Lecture Notes in Mathematics 1922

Editor

Avner Friedman

Mathematical Biosciences Institute
Ohio State University
231 West 18th Avenue
Columbus, OH 43210-1292
USA

e-mail: afriedman@math.ohio-state.edu
afriedman@mbi.ohio-state.edu

Library of Congress Control Number: 2007933684

Mathematics Subject Classification (2000): 35J20, 35J60, 35K55, 35K57, 62P10, 62P12, 92B99, 92D10, 92D15, 92D25, 92D40

ISSN print edition: 0075-8434
ISSN electronic edition: 1617-9692
ISBN 978-3-540-74328-6 Springer Berlin Heidelberg New York
DOI 10.1007/978-3-540-74331-6

Typesetting by the author and SPi using a Springer LaTeX macro package

Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 12109630 41/SPi 5 4 3 2 1 0

2

Large-Scale Phylogenetic Analysis of Emerging Infectious Diseases

D. Janies[1] and D. Pol[1,2]

[1] Department of Biomedical Informatics,
The Ohio State University, Columbus, OH 43210, USA
email: Daniel.Janies@osumc.edu
[2] Mathematical Biosciences Institute,
The Ohio State University, Columbus, OH 43210, USA
email: dpol@mbi.osu.edu

Summary. Microorganisms that cause infectious diseases present critical issues of national security, public health, and economic welfare. For example, in recent years, highly pathogenic strains of avian influenza have emerged in Asia, spread through Eastern Europe, and threaten to become pandemic. As demonstrated by the coordinated response to Severe Acute Respiratory Syndrome (SARS) and influenza, agents of infectious disease are being addressed via large-scale genomic sequencing. The goal of genomic sequencing projects are to rapidly put large amounts of data in the public domain to accelerate research on disease surveillance, treatment, and prevention. However, our ability to derive information from large comparative genomic datasets lags far behind acquisition. Here we review the computational challenges of comparative genomic analyses, specifically sequence alignment and reconstruction of phylogenetic trees. We present novel analytical results on two important infectious diseases, Severe Acute Respiratory Syndrome (SARS) and influenza.

SARS and influenza have similarities and important differences both as biological and comparative genomic analysis problems. Influenza viruses (Orthymxyoviridae) are RNA based. Current evidence indicates that influenza viruses originate in aquatic birds from wild populations. Influenza has been studied for decades via well-coordinated international efforts. These efforts center on surveillance via antibody characterization of the hemagglutinin (HA) and neuraminidase (N) proteins of the circulating strains to inform vaccine design. However, we still do not have a clear understanding of (1) various transmission pathways such as the role of intermediate hosts like swine and domestic birds and (2) the key mutation and genomic recombination events that underlie periodic pandemics of influenza. In the past 30 years, sequence data from HA and N loci has become an important data type. In the past year, full genomic data has become prominent. These data present exciting opportunities to address unanswered questions in influenza pandemics.

SARS is caused by a previously unrecognized lineage of coronavirus, SARS-CoV, which like influenza has an RNA based genome. Although SARS-CoV is widely believed to have originated in animals, there remains disagreement over the candidate animal source that lead to the original outbreak of SARS. In contrast to the long history of the study of influenza, SARS was only recognized in late 2002 and the virus that causes SARS has been documented primarily by genomic sequencing.

In the past, most studies of influenza were performed on a limited number of isolates and genes suited to a particular problem. Major goals in science today are to understand emerging diseases in broad geographic, environmental, societal, biological, and genomic contexts. Synthesizing diverse information brought together by various researchers is important to find out what can be done to prevent future outbreaks [JON03]. Thus comprehensive means to organize and analyze large amounts of diverse information are critical. For example, the relationships of isolates and patterns of genomic change observed in large datasets might not be consistent with hypotheses formed on partial data. Moreover when researchers rely on partial datasets, they restrict the range of possible discoveries.

Phylogenetics is well suited to the complex task of understanding emerging infectious disease. Phylogenetic analyses can test many hypotheses by comparing diverse isolates collected from various hosts, environments, and points in time and organizing these data into various evolutionary scenarios. The products of a phylogenetic analysis are a graphical tree of ancestor–descendent relationships and an inferred summary of mutations, recombination events, host shifts, geographic, and temporal spread of the viruses. However, this synthesis comes at a price. The cost of computation of phylogenetic analysis expands combinatorially as the number of isolates considered increases. Thus, large datasets like those currently produced are commonly considered intractable. We address this problem with synergistic development of heuristics tree search strategies and parallel computing.

