several applications such as diagnosis and drug and vaccine development. In this study, we have identified sequence fragments that were conserved across the majority of available DENV sequences, analyzed their historical dynamics, and evaluated their relevance as candidate vaccine targets, using various bioinformatics-based methods and immune assay in human leukocyte antigen (HLA) transgenic mice. This approach provides a framework for large-scale and systematic analysis of other human pathogens.

syndrome (DSS) that may be fatal. It is estimated that more than 100 million people are infected each year, with up to several hundred thousand DHF/DSS cases [3]. To date, there is no licensed prophylactic vaccine and no specific therapeutic formulation available.

Adaptive immune responses include cellular responses to short peptides derived from self and foreign proteins by proteolysis. The peptides are presented to T-cell receptors (TCRs) by major histocompatibility complex (MHC) molecules, referred to as human leukocyte antigen (HLA) molecules in humans. HLA class I and class II molecules bind and present peptides to CD8 and CD4 T-cells, respectively, that play a critical role in antigen (Ag)-specific cytotoxic responses and the induction and maintenance of Ag-specific memory responses [4–6]. Peptides that are recognized by the T cells and trigger an immune response are referred to as T-cell determinants. One problem in developing a tetravalent DENV vaccine is the viral diversity [7], with rather low intra-type, but high inter-type variability, resulting in type-specific and type cross-reactive T-cell determinants [8]. This variability of related structures gives rise to a large number of variant peptide sequences with one or more amino acid differences that may function as alternative determinants, or altered peptide ligands [9], and affect anti-DENV host immunity [10,11]. There is abundant evidence that interactions of memory T cells with peptide ligands bearing amino acid substitutions at TCR contact residues may alter T-cell activation and effector function [9,12–15]. Even a single amino acid substitution can impair the function of T cells in a variety of ways, producing profoundly different phenotypes that range from modified stimulatory function to complete inhibition [14]. These findings suggest that infection or immunization with multiple DENV types, as is the case with some tetravalent vaccines, may lead to T-cell responses to variant peptides that might be deleterious. There is also the possibility that the altered-ligand phenomenon and cross-reactive T-cell responses, referred to as original antigenic sin, may play a role in DHF/DSS [7,11,16,17]. Although the etiology of DHF and DSS is only partially understood, this consideration may have profound implications for the safety and efficiency of candidate vaccines.

The objective of this study was to search for sequence regions conserved across the majority of DENVs and representing potential immune targets [18]. Bioinformatics-based approaches were used to (a) extract all DENV sequences available in public databases, (b) identify and examine the structure-function relationship and distribution in nature of sequences that are highly conserved in the majority of DENVs (referred to as pan-DENV sequences), (c) analyze the variability of DENV sequences, and (d) examine the immune relevance of the conserved sequences as potential T-cell determinants that would be applicable to the majority of the human population worldwide [19]. We have also correlated the conserved DENV sequences to previously reported T-cell determinants and further identified novel candidate T-cell determinants by analyzing HLA-restricted immune responses in HLA transgenic mice.

Methods

Methodology overview

The bioinformatics approaches and rationale for the methodology adopted in this study have been previously described [20] and are summarized in **Figure 1**.

Data collection and sequence organization

DENV protein sequences were retrieved from the NCBI Entrez protein database in December 2005, and again in December 2007 for validation purposes, by use of a taxonomy ID search via the NCBI taxonomy browser [21]. The taxonomy IDs for DENV-1 to -4 were 11053, 11060, 11069 and 11070, respectively. The data for 2007 were processed separately from the 2005 dataset, but using identical procedures.

The sequences of the DENV proteins C, prM, E, NS1, NS2a, NS2b, NS3, NS4a, NS4b and NS5 were extracted from the database records (**Dataset S1**) by multiple sequence alignments, and application of the known cleavage sites obtained from the annotation of the GenPept [21] reference polyprotein sequences of DENV-1 to -4 (AAF59976, P14340, AAM51537, AAG45437, respectively), and from the literature [22]. Grouping of the sequences of each DENV type was performed by BLAST [23] followed by CLUSTALX 1.83 [24] multiple sequence alignments. Both full-length and partial sequences of each DENV protein were used for analysis, and identical sequences were not removed from datasets, unless otherwise indicated. All multiple sequence alignments were manually inspected and corrected for misalignments.

Identification of pan-DENV sequences

The DENV protein sequences were examined by a consensus-sequence based approach [25] to identify sequence fragments that were common across the 4 types. The consensus sequences for the proteins of each type (intra-type consensus) were first derived by multiple sequence alignments to select the predominant residue at each amino acid position. The 4 intra-type consensus sequences for a given protein (one from each type) were then aligned to reveal sequence fragments identical across each of the types that were at least 9 amino acids long. This minimum length was chosen because it represents the binding core length of a majority of HLA-restricted T-cell determinants [26]. Only sequence fragments that were identical in at least 80% of the sequences of each of the 4 types were retained for further analyses. Peptides with residue X in the alignment were ignored from the percentage representation (*i.e.* frequency) computation. The 80% intra-type representation cut-off was chosen because 44 of the 46 sequence fragments that were common across the 4 DENV types exhibited intra-type representation of $\geq$81%, and those two that did not had significantly lower representation ($\sim$56–67%) in one of the 4 types.

Information entropy analysis of pan-DENV sequences

Shannon information entropy [20,27] was used to study the diversity of DENV protein sequences within each type (intra-type

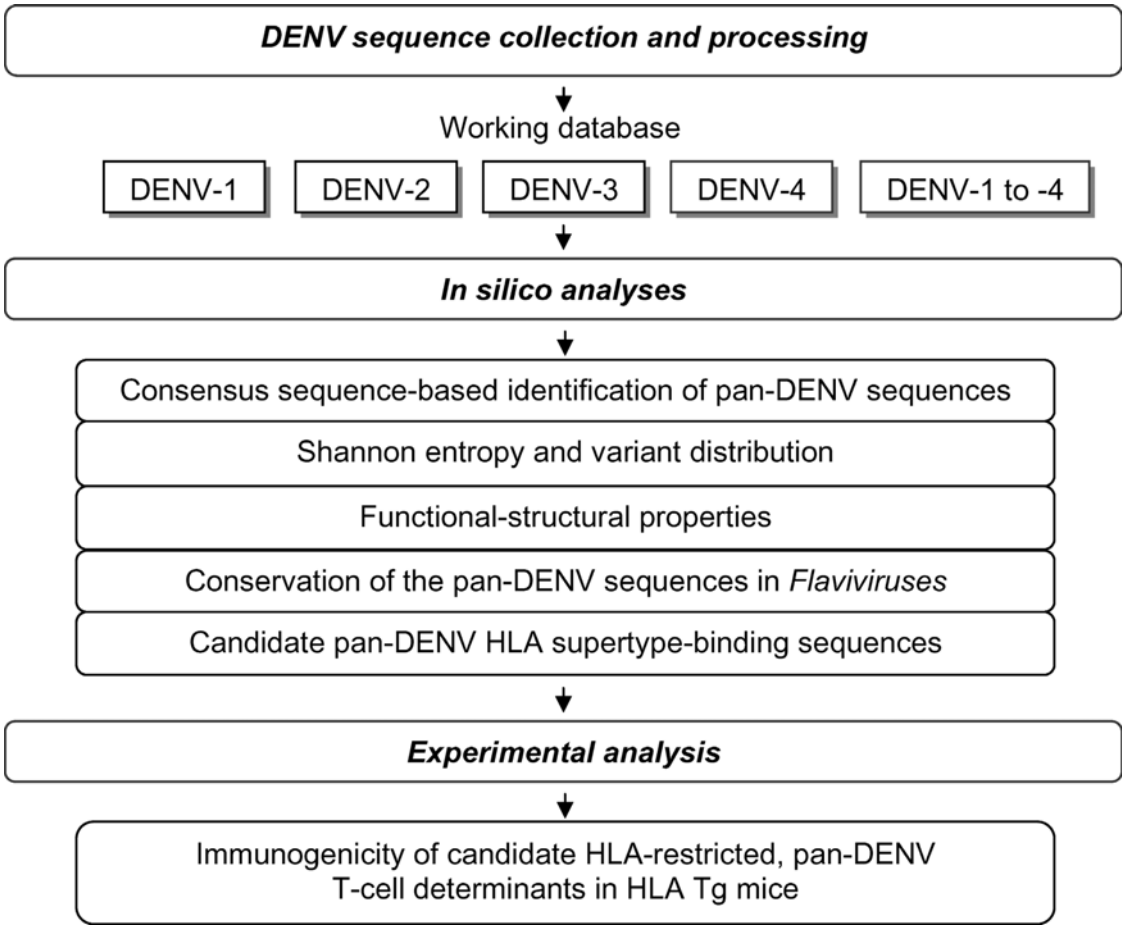


Figure 1. Overview of the bioinformatics and experimental approaches employed for the identification and analysis of the pan-DENV sequences.
doi:10.1371/journal.pntd.0000272.g001

diversity) and across all DENVs (pan-DENV diversity) and to assess the predicted evolutionary stability of the identified pan-DENV sequences. All entropy analyses were carried out by using the in-house developed Antigenic Variability Analyser tool (AVANA) [28]. For immunological applications, the entropy measure for antigenic sequences was based on nonamer peptides [26], centered at any given position in the alignment. Applying Shannon's formula, the nonamer peptide entropy *H(x)* at any given position x in the alignment was computed by

$$H(x) = -\sum_{i=1}^{n(x)} p(i, x) \log_2 p(i, x)$$

where *p(i, x)* is the probability of a particular nonamer peptide *i* being centered at position *x*. The entropy value increases with *n(x)*, the total number of peptides observed at position *x*; it is also sensitive to the relative frequency of the peptides; such that it decreases when one peptide is clearly dominant (*i.e.* the position is conserved). Only sequences that contain a valid amino acid at position *x* were used for the entropy computation, and the alignment gaps were ignored. Although gaps tend to occur in high-diversity regions, proteins that have a high fraction of gaps have reduced statistical support, yielding an artificially low entropy value; for this reason, positions where more than 50% of sequences contained a gap were discarded. Because of the statistical nature of the entropy measure, both complete protein and shorter fragment sequences were used in this computation. The first and last 4 positions in the alignment of each protein were not assigned any peptide entropy value as they cannot be the center of a nonamer.