2.1 Introduction

Phylogenetics is the study of the evolutionary relationships of genes and organisms, thus providing a retrospective analysis of biological change and adaptation over time. Phylogenetic trees are represented by acyclic graphs in which the leaves of these graphs represent the observed biological entities (taxa) being compared (e.g., sequences of genes, genomes, and/or anatomy of individuals, isolates or cultivars, species, or any higher level taxonomic unit). The internal nodes of the tree are interpreted as a nested set of hypothetical evolutionary ancestors of the entities under consideration as depicted in Fig. 2.1. Once a tree is complete, changes such as mutations and host shift can be traced along branches of the tree that contain important disease causing strains. This retrospective analysis of features provides means of finding mutations that are diagnostic of pathogens, correlating phenotypes and genotypes, and predicting strains that are important for vaccine design.

2.1.1 Modern School of Phylogenetics

The classification of organisms dates back to Aristotle [ARI343]. However, it was only a few decades ago that the theoretical foundations of the field of phylogenetics as it is practiced today were established.

The modern school of phylogenetics arose from the application of the ideas, termed cladistics, originally proposed by Hennig [HEN66]. Cladistics lead

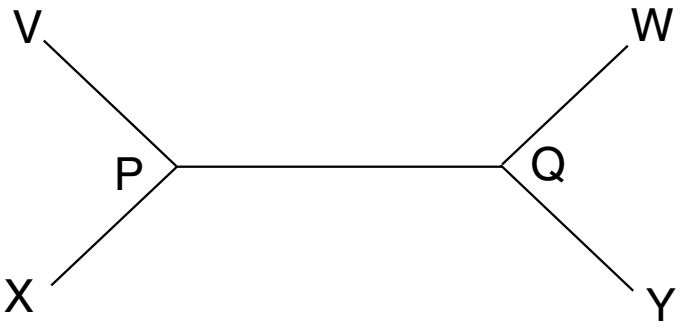

Fig. 2.1. Phylogenetic tree of four taxa labeled V, W, X, and Y and two hypothetical ancestors labeled P and Q

biologists to use shared derived similarities (termed synapomorphies) that distinguish various natural groups of organisms. Nested sets of natural groups of organisms based on synapomorphies are then used to discover the evolutionary relationships between organisms and reconstruct patterns of modification in the features of organisms. Subsequently, these principles have been used to develop optimization techniques to find the most justifiable sets of synapomorphies in large datasets. Optimization techniques are necessary with most large and real world datasets as they often contain several, often conflicting, evolutionary signals (treated below).

In contrast, advocates of another way of thinking, termed phenetics [SNE73], group organisms based on gross measures of similarity. Groups are based on measures evolutionary distance rather than the concept of shared derived characters. In modern practice, similarity methods espousing phenetic concepts are used in searches of nucleotide databases and some multiple alignment methods. Clustering algorithms in which least distant groups are clustered first and then distant clusters are connected are termed distance methods, in phylogenetics. Distance methods typically produce a single tree and cannot, on their own, trace patterns of change in the features of organisms as they convert raw data to distances. Next we discuss how various viewpoints have influenced methods, algorithms, and implementations in phylogenetics.

2.1.2 Phylogenetic Methods

A wide variety of methods have been proposed in order to infer the phylogenetic relationships of organisms. Most methods are based on minimizing edit cost (such as a Hamming distance) to transform one string of nucleotides or organismal characters into another. Phylogenetic methods can be further classified in two different categories: distance-based and character-based. In this paper, we compare and contrast the applications of distance and character-based methods used in infectious disease research. We illustrate applications

of these technique to study the evolutionary relationships of groups of RNA viruses and the patterns of mutations and phenotypic changes that can be reconstructed.