In theory, nonamer entropy values can range from 0, for a completely conserved nonamer peptide in all sequences analyzed, to 39 ($\log_2 20^9$); in practice, however, the upper bound is very much lower for alignments of closely related sequences. For finite-size sets of sequences, entropy computations are affected by the sequence count in the alignment. For an alignment of N sequences, alignment size bias is proportional to 1/N [29]. This relationship allows a correction for size bias by applying to each alignment a statistical adjustment that estimates entropy values for an infinitely-sized alignment with analogous peptide distribution. To obtain such an estimate, the alignment was repeatedly randomly sampled to create smaller alignments of varying size, whose entropy was measured. At each alignment position, the entropy of these subset alignments of size N was plotted against 1/N, using a linear regression to extrapolate the entropy estimate for $N \rightarrow \infty$. The regression's coefficient of determination (r^2) was used as a goodness-of-fit of the resulting estimate. In this study, size bias correction was applied to all entropy calculations, so that alignment sequence counts could be ignored in comparisons. All entropy values reported are therefore infinite-size set estimates, rather than the values directly computed from the alignments.

Nonamer variant analysis of pan-DENV sequences

Data from information entropy analysis were used to study the distribution of the representation of nonamer variant peptides in DENV sequences, within and across the types. For any given position x in the alignment, the combined representation of all nonamers, excluding the predominant peptide, was computed. The predominant nonamer was the peptide that was contained in the majority of the sequences at the position in the alignment. All the other peptides that differed by at least one amino acid from the predominant nonamer were defined as variants.

Functional and structural analyses of pan-DENV sequences

The known and putative structural and functional properties of pan-DENV sequences were searched in the literature and by use of the Prosite [30], via ScanProsite [31], and Pfam [32] databases. When possible, the sequences were mapped on the three-dimensional (3-D) structures of available DENV Ag in the PDB database [33] by use of ICM-Browser version 3.3 (www.molsoft.com). X-ray diffraction 3-D structures were visualized by use of the Corey, Pauling and Koltun (cpk) representation in the ICM-Browser.

Identification of pan-DENV sequences common to other viruses and organisms

Pan-DENV sequences that overlapped at least 9 consecutive amino acid sequences of other viruses and organisms were identified by performing BLAST search against viral protein sequences reported at NCBI (as of July 2007), excluding DENV sequences (parameters set: limit by Entrez query "txid10239[Organism:exp] NOT txid12637[Organism:exp]"; automatically adjust parameters for short sequences option enabled; low-complexity filter disabled; alignments: 20,000), and against protein sequences of all organisms excluding viruses (parameters set: limit by Entrez query "Root[ORGN] NOT Viruses[ORGN] NOT txid81077[ORGN]"; automatically adjust parameters for short sequences option enabled; low-complexity filter disabled; alignments: 20,000). The keyword "NOT txid81077 [ORGN]" was used to remove artificial sequence hits.

Identification of known and predicted pan-DENV HLA supertype binding sequences

Both literature search and query against the Immune Epitope Database [34] (www.immuneepitope.org) were performed to detect reported immunogenic, human T-cell determinants (both class I and II) of DENV that either fully or partially overlapped with the pan-DENV sequences. In addition, dedicated algorithms based on several prediction models were used to identify candidate putative HLA-binding sequences to multiple HLA class I and II supertype alleles within the pan-DENV sequences. Putative HLA supertypes class I-restricted peptides were identified by use of NetCTL [35], Multipred [36], ARB [37], and class II-restricted peptides by Multipred and TEPITOPE [38]. Further, the intra-type representation of the putative T-cell determinants was analyzed.

The NetCTL 1.2 algorithm (www.cbs.dtu.dk/services/NetCTL/) predicts peptides restricted by 12 HLA class I supertypes (A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58 and B62). The algorithm integrates the predictions of HLA binding, proteasomal C-terminal cleavage and transport efficiency by the transporter associated with antigen processing (TAP) molecules. HLA binding and proteasomal cleavage predictions are performed by an artificial neural networks (ANN) method, while TAP transport efficiency is predicted using a weight matrix method. The parameters used for NetCTL prediction were: 0.15 weight on C terminal cleavage (default), 0.05 weight on TAP transport efficiency (default), and 0.5 threshold for HLA supertype binding, which was reported to be optimal (sensitivity (SN), 0.89 and specificity (SP), 0.94) in a large benchmark study containing more than 800 known class I T-cell determinants [35].

The TEPITOPE software (2000 beta version; courtesy of J. Hammer) utilizes quantitative matrix-based motifs, obtained from experimental scanning of the binding of P1-anchored designer peptides to soluble HLA-DR molecules in *in-vitro* competition assays, to predict peptides binding to 25 common HLA-DR alleles (DRB1*0101, *0102, *0301, *0401, *0402, *0404, *0405, *0410, *0421, *0701, *0801, *0802, *0804, *0806, *1101, *1104, *1106, *1107, *1305, *1307, *1311, *1321, *1501, *1502, and DRB5*0101) [38,39]. The parameters for TEPITOPE predictions were: 5% quantitative threshold and putative determinants with a 10-fold inhibitory residue included. Nonamer peptides predicted to bind at least 10 out of the 25 HLA-DR alleles were selected as putative supertype-restricted determinants.

Multipred (research.i2r.a-star.edu.sg/multipred/) is a computational system for the prediction of peptides that bind to HLA class I supertypes A2 and A3 and class II HLA-DR supertype [36]. The HLA alleles selected to represent these supertypes by Multipred were as follows: A2 supertype, A*0201, *0202, *0203, *0204, *0205, *0206, *0207 and *0209; A3 supertype, A*0301, *0302, *1101, *1102, *3101, *3301 and *6801; DR supertype, DRB1*0101, *0301, *0401, *0701, *0801, *1101, *1301, and *1501. Hidden Markov model (HMM) and ANN methods are the predictive models of Multipred; both have been optimized and show similar performances [36]. The sum thresholds used for prediction of peptides restricted to the three HLA supertypes by ANN and HMM methods were: A2, 31.33 (ANN; SN = 0.80 and SP = 0.83) and 47.08 (HMM; SN = 0.80 and SP = 0.78); A3, 24.53 (ANN; SN = 0.90 and SP = 0.95) and 37.58 (HMM; SN = 0.80 and SP = 0.87); and DR, 23.42 (ANN; SN = 0.90 and SP = 0.92) and 51.08 (HMM; SN = 0.90 and SP = 1.00). Consensus predictions of the two methods were taken as final predictions for each HLA supertype.

The ARB matrix method (epitope.liai.org:8080/matrix/matrix_prediction.jsp) is based on a matrix of coefficients to predict IC50 values [37]. The HLA class I alleles predicted by ARB were grouped according to the current supertype classification [19,40] and supertypes containing more than two alleles were selected, namely A2 (A*0201, *0202, *0203, *0206, and *6802), A3 (A*0301, *1101, *3101, *3301 and *6801), B7 (B*0702, A*3501, *5101, *5301, and *5401), and B44 supertypes (B*4001, *4002, *4402, *4403, and *4501). The prediction threshold value chosen for optimum sensitivity and specificity was IC50$\leq$1000 nM and nonamer peptides predicted to bind 3 or more alleles of the supertype were considered as putative promiscuous HLA supertype-restricted determinants.

ELISpot analysis of HLA-DR restricted determinants in pan-DENV sequences

All experiments were approved by the Johns Hopkins University Institutional Animal Care and Use Committee. Murine H-2 class II-deficient, HLA-DR2 [41], HLA-DR3 [42,43], HLA-DR4 (referred to as DR4/IE) [44] and HLA-DR4/human CD4 (huCD4) [45,46] Tg mice were used, bred and maintained in the Johns Hopkins University School of Medicine Animal Facility. Specific pathogen-free (SFP) colonies were maintained in a helicobacter-negative mice facility. The HLA-DR expression of the experimental transgenic mice was evaluated by flow cytometry.

Mice were immunized subcutaneously at the base of the tail, twice at two weeks interval, with pools of overlapping peptides covering the DENV-3 protein (15–17 aa, overlapping by 10–11 aa) (Schafer-N Inc., Copenhagen, Denmark; BEI Resources,

Manassas, VA). Peptide pools (73–155 peptides per pool) contained 1 μg of each peptide and were emulsified (1:1) in TiterMax adjuvant (TiterMax USA, Inc.). An aqueous preparation of TiterMax (1:1) was used as a negative control. Two weeks after the second immunization, the mice were sacrificed and HLA-DR-restricted CD4 T cell reponses were assessed by *ex vivo* IFN-γ ELISpot assay using CD8-depleted splenocytes. Each target peptide was tested in duplicate. Spot-forming cell (SFC) counts were normalized to 10^6 cells. The results were considered significant when the average SFC minus two standard deviations (SD) was greater than the average of the background plus two SD; and the average values were greater than 10 SFC per 10^6 splenocytes. The initial screening assays were performed with peptide matrices [47], followed by assays with the relevant individual peptides (Nascimento *et al.*, manuscript in preparation).

Results

Dengue virus type protein datasets

A total of 9,512 and 12,404 complete and partial DENV protein sequences were collected from the NCBI Entrez protein database of December 2005 and 2007, respectively, representing an increase of approximately 30% (2892 sequences) in the 24-months interval (Table 1). The total number of sequences (2007) varied from 4,011 for DENV-2 to 1,415 for DENV-4 and from 3,845 for E to 523 for NS4a proteins. Most of the individual protein sequences originated from DENV strains that were unique variants with respect to the entire polyprotein, but were identical to other strains with respect to individual proteins [48].

Conserved pan-DENV sequences

The consensus-sequence approach [20,25] identified a total of 44 pan-DENV sequences of at least 9 amino acids that were present in ≥80% of all sequences of each DENV type for both 2005 and 2007 datasets (**Figure 2; Table S1**). Strikingly, 34 of the 44 (~77%) were conserved in ≥95% of all reported DENV sequences. The size of the pan-DENV sequences ranged from 9 to 22 amino acids, with a combined size of 514 residues, corresponding approximately to 15% of the complete DENV polyprotein (~3390 amino acids) (**Table 2**). The vast majority (42/44) of the pan-DENV sequences were localized in the NS proteins, with 17, 12, 7 and 5 sequences found in NS5, NS3, NS1 and NS4b, respectively, and 1 in the NS4a protein. Notably, the remaining two pan-DENV sequences were localized in the E protein. No region of at least 9 amino acids and conserved in ≥80% of the sequences of each DENV type was found in the C, prM, NS2a and NS2b proteins. The largest size of the combined pan-DENV sequences was in the NS5 protein, representing a total of 215 amino acid positions covering ~24% of the protein, followed by NS3, NS1 and NS4b with 122, 74 and 69 amino acid positions covering ~20, ~21 and ~28% of the corresponding proteins, respectively. The two pan-DENV sequences in the E protein had a combined size of only 25 amino acids, corresponding to ~5% of the protein.