Distance-based Methods

Among the distance-based methods, the most commonly used is Neighbor-Joining [SAI87]. Distance based methods require a precomputed multiple alignment of DNA or amino acid sequences drawn from homologous genes. The most similar pair of taxa (as represented by sequences) are clustered. The clustered pair is then considered as a single taxon and the next most similar pair of taxa is clustered until only the last taxon is joined and the tree is completed. Although the use of distance-based methods is relatively common in analysis of organisms that cause infectious disease, several authors have criticized the performance of this method for phylogenetic reconstruction (see [FAR96]). One strategic flaw of the method is that it is computationally greedy. Distance methods form the most similar clusters instantaneously without considering locally suboptimal paths that may lead to a better global optimum.

Character-based Methods

Other methods of phylogenetic analysis focus on characters, which are typically polymorphisms, recognized in columns of aligned nucleotides or amino acids from sequences of interest or investigator encoded characters of polymorphic phenotypes.

Character-based methods seek to find the phylogenetic trees that optimize a particular criterion. Major optimality criteria include parsimony [FAR83] and maximum likelihood [FEL73, FEL81]. Bayesian analysis [RAN96, LI00, HUE02] uses a maximum likelihood optimality criterion but incorporates the probability, termed the prior, that a hypothesis is correct in the absence of data.

The unifying feature of character-based methods is that they examine many randomly generated trees each representing an evolutionary hypothesis of character transformations and organismal relationships. As a character based analysis progresses, edit costs are calculated for transformations that candidate tree imply and optimal trees are stored for further consideration and refinement. The concept of optimality can be associated with cladistics or maximum likelihood but not distance methods. Distance techniques lack a measure of tree quality and means to compare trees.

Cladistics employs parsimony as an optimality criterion. The core concept of cladistics is that the least number of transformations in the data implies the most defensible hypothesis. In cladistics, various edit costs can be applied to different genomic and phenotypic transformations. In the case of weighted parsimony the goal of tree search is to minimize weighted costs.

Under the maximum likelihood criterion the probability of the data, given the tree, calculated with a model for nucleotide or amino acid substitution is optimized. The related technique of Bayesian phylogenetic inference uses maximum likelihood to evaluate trees. Bayesian analysis aims to capture a posterior probability distribution of trees. Typically the results of a Bayesian analysis are displayed not as an optimal tree but rather as the probability that a set of evolutionary relationships is "true," given the prior probabilities, the substitution model, and data.

All character-based methods of molecular phylogenetics [cladistics, maximum likelihood (and related Bayesian methods)] rely on explicit assumptions about ancestral character states to polarize transformations of phenotypes and genotypes that can be reconstructed from data. As an example of such assumptions, in character based analyses is the outgroup criterion (treated below). In contrast, distance based methods do not use an outgroup criterion. Distance based methods do not use the outgroup criterion.

Parsimony

Parsimony is a widely used optimality criterion. This criterion is associated with the concept that simpler explanations provide for more supportable hypotheses. In phylogenetics, the most parsimonious tree(s) is that which implies the minimum number of transformations in sequence and/or phenotypic character states among organisms of interest. The biological justification of this use of parsimony is that descent with modification from a common ancestor is a primary pattern of organismal diversification and the record of transformations can be used to reconstruct that pattern. As such, the tree(s) that minimizes the overall number of independent transformations (convergences or reversals in character state) that are needed to explain the observed data are to be preferred [FAR83]. Recombination and horizontal gene transfer as seen among RNA viruses are violation of the assumption of ancestor to descendent evolution, not parsimony per se. Some novel techniques for discovery and understanding of reassortment and horizontal gene transfer have been developed under the parsimony criterion [WAN05, WHE05].

The parsimony score for a tree is measured based on the number of transformations implied by the tree, known as the *tree length* [FAR70]. The tree length is the sum, over all edges, of the Hamming distances between the labels at the endpoints of the edge [RIC97]. The labels located at the leaves of the tree are the observed characteristics (either genotypic of phenotypic) of the organisms being analyzed. The internal nodes are labeled in order to minimize the tree length of each tree being evaluated.

Given a tree and a matrix of features or aligned sequences for each taxon, the tree length is calculated using the Fitch algorithm [FIT71]. This algorithm works in polynomial time with the amount of data being analyzed (both in the number of characters and taxa). Thus for a sequence alignment of thousands of taxa, each of which is labeled with thousands of nucleotides, the tree length

of *a particular tree* can be computed using modern implementations of the Fitch algorithm [GOL03] in fractions of a second.