In large-scale genomic analyses such as this study, biases may result from the collection of completely or partially overlapping redundant sequences, corresponding to identical or highly similar circulating DENV isolates sequenced by various dengue surveillance programs in different countries. Although to some extent this redundancy may be accepted as a reflection of the incidence of the corresponding DENV isolates in nature, we assessed its potential bias effect by repeating the analysis of conservation after discarding duplicate sequences from the datasets. The analysis of unique sequences identified all the pan-DENV sequences that were identified when including duplicates (**Figure 2**), except for $NS1_{12–20}$, $NS1_{25–35}$ and $NS5_{597–616}$. Therefore, the presence of duplicates in the DENV datasets did not significantly affect the results. Although the removal of duplicates does not fully compensate for biases in the datasets, the removal of highly similar sequences, which may have been generated from relatively large sequencing efforts in single outbreaks, was deemed undesirable, since such arbitrary selection would introduce additional biases.

Evolutionary diversity of DENV protein nonamer peptide sequences

The evolutionary diversity of each DENV type, and the 4 types combined, was studied by use of Shannon information entropy

Table 1. Number and distribution of reported DENV protein sequences.

DENV protein[b]	No. of sequences[a]										
	DENV-1		DENV-2		DENV-3		DENV-4		*Total*		
	2005	2007	2005	2007	2005	2007	2005	2007	2005	2007	Increase
C	194	298	266	311	414	547	117	122	991	1278	287
prM	206	311	353	404	458	590	207	225	1224	1530	306
E	852	1051	1277	1518	716	910	338	366	3183	3845	662
NS1	410	565	640	752	201	308	142	159	1393	1784	391
NS2a	150	238	132	173	90	169	121	125	493	705	212
NS2b	136	224	130	163	104	183	40	44	410	614	204
NS3	98	186	145	178	216	297	30	34	489	695	206
NS4a	91	178	128	162	70	151	28	32	317	523	206
NS4b	89	176	129	163	70	150	109	113	397	602	205
NS5	92	179	151	187	181	267	191	195	615	828	213
Total	2318	3406	3351	4011	2520	3572	1323	1415	9512	12404	2892

[a]Collected from the NCBI Entrez protein database
[b]Manually processed after multiple sequence alignments and use of the known DENV cleavage sites
doi:10.1371/journal.pntd.0000272.t001

Figure 2. Pan-DENV sequences and their representations in the 4 DENV types. The 44 pan-DENV sequences of at least 9 amino acids that were found present in ≥80% of the recorded sequences of each DENV type are shown. The representation values are shown for the 2005 dataset; see Table S1 for values of both 2005 and 2007 datasets. Amino acid positions were numbered according to the sequence alignments of the 4 DENV types. The corresponding proteins are indicated on the left.
doi:10.1371/journal.pntd.0000272.g002

Table 2. Distribution and size of the pan-DENV sequences.

DENV protein	Size (aa)	Pan-DENV sequences[a]		
		No.	Size[b]	% of protein[c]
C	113–115	0	0	0
prM	166	0	0	0
E	493–495	2	25	5
NS1	352	7	74	21
NS2a	218	0	0	0
NS2b	130	0	0	0
NS3	618–619	12	122	20
NS4a	150	1	9	6
NS4b	245–249	5	69	28
NS5	900–904	17	215	24
Total	3387–3398	44	514	15

[a]Sequences of at least 9 amino acids that were represented in ≥80% of all DENV sequences of each type

[b]Combined amino acid size of all pan-DENV sequences in the protein

[c]Percentage of the combined pan-DENV sequence size over that of the corresponding protein size

doi:10.1371/journal.pntd.0000272.t002

[27], modified to examine the variability of nonamer peptide sequences, as described in the *Methods*. The entropy of the proteome of the recorded viruses of each type showed numerous long regions of low entropy (≤1), reflecting the relatively high degree of intra-type sequence conservation, in particular in the NS3, NS4b and NS5 proteins (**Figure 3A–D)**. Overall, the average intra-type nonamer entropy values of the individual protein sequences of DENV-1, -2, -3 and -4 ranged from 0.2 for the DENV-4 NS4b to 1.0 for DENV-2 prM (**Figure S1**). Of note, however, were the marked differences in the relative degree of entropy of each protein between the 4 DENV types. For example, NS4b had the least diversity of the proteins of 3 types, but was replaced in DENV-2 by NS2b, which was the second most variable in DENV-3. The consequence of the differences in the sequences of each protein between the 4 types was a marked increase in the peptide entropy across the DENV 1-4 proteomes (**Figure 3E),** with average peptide entropy ranging from 1.6 for NS3 to 2.6 for NS2a (**Figure S1**), except for 44 sharply defined regions of low nonamer entropy (≤0.5) where the sequences were highly conserved in all DENVs (**Figure 3E)**, with no significant difference between the 2005 and 2007 datasets (**Table S2**). Majority of the pan-DENV sequences had entropy values of ≤0.3, corresponding to the intra-type representation of ≥90%. Thus, the congruent consensus- and entropy-based analyses of the DENV nonamer peptides revealed highly conserved and evolutionarily stable pan-DENV sequences distributed in several viral proteins, despite the marked viral diversity defining multiple DENV types, genotypes and variants [49].

Representation of DENV variant nonamer peptide sequences

The combined representation of variant peptides that differed by at least one amino acid from the predominant peptide was also

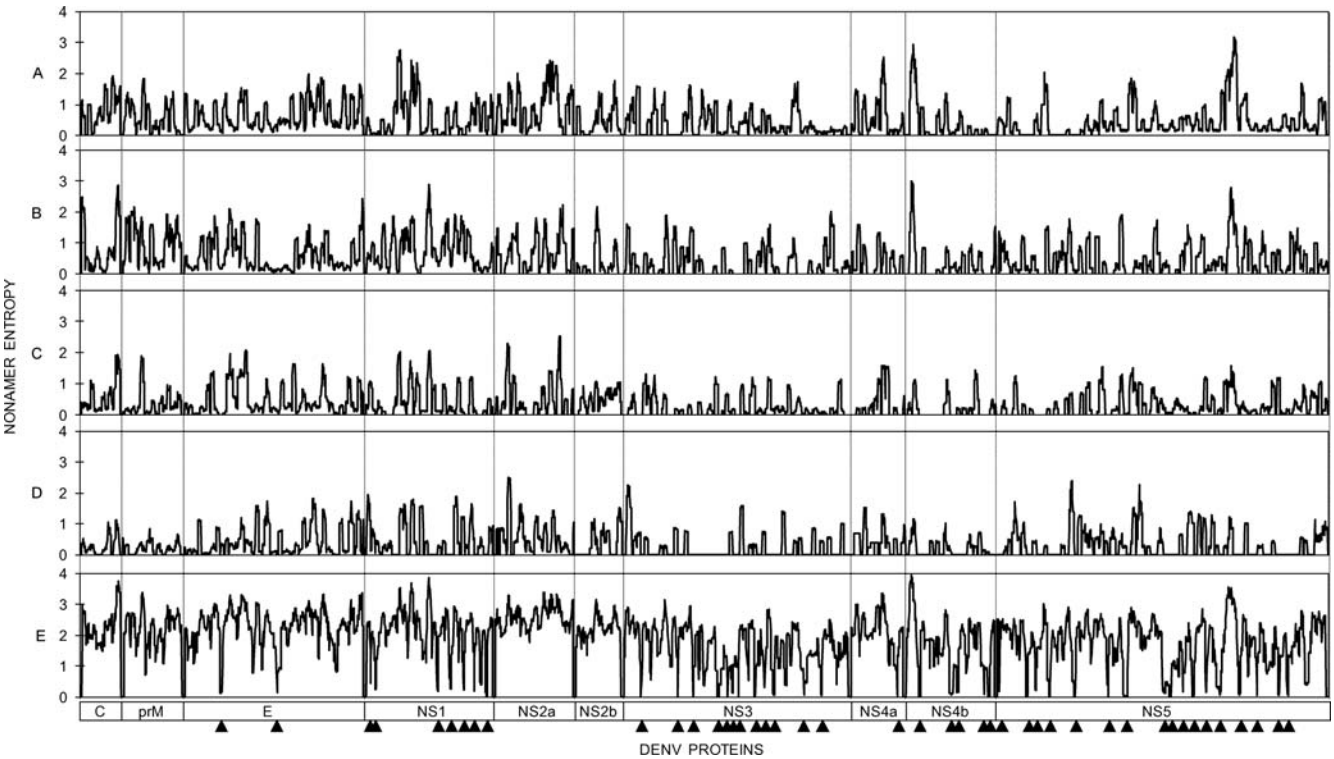


Figure 3. Shannon entropy of nonamer peptides within and across DENV types sequences. The entropy values were computed from the alignments of DENV sequences using the Antigenic Variability Analyzer software, as described in the *Methods*. Values were plotted for DENV-1 (A), DENV-2 (B), DENV-3 (C), DENV-4 (D), and all 4 DENV types (E) sequences (2005 dataset). Entropy values around protein cleavage sites are non significant, since the corresponding positions cannot be the center of a nonamer (*see Methods*). The triangles below indicate the locations of the pan-DENV sequences in the corresponding proteins.

doi:10.1371/journal.pntd.0000272.g003

analyzed at each nonamer position. Examples of this analysis for DENV-3 proteins are shown in **Table 3**. Nonamers that lack entropy (zero entropy) have one sequence in all of the recorded virus isolates, and therefore have no variants. Positions with high entropy can contain many different variant peptides, each at lesser (or equal) frequency than the predominant peptide. The combined representation of variant peptides at each nonamer position across the proteome of each individual DENV type was generally low, representing less than 10% of the corresponding sequences, except for some positions where it was more than 50% (**Figure 4A–D**). Notably, the nonamer position with the highest combined variant representation for each DENV type was found in the nonstructural proteins and not the structural ones, with representation values ranging from ~61 to ~78% (DENV-1 NS5, DENV-2 NS5, DENV-3 NS2a, and DENV-4 NS1 and NS3 proteins). When representations of variants across all DENVs were calculated, the majority of all nonamer sites contained variants that together represented ~60–85% of the total DENV sequences at that site (the highest representation of ~85% was in the NS1 protein) (**Figure 4E**). This was in striking contrast to the 0 to ~5% combined representation of variants at each nonamer position in the pan-DENV sequences, with no significant difference between the 2005 and 2007 datasets (**Table S2**). The majority of all nonamer sites in the pan-DENV sequences lacked variant or contained variants that together represented <1% of all recorded DENVs. These data further illustrate the extremely high genetic stability of the 44 pan-DENV sequences, among all recorded DENV sequences and demonstrate that irrespective of the high variability between the sequences of the 4 DENV types, the representation of variants in the pan-DENV sequences was almost negligible.

Functional and structural correlates of the pan-DENV sequences

Highly conserved protein sequences are likely to represent critical sites and domains [50]. A search of the literature and the Prosite and Pfam databases [30,32] revealed that 27 of the 44 pan-DENV sequences were associated with biological activities (**Table S3**); the functional significance of the remaining 17 pan-DENV sequences was not known. The two pan-DENV sequences in the E protein corresponded to the fusion peptide (positions 98 to 110) and dimerisation domain [51,52]. In NS3, one pan-DENV sequence corresponded to the peptidase family S7 (*Flavivirus* serine protease) domain and comprised the His-51 catalytic residue [53], 3 sequences corresponded to known/putative *Flavivirus* Asp-Glu-Ala-Asp/His (DEAD/H) domain associated with ATP-dependent helicase activity [54], and two sequences were predicted to be required for cell attachment and targeting signal for microbodies. In NS5, one pan-DENV sequence corresponded to the conserved methyltransferase (MTase) *S*-adenosyl-L-methionine binding motif I (positons 77–86) involved in viral RNA capping [55], and two sequences corresponded to RNA dependent RNA polymerase (RdRp) domain [56]. Furthermore, 6 of the 27 pan-DENV sequences were predicted to exhibit post-translational modification(s), including N-glycosylation, protein kinase C and casein kinase II phosphorylation, N-myristoylation and/or amidation (**Table S3**).

It is generally recognized that amino acids buried inside proteins are subject to greater interactions and packing constraints [57] than those exposed on the outer surface. Although none of the DENV protein structures in the protein data bank (PDB) [33] was full-length, 19 of the 44 pan-DENV sequences could be mapped on the available crystallographic models of the E ectodomain (Accession No. 1OAN; 394 out of 493–495 residues), NS3 (1BEF and 2BMF, 181 and 451 out of 618–619 residues, respectively) and NS5 fragments (1R6A, 295 out of 900–904 residues). Eleven of the 19 pan-DENV sequences were buried, 2 partially exposed and 6 exposed at the surface of the corresponding structures (**Figure S2**). However, these results should be considered preliminary until full-length 3-D structures are available.

Distribution of pan-DENV sequences in nature

Twenty-seven (27) of the 44 pan-DENV sequences overlapped at least 9 amino acid sequences of as many as 64 other viruses of the family *Flaviviridae*, genus *Flavivirus* (**Figure 5**). *Zika virus* shared 22 of the 27 sequences; *Ilheus* and *Kedougou* viruses, 18; and representatives of some of the significant human pathogens, *West Nile*, *St. Louis encephalitis*, *Japanese encephalitis*, *Yellow fever* and *Tick-borne encephalitis* viruses, shared from 16 to 9 pan-DENV sequences. Thirteen (13) of the 27 sequences represented NS5, of which 9 were present in at least

Table 3. Examples of the distribution of variant nonamer peptides in DENV-3.

DENV-3 protein	Nonamer position	No. of sequences	Nonamer peptides[a]	Representation of peptides	Combined % representation of variants[b]	Nonamer entropy[c]
E	14	479	DFVEGLSGA	479 (100%)	0	0
NS2a	176	64	LAGISLLPV	25 (39%)	61	2.4
			LAGVSLLPV	11 (17%)		
			LAGVSLLPL	9 (14%)		
			LAVISLLPV	9 (14%)		
			LAGISLLPL	6 (9%)		
			LAGISLFPV	2 (3%)		
			LAGISLMPV	2 (3%)		
NS4a	86	68	SIGLICVVA	39 (57%)	43	1.5
			SIGLICVIA	19 (28%)		
			SIGLICVIV	8 (13%)		
			SIGLICVAA	2 (3%)		

[a]The predominant peptide is underlined
[b]Variants include all the peptides at the position, except the predominant
[c]Entropy value of all the peptides at the position (predominant peptide included)
doi:10.1371/journal.pntd.0000272.t003

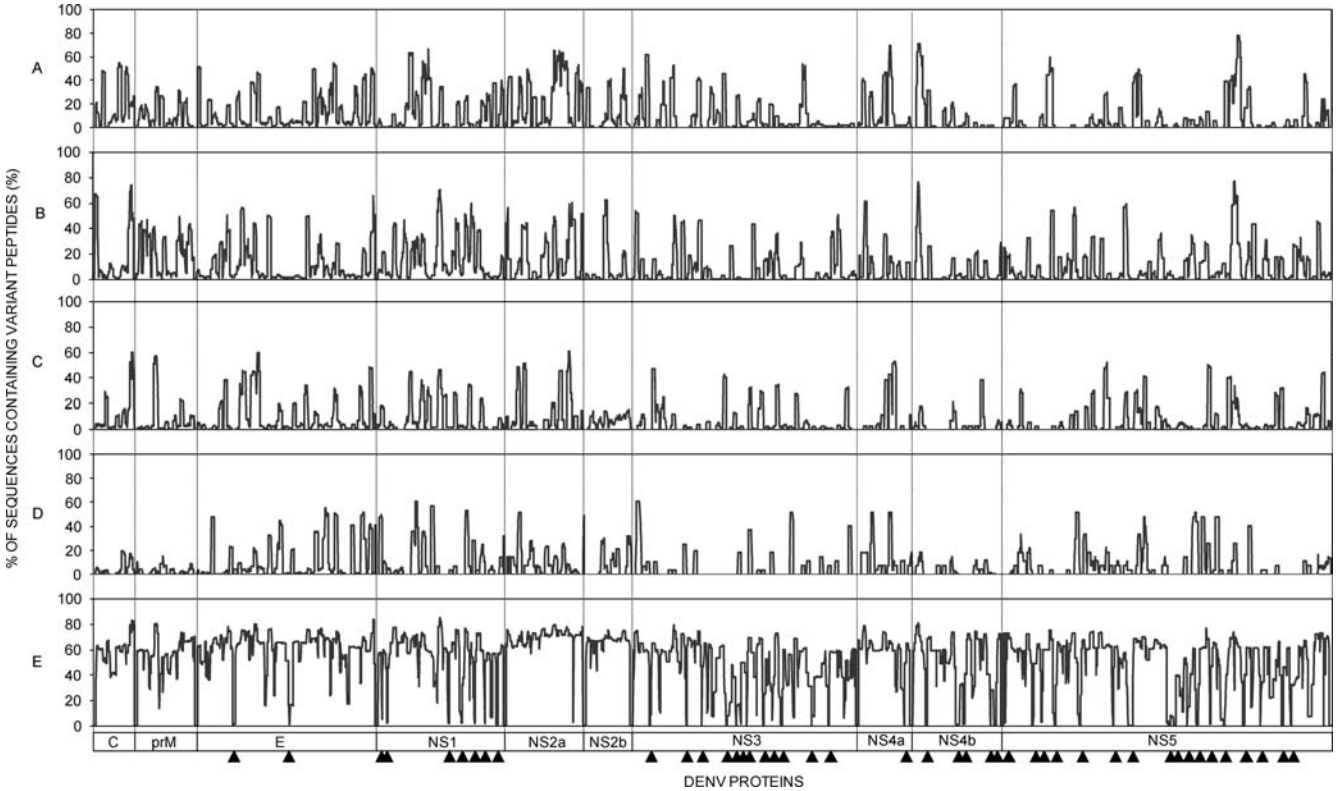


Figure 4. Percentage representations of variant nonamer peptides within and across DENV types sequences. The percentage of sequences that contained variant peptides at each nonamer position are shown for DENV-1 (A), DENV-2 (B), DENV-3 (C), DENV-4 (D), and all 4 DENV types (E) (2005 dataset). Values around protein cleavage sites are non significant (see Figure 3). The triangles below indicate the locations of the pan-DENV sequences in the corresponding proteins.
doi:10.1371/journal.pntd.0000272.g004

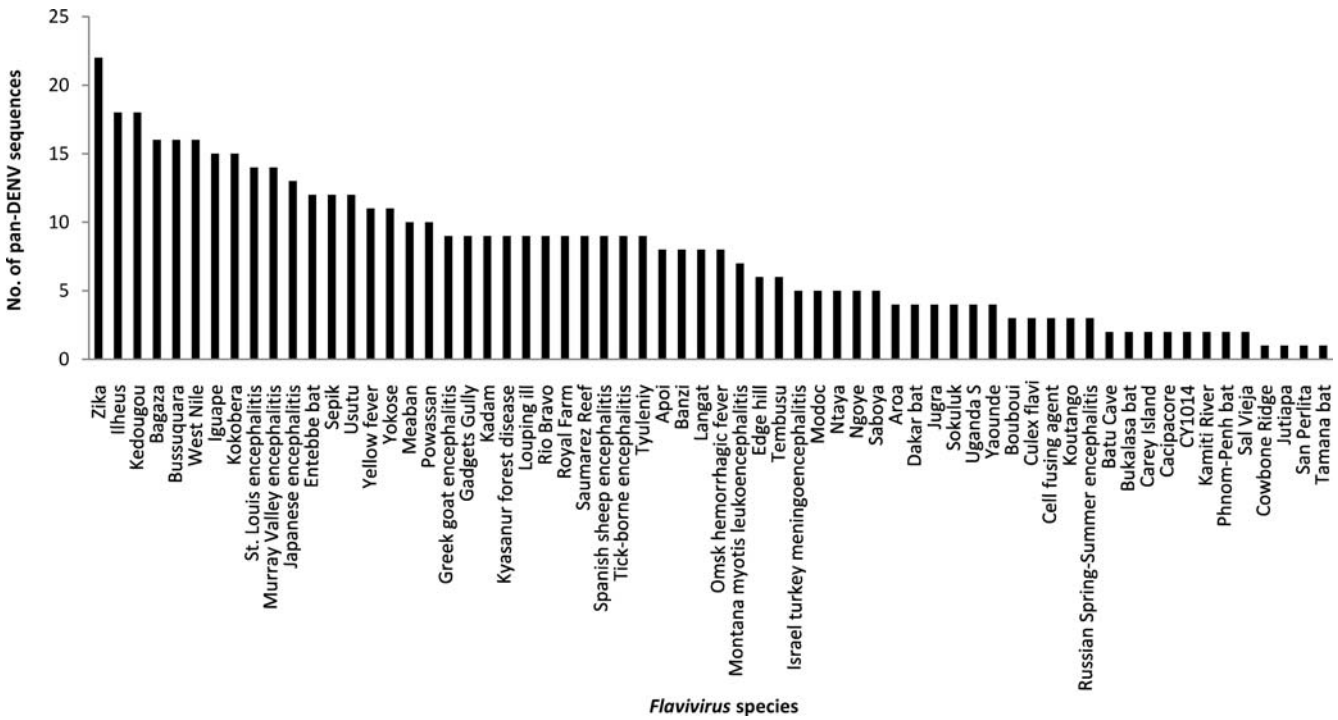


Figure 5. Number of pan-DENV sequences conserved in the different *Flaviviruses.*
doi:10.1371/journal.pntd.0000272.g005