Inferring Evolutionary Events on a Tree

Given a tree and a data matrix of sequences and features, the parsimony method can pinpoint the branches on which certain evolutionary events are inferred to occur between ancestor or descendent. In an infectious disease context, these events can be a shift by a viral lineage from animal to human host. In the case of standard nucleotide sequence analysis, transformation events include substitution mutations (replacement of a given nucleotide by other) and nucleotide insertions and deletion mutations. Some analyses invoke more complex parsimony models with weighted recombination and horizontal transfer events, as well as differentially weighting certain classes of mutation such as transversions (pyrimidine–purine shifts), transitions (pyrimidine–pyrimidine or purine–purine shifts), or insertion–deletion events [WHE05].

Note that in using the Fitch algorithm to optimize a phylogenetic tree, both the tree length and the branch in which a particular transformation event is inferred to occur can be calculated in unrooted or rooted trees. The results of these calculations are independent of the root chosen for the tree. For example, in an unrooted tree relating four taxa known from their nucleotide sequences (see Fig. 2.2), the Fitch algorithm can be used to identify a specific branch of a tree in which a transformation occurs (e.g., a mutation between nucleotides C and T of the third sequence position occurring in the only internal edge of the tree in Fig. 2.3).

However, the polarity of a transformation event is dependent on how the tree is rooted. Inferring polarity of change requires an external criterion, termed the outgroup.

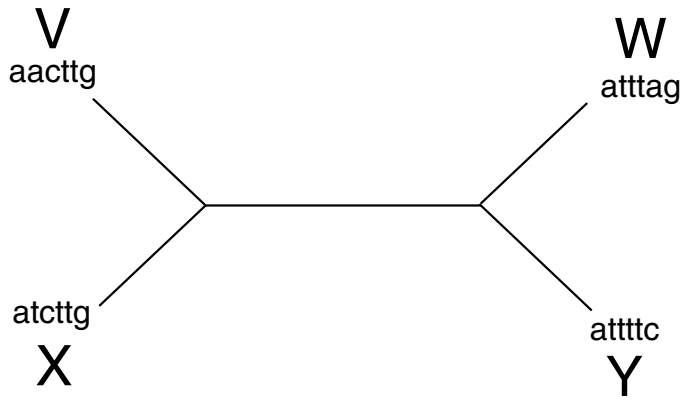

Fig. 2.2. Phylogenetic tree of four taxa V, W, X, and Y as in Fig. 2.1 but with the addition of nucleotide sequences observed for each taxon

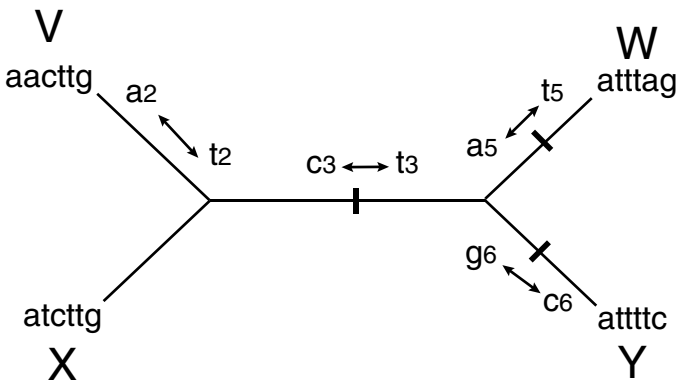

Fig. 2.3. Unrooted phylogenetic tree of four taxa and observed nucleotide sequences as in Fig. 2.2. Here in Fig. 2.3, mutations inferred on various branches are indicated by the nucleotide state and sequence position (in subscript). The number of mutations is four thus the tree length would be four. However, the polarity of mutations cannot be inferred, hence the bidirectional arrows

Polarity of Change and the Outgroup Criterion

In character-based methods explicit assumptions of ancestral character states are set up by the investigator via the designation of at least one taxon as the outgroup. A good outgroup is known to be closely related to the taxa of interest (termed the ingroup). However, the outgroup must be clearly not a member of the ingroup. The underlying logic of the outgroup criterion is that the transformation events that occurred at evolutionary origin of the ingroup can be identified by comparison to modern organisms of another clade but with which the ingroup shares a common ancestor. The common ancestor is a hypothetical organism that provides a baseline set of character states from which polarity determinations can be made. Thus the outgroup method, like Bayesian inference, incorporates some previous knowledge of the relationships of the organisms. If the phylogenetic results show that the ingroup includes some members of the outgroup the previous knowledge must be reevaluated.