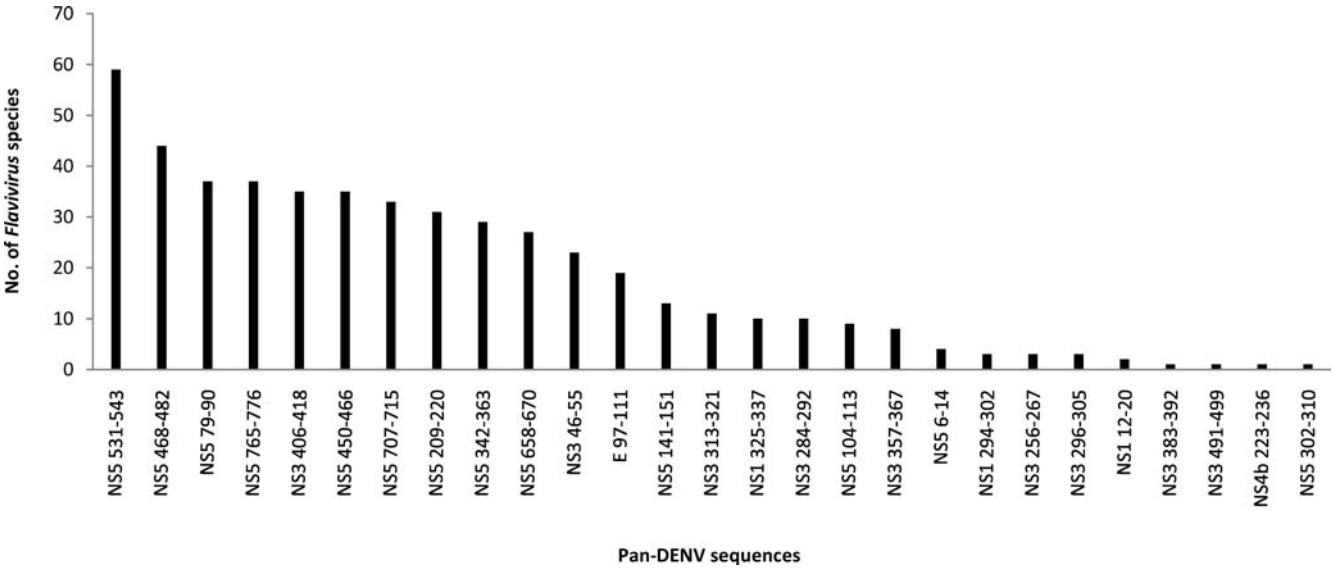


Figure 6. Number of *Flaviviruses* shared by the pan-DENV sequences.
doi:10.1371/journal.pntd.0000272.g006

27 *Flavivirus* species; 9 represented NS3, of which two were found in 35 and 23 species; one E sequence was found in 19 species; and the remaining were in NS1 and NS4b (**Figure 6; Table S4**). Five (5) of the 27 were associated with known biological activities ($NS5_{79–90}$ MTase, $NS5_{658–670}$ RdRp, $NS3_{46–55}$ peptidase S7, $NS3_{284–292}$ DEAD/H and $E_{97–111}$ dimerisation/fusion domains). Interestingly, two sequences, $NS3_{406–418}$ and $NS5_{597–616}$, overlapped 9 amino acid sequences of the cell fusing agent virus polyprotein-like protein from the mosquito *Aedes albopictus* [58], and the phage-related tail fibre protein-like protein from the bacteria *Chromohalobacter salexigens DSM 3043*, respectively.

The representation of many of the pan-DENV sequences was high among known sequences of several of the highly studied *Flaviviruses* (**Table S4**): *St. Louis encephalitis*, *West Nile*, *Japanese encephalitis*, *Murray Valley encephalitis*, *Usutu*, *Kokobera*, *Ilheus*, *Tick-borne encephalitis*, *Langat*, *Omsk hemorrhagic fever*, *Louping ill*, *Powassan*, *Kyasanur forest disease and Yellow fever* viruses. Protein sequence data for the rest of the *Flaviviruses* that shared pan-DENV sequences was limited (<10 sequences) in the public database. Seven of the 27 pan-DENV sequences, $NS1_{12–20}$, $NS3_{256–267}$, $NS3_{383–392}$, $NS3_{491–499}$, $NS4b_{223–236}$, $NS5_{6–14}$ and $NS5_{302–310}$, were present in a few species with less than 10 reported total sequences (**Table S4**).

Known and predicted HLA supertype-restricted, pan-DENV T-cell determinants

Literature survey and database search revealed that 10 of the pan-DENV sequences (9 in NS3, one in E) overlapped at least 9 amino acids of 15 previously reported DENV T-cell determinants immunogenic in human, with their HLA restriction, when known, showed both class II (DR*15, DPw2) and class I (A*11) specificities (**Table 4**). Further evaluation of the immune-relevance of the pan-DENV sequences included a search for candidate putative promiscuous HLA supertype-restricted T-cell determinants within these regions by use of several computational algorithms: NetCTL [35], Multipred [36], ARB [37] and TEPITOPE [38]. Overall, 34 of 44 (~77%) pan-DENV sequences (**Figure 7**), identified in the NS5, NS3, NS1, E and NS4a proteins were predicted to contain 100 supertype-restricted binding nonamers (**Table S5**). The majority (88/100) of the predicted promiscuous HLA-binding nonamers were present in ≥95% of the sequences of each DENV type (**Table S6**). Thirty-one (~91%) of the 34 putative supertype pan-DENV sequences contained HLA-binding nonamers for multiple HLA supertypes. Clusters (hotspots) of two or more overlapping HLA-binder nonamer core peptides were present in 27 (~79%) of the 34 putative supertype pan-DENV sequences. About half (14/27) of these clusters contained three or more nonamer binders overlapping by 8 amino acids, covering most or the entire corresponding conserved region.

Immunogenicity of HLA-DR-restricted pan-DENV sequences in HLA Tg mice

The immunogenicity of the pan-DENV sequences was also analyzed by assay of peptide-specific HLA-restricted T-cell responses in murine H-2 class II-deficient, HLA-DR Tg mice expressing 3 prototypic HLA-DR alleles, corresponding to the divergent subgroups HLA-DR2 (DRB1*1501), HLA-DR3 (DRB1*0301), and HLA-DR4 (DRB1*0401). Mice were immunized with pools of overlapping peptides covering the sequences of the E, NS1, NS3, and NS5 proteins of DENV-3, and HLA-DR-restricted CD4 T-cell responses were assessed by IFN-γ ELISpot assays using CD8-depleted splenocytes. Thirty peptides eliciting positive T-cell responses in the HLA Tg mice contained 9 or more consecutive amino acids of 22 pan-DENV sequences, that were localized in the NS5 (11), NS3 (6), NS1 (4), and E proteins (one) (**Table 5**). Overall, 9, 10 and 18 peptides elicited positive responses in HLA-DR2, -DR3, and/or -DR4 Tg mice, respectively; 20 corresponded to sequences of NS5, 10 of NS3, 6 of NS1, and one of E. Furthermore, at least 7 of the pan-DENV sequences, all localized in the NS5 and NS1 proteins, contained promiscuous T-cell determinants for multiple HLA-DR alleles (**Table 5**). These data, together with those previously reported **(Table 4),** showed that a minimum of 26 of the 44 pan-DENV sequences, distributed predominantly in the NS5 and NS3 proteins, and to a lesser extent in NS1 and E, contained numerous HLA-restricted class II and/or class I determinants demonstrated by assays of T-cell responses *in vivo*.

Discussion

In this study, we identified and characterized pan-DENV sequences that were highly conserved in all recorded DENV

Table 4. Reported human T-cell determinants in the pan-DENV sequences.

DENV protein	Pan-DENV sequence[a]	Immunogenic T-cell determinants[b]			
		Sequence[c]	T subset	HLA Ag	Reference(s)
E	$_{252}$VLGSQEGAMH$_{261}$	KKQDVVVLGSQEGAM	-	-	[76]
NS3	$_{46}$FHTMWHVTRG$_{55}$	TFHTMWHVTRGAVLM	CD4	-	[76]
	$_{148}$GLYGNGVVT$_{156}$	KVVGLYGNGVVTRSG	CD4	DR*15	[76]
	$_{189}$LTIMDLHPG$_{197}$	KRLTIMDLHPGAGKT	CD4	-	[72]
		RKLTIMDLHPGSGKT	CD4	-	[72]
		RKLTIMDLHPGAGKT	CD4	-	[72]
		RNLTIMDLHPGSGKT	CD4	-	[72]
	$_{256}$EIVDLMCHATFT$_{267}$	EHTGREIVDLMCHAT	CD4	-	[76]
		EIVDLMCHATFTMRL	CD4	-	[76]
		EIVDLMCHAT	CD4	DPw2	[77,78]
	$_{284}$MDEAHFTDP$_{292}$	LIIMDEAHFTDPASI	-	-	[76]
	$_{313}$IFMTATPPG$_{321}$	AGIFMTATPPGSRDP	-	-	[76]
	$_{357}$GKTVWFVPSIK$_{367}$	TVWFVPSIK	CD8	A*11	[16]
	$_{383}$VIQLSRKTFD$_{392}$	KKVIQLSRKTFDSEY	-	-	[76]
	$_{406}$VVTTDISEMGANF$_{418}$	NDWDFVVTTDISEMG	-	-	[76]

[a]Amino acid positions numbered according to the sequence alignments of the 4 DENV types
[b]Dashes, not determined
[c]Sequences present in the pan-DENV sequences are underlined
doi:10.1371/journal.pntd.0000272.t004

isolates. The large number of sequences analyzed (12,404 as of December 2007), and their wide distribution in terms of geography and time (1945–2007) (data not shown), offered information for a broad survey of DENV protein diversity in nature. The 44 pan-DENV protein sequences of at least 9 aa, covering 514 aa or about 15% of the complete DENV polyprotein of ~3390 aa, were conserved in at least 80% of all recorded DENV sequences, and 34 of the 44 (~77%) were conserved in ≥95% of DENV sequences. All the 44 were in the non-structural proteins except for the two E sequences. These conserved sequences have shown remarkable stability over the entire history of DENV sequences deposited in the NCBI Entrez protein database, as illustrated by their low peptide entropy values and variant frequencies. In addition, 27 of the pan-DENV sequences were conserved in 64 other *Flaviviruses*, as further evidence of prolonged evolutionary stability within this genus, as previously discussed [59–61]. Two are also present in the proteomes of the *Aedes albopictus* mosquito and the bacteria *Chromohalobacter salexigens*, possibly in keeping with recent reports of the genetic recombination between phyla [58]. It is likely that these pan-DENV sequences have been under selection pressure to fulfill critical biological and/or structural properties, some of which have been identified for the E (fusion peptide, dimerization domain), NS3 (peptidase S7, DEAD/H domains) and NS5 proteins (MTPase, RdRp domains) [51–56]. Hence, these conserved sequences are unlikely to significantly diverge in newly emerging DENV isolates in the future, and represent attractive targets for the development of specific anti-viral compounds and vaccine candidates.