The outgroup taxon is included in the data matrix of the phylogenetic analysis and the entire data set is analyzed simultaneously. The phylogenetic position and relationships of the outgroup are determined by the optimality criterion. In the case of the parsimony method, the outgroup is treated as any other taxon and is positioned in the tree in the position that minimized tree length. Once the phylogenetic affinities outgroup are established, the outgroup can be used to root the tree, and the polarities of the transformations can be established (note the unidirectional arrows in Fig. 2.4). If chosen carefully, the outgroup will not be clustered with any of the ingroup. Model based methods can also be used in reconstruction of ancestral character states (e.g., [CHA00, THR04]).

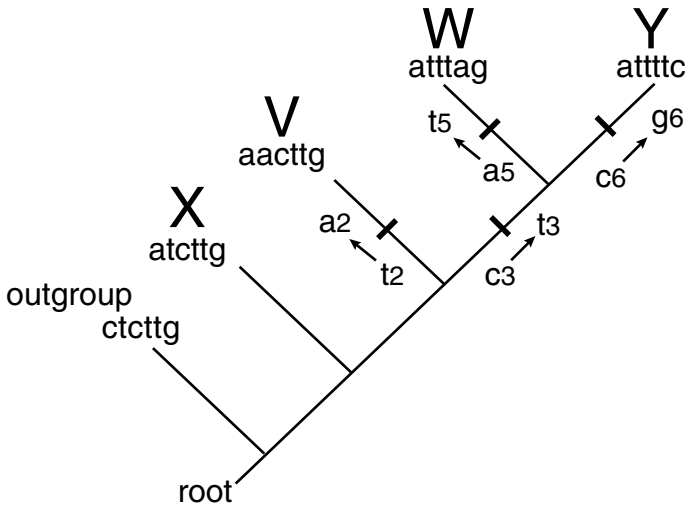

Fig. 2.4. Rooted phylogenetic tree of five taxa. Four of the taxa are the same as in Figs. 2.2 and 2.3 but here we add an outgroup and polarize the mutations, hence the unidirectional arrows. Labels at the leaves of the tree are the observed nucleotide sequences (see Fig. 2.2). Mutations are marked on the branches where they are inferred to occur

To communicate the choice of outgroup taxon or taxa and clarify the relationships of the taxa, character based trees are often drawn as directed acyclic graphs with the root positioned on the branch between the outgroup and ingroup.

In the example diagrammed in Fig. 2.4, a mutation is inferred to occur in the third sequence position from the ancestral state of C to the derived state of T. The presence of a T in the third sequence position is a synapomorphy, a derived character state that can be used to distinguish the members of the group formed by the taxa W and Y. In contrast, the presence of a C in the third sequence position in taxa X and V cannot be used to distinguish these taxa since a C is also present in the third position in the outgroup. In this case the third position C is a primitive similarity of X and Y or a symplesiomorphy. The other mutations occurring in sequence positions 2, 5, and 6 are found only in one taxon and thus cannot be used to infer relationships. These are termed autapomorphies. Sequence position 4 is inferred to have not changed in this example and is thus of no value in discovering groups. In cladistics only the shared derived characteristics, synapomorphies, are used to diagnose a group.

Although other criteria have been proposed to root phylogenetic trees, the outgroup criterion is the least arbitrary. As a result, outgroup rooting is widely used for character-based phylogenetic analyses [NIX94].

Problems of the Outgroup Criterion. As seen, the use of the outgroup taxon provides an informative way test hypotheses on the content of natural groups and to root the phylogenetic tree in a way that allows interpretation of the polarity of change of evolutionary events.

However, the choice of an outgroup taxon is key to the success of this method. If the nucleotide sequences of a candidate outgroup are divergent from the sequences of the ingroup taxa, the phylogenetic position of the outgroup might be hard to establish [WHE90]. Therefore, the choice of the outgroup requires judicious selection and searches for organisms that (1) are safely outside the ingroup but (2) that have comparable data [WHE90].