There also is evidence that many of the conserved sequences are immunologically relevant. A majority (26/44) contained at least 9 amino acids overlapping with a total of 45 peptides that have been reported to be immunogenic in humans and/or HLA-DR Tg mice. In addition, putative T-cell determinants for 12 major HLA class I supertypes and for class II DR supertype, with broad application to the immune responses of human population worldwide, were predicted by computational analysis. Some of the putative T-cell determinants were predicted to be promiscuous to multiple HLA supertypes, in addition to multiple alleles of a given HLA supertype. Such a degree of promiscuity has previously been observed for DENV [62] and HIV peptides [63], among others. The existence of conserved T-cell determinants specific for multiple HLA supertypes further supports their evaluation as vaccine targets, since they would provide broader population coverage [63]. Many of the predicted HLA binding nonamers were localized in clusters, as we have also observed in HLA Tg mice immunized with WNV proteins and DNA encoding the SARS coronavirus N protein [64], and has been reported in studies of *human immunodeficiency virus* (HIV) type 1 proteins [65–68], the outer membrane protein of *Chlamydia trachomatis* [69], and other antigens [64].

The significant sequence variations between the proteins of the 4 DENV types represent a cardinal issue for the development of a tetravalent DENV vaccine that provides robust protection against each DENV type. Subtle amino acid substitutions within T-cell determinants restricted by a given HLA allomorph, such as in the event of sequential heterologous infections, or between a vaccine formulation and a subsequent natural infection [7], can dramatically alter the phenotype of the specific T cells, resulting in a wide range of effects from agonism to antagonism [9,12–15]. Because of the extent of intra-type (1 to 21%) and inter-type (14 to 67%) amino acid variability among DENV isolates [48], many nonamer T-cell determinants contain single or multiple amino acid difference(s). When the 4 DENV types were analyzed together, a majority of the nonamer positions across the full proteome exhibited variants that together were present in ~60 to ~85% of all sequences. The frequencies of variant peptides across the 4 DENV types suggest that vaccine strategies incorporating whole DENV immunogens, such as inactivated and recombinant subunit

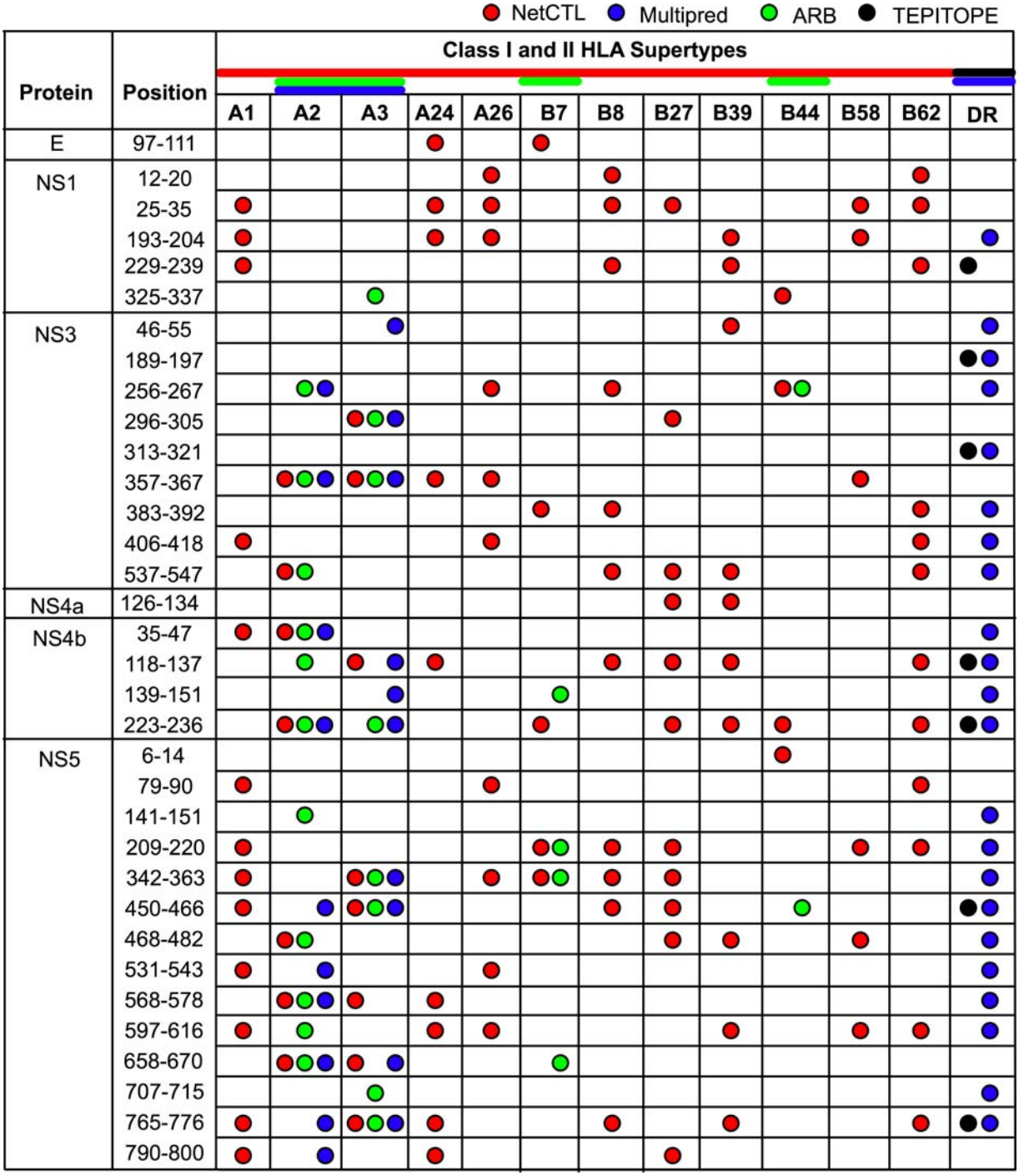


Figure 7. Candidate putative HLA supertype-restricted, pan-DENV T-cell determinants predicted by computational algorithms. Amino acid positions of the pan-DENV sequences are numbered according to the sequence alignments of the 4 DENV types; the corresponding DENV proteins are indicated on the left. Predicted HLA-restricted T-cell determinants were identified using NetCTL, Multipred, ARB and TEPITOPE algorithms (see Methods).
doi:10.1371/journal.pntd.0000272.g007

vaccines, live attenuated viruses, or chimeric viruses expressing structural DENV genes, are likely to elicit T-cell responses to altered peptide ligands. This phenomenon is also likely to occur in individuals exposed to several *Flaviviruses*, such as DENV, JEV and YFV that are co-circulating in regions of Asia, India or South America, or following vaccination [70].

While the immune correlates of DENV protection remain poorly documented, there is evidence that both neutralizing antibody and specific T-cell responses are required [7,71]. The incorporation of defined HLA-restricted T-cell determinants within DENV vaccine candidates might improve vaccine efficiency by increasing T-cell help to sustain a robust, long-lived

Table 5. Immunogenicity of the pan-DENV sequences in HLA-DR transgenic mice.

DENV protein	Pan-DENV sequence[b]	Ag-specific CD4 T-cell responses[a]			
		Peptide sequences (DENV-3)[c]	IFN-γ-SFC/10^6splenocytes±SD[d]		
			DR2	DR3	DR4
E	$_{252}$VLGSQEGAMH$_{261}$	PEVVVLGSQEGAMHT	-	-	88±34
NS1	$_{193}$AVHADMGYWIES$_{204}$	AVHADMGYWIESQKN	-	17±1	-
	$_{229}$HTLWSNGVLES$_{239}$	WPKSHTLWSNGVLES	-	129±3*	-
		HTLWSNGVLESDMII	-	131±103	37±3
	$_{266}$GPWHLGKLE$_{274}$	HTQTAGPWHLGKLE	-	333±6	-
	$_{294}$RGPSLRTTT$_{302}$	TRGPSLRTTTVSGKL	-	-	11±4
NS3	$_{189}$LTIMDLHPG$_{197}$	KKRNLTIMDLHPGSG	-	-	50±16
	$_{296}$AARGYISTRV$_{305}$	ASIAARGYISTRVGM	40±14	-	-
		ARGYISTRVGMGEAA	9±4	-	-
	$_{313}$IFMTATPPG$_{321}$	EAAAIFMTATPPGTA	-	-	474±116
		IFMTATPPGTADAFP	-	-	323±287
	$_{357}$GKTVWFVPSIK$_{367}$	TDFAGKTVWFVPSIK	48±15	-	-
		GKTVWFVPSIKAGND	396±14	-	-
	$_{383}$VIQLSRKTFD$_{392}$	KKVIQLSRKTFDTEY	-	21±3	-
	$_{406}$VVTTDISEMGANF$_{418}$	FVVTTDISEMGANFK	-	-	408±104
		TDISEMGANFKADRV	-	152±33	-
NS5	$_{302}$TWAYHGSYE$_{310}$	DENPYKTWAYHGSYEVK	126±10*	-	14±5
		TWAYHGSYEVKATGSA	161±20*	-	63±17
	$_{342}$AMTDTTPFGQQRVFKEKVDTRT$_{363}$	MVTQMAMTDTTPFGQQR	-	-	28±0*
	$_{450}$CVYNMMGKREKKLGEFG$_{466}$	GSCVYNMMGKREKKLGE	-	-	13±2
	$_{505}$SGVEGEGLH$_{513}$	NSYSGVEGEGLHKLGYI	-	-	184±15
	$_{531}$YADDTAGWDTRIT$_{543}$	KIPGGAMYADDTAGWDT	-	-	46±3
	$_{568}$IFKLTYQNKVV$_{578}$	ANAIFKLTYQNKVVKVQ	577±384	-	24±9*
	$_{597}$DQRGSGQVGTYGLNTFTNME$_{616}$	VMDIISRKDQRGSGQVG	-	88±1	-
	$_{658}$RMAISGDDCVVKP$_{670}$	VERLKRMAISGDDCVVK	-	159±24	16±6
		MAISGDDCVVKPIDDRF	-	249±39	-
	$_{707}$VPFCSHHFH$_{715}$	DWQQVPFCSHHFHELIM	32±8*	34±11	-
	$_{765}$LMYFHRRDLRLA$_{776}$	MYFHRRDLRLASNAI	75±16*	-	33±9
	$_{790}$PTSRTTWSIHA$_{800}$	VHWVPTSRTTWSIHAHH	-	-	83±1
		SRTTWSIHAHHQWMTTE	-	-	122±46

[a]Assessed by IFN-γ ELISpot assay in HLA-DR2 (DRB1*1501), HLA-DR3 (DRB1*0301) and HLA-DR4 (DRB1*0401) Tg mice immunized with DENV-3 peptides (see Methods)
[b]Amino acid positions numbered according to the sequence alignments of the 4 DENV types
[c]Sequences present in the pan-DENV sequences are underlined
[d]SFC, spot-forming cells; SD, standard deviation. Representative results from at least two immunized Tg mice are shown, except when indicated by an asterisk
doi:10.1371/journal.pntd.0000272.t005

immunity, and possibly through direct cytostatic and cytotoxic effects on infected cells. For tetravalent formulations, it may be relevant to focus primarily on sequences that are conserved in all 4 DENV types and to avoid the regions of T-cell immunity that are highly variable, unless they are strictly type-specific [17,72]. The two pan-DENV E sequences (positions 97–111 and 252–261) and the exposed domain III of the E antigen (positions 300–400) [73,74], are also candidate sequences for neutralizing antibody responses. An additional criterion for the selection of T-cell targets is the need for determinants with broad HLA representation, as it has been emphasized in the recognition of HLA supertypes [18–20]. Further investigations are needed to validate the immunogenicity of the candidate T-cell determinants in human subjects, and to identify sequences associated with deleterious T-cell responses.

The global approach described herein provides a framework and methodology for large-scale and systematic analysis of conserved sequences of other pathogens, in particular for rapidly evolving viruses such as influenza A virus [75] and HIV [63]. These studies will offer insights into their diversity and evolutionary history, together with providing critical data for rational vaccine development, structure-based design of candidate inhibitory compounds, and improvement of the current diagnostic methods.

Supporting Information

Figure S1 Average nonamer peptide entropy for each protein of each DENV type and all the four types combined. The values are shown for the 2005 dataset.
Found at: doi:10.1371/journal.pntd.0000272.s001 (0.70 MB TIF)

Figure S2 Molecular location of 19 pan-DENV sequences (in red) on the protein's 3-D structure. These sequences were mapped on the available crystallographic models of the E ectodomain (PDB Accession No. 1OAN; 394 out of 493-495 residues), NS3 (1BEF and 2BMF, 181 and 451 out of 618-619 residues, respectively) and NS5 fragments (1R6A, 295 out of 900-904 residues). The major portions of eleven of the 19 pan-DENV sequences were buried (NS3-$_{148}$GLYGNGVVT$_{156}$, $_{256}$EIVDLMCHATFT$_{267}$, $_{284}$MDEAHFTDP$_{292}$, $_{296}$AARGYISTRV$_{305}$, $_{313}$IFMTATPPG$_{321}$, $_{357}$GKTVWFVPSIK$_{367}$, $_{406}$VVTTDISEMGANF$_{418}$, and $_{491}$EAKMLLDNI$_{499}$; NS5-$_{79}$DLGCGRGGWSYY$_{90}$, $_{141}$DTLLCDIGESS$_{151}$ and $_{209}$PLSRNSTHEMYW$_{220}$), 2 were partially buried/exposed (NS3-$_{46}$FHTMWHVTRG$_{55}$ and $_{537}$LMRRGDLPVWL$_{547}$) and the remaining 6 were exposed (E-$_{97}$VDRGWGNGCGLFGKG$_{111}$ and $_{252}$VLGSQEGAMH$_{261}$; NS3-$_{189}$LTIMDLHPG$_{197}$ and $_{383}$VIQLSRKTFD$_{392}$; NS5-$_{6}$GETL-GEKWK$_{14}$ and $_{104}$TKGGPGHEEP$_{113}$) at the surface of the corresponding structures.
Found at: doi:10.1371/journal.pntd.0000272.s002 (9.65 MB DOC)

Table S1 The intra-type percentage representation of pan-DENV sequences.
Found at: doi:10.1371/journal.pntd.0000272.s003 (0.10 MB DOC)

Table S2 Pan-DENV sequences, entropy and representation of variants.
Found at: doi:10.1371/journal.pntd.0000272.s004 (0.08 MB DOC)

Table S3 Functional and structural properties of pan-DENV sequences.
Found at: doi:10.1371/journal.pntd.0000272.s005 (0.06 MB DOC)

Table S4 Distribution of pan-DENV sequences in nature.
Found at: doi:10.1371/journal.pntd.0000272.s006 (0.12 MB DOC)

Table S5 Candidate putative HLA supertype-restricted binding nonamer peptides in pan-DENV sequences, predicted by immunoinformatic algorithms.
Found at: doi:10.1371/journal.pntd.0000272.s007 (0.20 MB DOC)

Table S6 Intra-type representation of candidate putative HLA supertype-restricted nonamer peptides predicted by immunoinformatics algorithms.
Found at: doi:10.1371/journal.pntd.0000272.s008 (0.22 MB DOC)

Dataset S1 GI numbers.
Found at: doi:10.1371/journal.pntd.0000272.s009 (0.86 MB XLS)

Alternative Language Abstract S1 Translation of the abstract into Chinese by Guang Lan Zhang.
Found at: doi:10.1371/journal.pntd.0000272.s010 (0.06 MB PDF)

Acknowledgments

The authors are grateful to Lars Fugger (Weatherall Institute of Molecular Medicine, Oxford, UK) and Arthur Vanderbark (Oregon Health and Science University, Portland), Chella S. David (Mayo Clinic, Rochester), and Grete Sønderstrup (Stanford University School of Medicine) for providing HLA-DR2, -DR3 and -DR4 Tg mice, respectively. Overlapping DENV-3 peptide arrays were obtained through by the NIH Biodefense and Emerging Infectious Disease Research Resources Repository, NIAID, NIH. The authors thank Paul Nordstrom August, T. Jahan, and Aslam Khan for their valuable suggestions and help with the illustrations. The authors are grateful to Yu Jianshi for his help in translating the abstract to Chinese.

Author Contributions

Conceived and designed the experiments: AMK OM EJMN KNS ATH JS JTA. Performed the experiments: AMK OM EJMN. Analyzed the data: AMK OM EJMN KNS ATH GLZ ETM TWT VB JS JTA. Contributed reagents/materials/analysis tools: AMK OM EJMN GLZ JTA. Wrote the paper: AMK OM JS JTA.

References

1. Rico-Hesse R (2003) Microevolution and virulence of dengue viruses. Adv Virus Res 59: 315–341.
2. Holmes EC (2006) The evolutionary biology of dengue virus. Novartis Found Symp 277: 177–187; discussion 187–192, 251–173.
3. Mackenzie JS, Gubler DJ, Petersen LR (2004) Emerging flaviviruses: the spread and resurgence of Japanese encephalitis, West Nile and dengue viruses. Nat Med 10: S98–109.
4. Esser MT, Marchese RD, Kierstead LS, Tussey LG, Wang F, et al. (2003) Memory T cells and vaccines. Vaccine 21: 419–430.
5. Zinkernagel RM, Hengartner H (2004) On immunity against infections and vaccines: credo 2004. Scand J Immunol 60: 9–13.
6. Pulendran B, Ahmed R (2006) Translating innate immunity into immunological memory: implications for vaccine development. Cell 124: 849–863.
7. Rothman AL (2004) Dengue: defining protective versus pathologic immunity. J Clin Invest 113: 946–951.
8. Livingston PG, Kurane I, Dai LC, Okamoto Y, Lai CJ, et al. (1995) Dengue virus-specific, HLA-B35-restricted, human CD8+ cytotoxic T lymphocyte (CTL) clones. Recognition of NS3 amino acids 500 to 508 by CTL clones of two different serotype specificities. J Immunol 154: 1287–1295.
9. Sloan-Lancaster J, Allen PM (1996) Altered peptide ligand-induced partial T cell activation: molecular mechanisms and role in T cell biology. Annu Rev Immunol 14: 1–27.
10. Welsh RM, Rothman AL (2003) Dengue immune response: low affinity, high febrility. Nat Med 9: 820–822.
11. Mongkolsapaya J, Duangchinda T, Dejnirattisai W, Vasanawathana S, Avirutnan P, et al. (2006) T cell responses in dengue hemorrhagic fever: are cross-reactive T cells suboptimal? J Immunol 176: 3821–3829.
12. Evavold BD, Sloan-Lancaster J, Allen PM (1993) Tickling the TCR: selective T-cell functions stimulated by altered peptide ligands. Immunol Today 14: 602–609.
13. Madrenas J, Germain RN (1996) Variant TCR ligands: new insights into the molecular basis of antigen-dependent signal transduction and T-cell activation. Semin Immunol 8: 83–101.
14. Kalergis AM, Nathenson SG (2000) Altered peptide ligand-mediated TCR antagonism can be modulated by a change in a single amino acid residue within the CDR3 beta of an MHC class I-restricted TCR. J Immunol 165: 280–285.
15. Nishimura Y, Chen YZ, Uemura Y, Tanaka Y, Tsukamoto H, et al. (2004) Degenerate recognition and response of human CD4+ Th cell clones: implications for basic and applied immunology. Mol Immunol 40: 1089–1094.
16. Loke H, Bethell DB, Phuong CX, Dung M, Schneider J, et al. (2001) Strong HLA class I–restricted T cell responses in dengue hemorrhagic fever: a double-edged sword? J Infect Dis 184: 1369–1373.
17. Mongkolsapaya J, Dejnirattisai W, Xu XN, Vasanawathana S, Tangthawornchaikul N, et al. (2003) Original antigenic sin and apoptosis in the pathogenesis of dengue hemorrhagic fever. Nat Med 9: 921–927.
18. Sette A, Livingston B, McKinney D, Appella E, Fikes J, et al. (2001) The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation. Biologicals 29: 271–276.
19. Sette A, Sidney J (1999) Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics 50: 201–212.
20. Khan AM, Miotto O, Heiny AT, Salmon J, Srinivasan KN, et al. (2006) A systematic bioinformatics approach for selection of epitope-based vaccine targets. Cell Immunol 244: 141–147.
21. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33: D39–45.
22. Osatomi K, Sumiyoshi H (1990) Complete nucleotide sequence of dengue type 3 virus genome RNA. Virology 176: 643–647.
23. McGinnis S, Madden TL (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32: W20–25.
24. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876–4882.

25. Novitsky V, Smith UR, Gilbert P, McLane MF, Chigwedere P, et al. (2002) Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? J Virol 76: 5435–5451.
26. Rammensee HG, Friede T, Stevanoviic S (1995) MHC ligands and peptide motifs: first listing. Immunogenetics 41: 178–228.
27. Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27: 379–423 and 623–656.
28. Miotto O, Heiny A, Tan TW, August JT, Brusic V (2008) Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis. BMC Bioinformatics 9 Suppl 1: S18.
29. Paninski L (2003) Estimation of entropy and mutual information. Neural Computation 15: 1191–1253.
30. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, et al. (2006) The PROSITE database. Nucleic Acids Res 34: D227–230.
31. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, et al. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34: W362–365.
32. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138–141.
33. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
34. Peters B, Sidney J, Bourne P, Bui HH, Buus S, et al. (2005) The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol 3: e91. doi:10.1371/journal.pbio.0030091.
35. Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, et al. (2005) An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. Eur J Immunol 35: 2295–2303.
36. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V (2005) MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucleic Acids Res 33: W172–179.
37. Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, et al. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics 57: 304–314.
38. Bian H, Hammer J (2004) Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE. Methods 34: 468–475.
39. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, et al. (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol 17: 555–561.
40. Sette A, Sidney J, Livingston BD, Dzuris JL, Crimi C, et al. (2003) Class I molecules with similar peptide-binding specificities are the result of both common ancestry and convergent evolution. Immunogenetics 54: 830–841.
41. Vandenbark AA, Rich C, Mooney J, Zamora A, Wang C, et al. (2003) Recombinant TCR ligand induces tolerance to myelin oligodendrocyte glycoprotein 35-55 peptide and reverses clinical and histological signs of chronic experimental autoimmune encephalomyelitis in HLA-DR2 transgenic mice. J Immunol 171: 127–133.
42. Strauss G, Vignali DA, Schonrich G, Hammerling GJ (1994) Negative and positive selection by HLA-DR3(DRw17) molecules in transgenic mice. Immunogenetics 40: 104–108.
43. Madsen L, Labrecque N, Engberg J, Dierich A, Svejgaard A, et al. (1999) Mice lacking all conventional MHC class II genes. Proc Natl Acad Sci U S A 96: 10338–10343.
44. Ito K, Bian HJ, Molina M, Han J, Magram J, et al. (1996) HLA-DR4-IE chimeric class II transgenic, murine class II-deficient mice are susceptible to experimental allergic encephalomyelitis. J Exp Med 183: 2635–2644.
45. Fugger L, Michie SA, Rulifson I, Lock CB, McDevitt GS (1994) Expression of HLA-DR4 and human CD4 transgenes in mice determines the variable region beta-chain T-cell repertoire and mediates an HLA-DR-restricted immune response. Proc Natl Acad Sci U S A 91: 6151–6155.
46. Cope AP, Patel SD, Hall F, Congia M, Hubers HA, et al. (1999) T cell responses to a human cartilage autoantigen in the context of rheumatoid arthritis-associated and nonassociated HLA-DR4 alleles. Arthritis Rheum 42: 1497–1507.
47. Roederer M, Koup RA (2003) Optimized determination of T cell epitope responses. J Immunol Methods 274: 221–228.
48. Khan AM, Heiny AT, Lee KX, Srinivasan KN, Tan TW, et al. (2006) Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus. BMC Bioinformatics 7 Suppl 5: S4.
49. Holmes EC, Burch SS (2000) The causes and consequences of genetic variation in dengue virus. Trends Microbiol 8: 74–77.
50. Valdar WS (2002) Scoring residue conservation. Proteins 48: 227–241.
51. Allison SL, Schalich J, Stiasny K, Mandl CW, Heinz FX (2001) Mutational evidence for an internal fusion peptide in flavivirus envelope protein E. J Virol 75: 4268–4275.
52. Modis Y, Ogata S, Clements D, Harrison SC (2004) Structure of the dengue virus envelope protein after membrane fusion. Nature 427: 313–319.
53. Murthy HM, Clum S, Padmanabhan R (1999) Dengue virus NS3 serine protease. Crystal structure and insights into interaction of the active site with substrates by molecular modeling and structural analysis of mutational effects. J Biol Chem 274: 5573–5580.
54. Xu T, Sampath A, Chao A, Wen D, Nanao M, et al. (2005) Structure of the Dengue virus helicase/nucleoside triphosphatase catalytic domain at a resolution of 2.4 A. J Virol 79: 10278–10288.
55. Egloff MP, Benarroch D, Selisko B, Romette JL, Canard B (2002) An RNA cap (nucleoside-2′-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. Embo J 21: 2757–2768.
56. Yap TL, Xu T, Chen YL, Malet H, Egloff MP, et al. (2007) Crystal structure of the dengue virus RNA-dependent RNA polymerase catalytic domain at 1.85-angstrom resolution. J Virol 81: 4753–4765.
57. Haydon DT, Woolhouse ME (1998) Immune avoidance strategies in RNA viruses: fitness continuums arising from trade-offs between immunogenicity and antigenic variability. J Theor Biol 193: 601–612.
58. Crochu S, Cook S, Attoui H, Charrel RN, De Chesse R, et al. (2004) Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes. J Gen Virol 85: 1971–1980.
59. Henchal EA, Putnak JR (1990) The dengue viruses. Clin Microbiol Rev 3: 376–396.
60. Kuno G, Chang GJ, Tsuchiya KR, Karabatsos N, Cropp CB (1998) Phylogeny of the genus Flavivirus. J Virol 72: 73–83.
61. Billoir F, de Chesse R, Tolou H, de Micco P, Gould EA, et al. (2000) Phylogeny of the genus flavivirus using complete coding sequences of arthropod-borne viruses and viruses with no known vector. J Gen Virol 81: 781–790.
62. Gagnon SJ, Zeng W, Kurane I, Ennis FA (1996) Identification of two epitopes on the dengue 4 virus capsid protein recognized by a serotype-specific and a panel of serotype-cross-reactive human CD4+ cytotoxic T-lymphocyte clones. J Virol 70: 141–147.
63. Wilson CC, McKinney D, Anders M, MaWhinney S, Forster J, et al. (2003) Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1. J Immunol 171: 5611–5623.
64. Gupta V, Tabiin TM, Sun K, Chandrasekaran A, Anwar A, et al. (2006) SARS coronavirus nucleocapsid immunodominant T-cell epitope cluster is common to both exogenous recombinant and endogenous DNA-encoded immunogens. Virology 347: 127–139.
65. Berzofsky JA, Pendleton CD, Clerici M, Ahlers J, Lucey DR, et al. (1991) Construction of peptides encompassing multideterminant clusters of human immunodeficiency virus envelope to induce in vitro T cell responses in mice and humans of multiple MHC types. J Clin Invest 88: 876–884.
66. Shankar P, Fabry JA, Fong DM, Lieberman J (1996) Three regions of HIV-1 gp160 contain clusters of immunodominant CTL epitopes. Immunol Lett 52: 23–30.
67. Surman S, Lockey TD, Slobod KS, Jones B, Riberdy JM, et al. (2001) Localization of CD4+ T cell epitope hotspots to exposed strands of HIV envelope glycoprotein suggests structural influences on antigen processing. Proc Natl Acad Sci U S A 98: 4587–4592.
68. Brown SA, Stambas J, Zhan X, Slobod KS, Coleclough C, et al. (2003) Clustering of Th cell epitopes on exposed regions of HIV envelope despite defects in antibody activity. J Immunol 171: 4140–4148.
69. Kim SK, DeMars R (2001) Epitope clusters in the major outer membrane protein of Chlamydia trachomatis. Curr Opin Immunol 13: 429–436.
70. Moran E, Simmons C, Vinh Chau N, Luhn K, Wills B, et al. (2008) Preservation of a critical epitope core region is associated with the high degree of flaviviral cross-reactivity exhibited by a dengue-specific CD4(+) T cell clone. Eur J Immunol 38: 1050–1057.
71. Whitehead SS, Blaney JE, Durbin AP, Murphy BR (2007) Prospects for a dengue virus vaccine. Nat Rev Microbiol 5: 518–528.
72. Mangada MM, Rothman AL (2005) Altered cytokine responses of dengue-specific CD4+ T cells to heterologous serotypes. J Immunol 175: 2676–2683.
73. Mota J, Acosta M, Argotte R, Figueroa R, Mendez A, et al. (2005) Induction of protective antibodies against dengue virus by tetravalent DNA immunization of mice with domain III of the envelope protein. Vaccine 23: 3469–3476.
74. Chin JF, Chu JJ, Ng ML (2007) The envelope glycoprotein domain III of dengue virus serotypes 1 and 2 inhibit virus entry. Microbes Infect 9: 1–6.
75. Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, et al. (2007) Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. PLoS ONE 2: e1190. doi:10.1371/journal.-pone.0001190.
76. Simmons CP, Dong T, Chau NV, Dung NT, Chau TN, et al. (2005) Early T-cell responses to dengue virus epitopes in Vietnamese adults with secondary dengue virus infections. J Virol 79: 5665–5675.
77. Kurane I, Dai LC, Livingston PG, Reed E, Ennis FA (1993) Definition of an HLA-DPw2-restricted epitope on NS3, recognized by a dengue virus serotype-cross-reactive human CD4+ CD8- cytotoxic T-cell clone. J Virol 67: 6285–6288.
78. Okamoto Y, Kurane I, Leporati AM, Ennis FA (1998) Definition of the region on NS3 which contains multiple epitopes recognized by dengue virus serotype-cross-reactive and flavivirus-cross-reactive, HLA-DPw2-restricted CD4+ T cell clones. J Gen Virol 79 (Pt 4): 697–704.