2.1.3 Sequence Alignment

The cases shown in Figs. 2.1–2.4 are based on the simplifying assumption that the genes sequenced for the taxa of interest have equal number of residues (i.e., amino acid or nucleotide sequences of the same length).

Frequently, in empirical studies of related organisms, homologous genes have sequences with different number of residues. Sequence length variation occurs in both coding and noncoding loci. The causes can be genetic drift, mutation, recombination, or horizontal transfer events. The phylogenetic analysis of molecular sequences, like that of all other comparative data, is based on schemes of putative homology that are then tested via phylogenetic analysis. Unlike some other data types, however, putative homologies in molecular data are not directly observable. Sequences from various organisms are often unequal in length. Hence, the correspondences among sequence positions are not evident and some sort of procedure is required to determine which regions are homologous. This procedure is typically multiple sequence alignment. Alignment inserts gaps to make the putatively corresponding residue line up into columns. These columns (characters) comprise the matrix used to reconstruct cladograms. The matrix is then submitted to phylogenetic analysis in the same manner as other forms of data such as morphological characters scored by an investigator. Thus the primary reason in phylogenetics to create an alignment has a strongly operational basis – to make it possible to submit these data to standard phylogeny programs that were designed to handle column vectors of morphological characters. Nevertheless, alignment followed by tree search is the standard procedure.

Two major options are currently available to analyze sequence data in a phylogenetic framework: a twostep analysis or a one-step analysis.

Two-Step Analyses

Phylogenetic analysis of large genomic datasets can present several nested NP-complete problems: multiple alignment, tree-search, and in some cases, gene order and complement differences among organisms. Just as in distance methods, in most character-based methods, alignments are precomputed before any

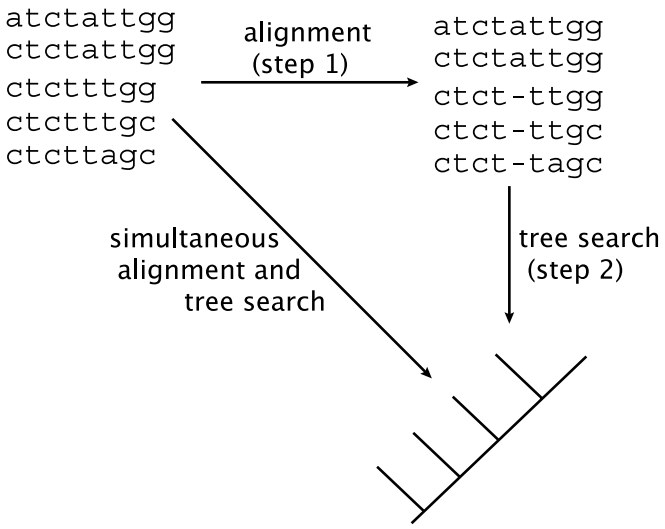

Fig. 2.5. One-step and two-step procedures for the analysis of DNA sequences with different number of nucleotides showing the analysis from raw DNA sequences of different length to the inferred optimal tree

phylogenetic analysis. The alignment procedure is usually done through algorithms that produce a matrix from the raw DNA sequences of the organisms being analyzed (Fig. 2.5). This data set is then analyzed (second step) in order to find the optimal tree (see Sect. 2.1.4).

The multiple alignment procedure ranges from easy in many coding loci to very difficult in noncoding loci such as functional RNAs and genes containing introns [MOR97]. In the case of some protein coding loci the alignment may be a nonissue if there are no significant length differences in sequences. However, various investigators who employ different primer sets and editing styles often produce various length sequences. Leading and trailing gaps produced by experimental artifact should not be counted in tree length calculations.

Results of multiple alignment of functional RNAs and genes containing introns can be sensitive to parameter choices [FIT83]. Important parameters include the addition order of taxa, relative costs of various classes of mutations (transversions, transitions, insertion-deletion), and differential costs applied to opening or extending regions of insertion–deletions. Analyses of different alignments of the same raw sequences can lead to different trees irrespective of tree search procedures [MOR97]. In such cases investigators must search parameter space [WHE95, PHI00] or otherwise justify their assumptions during alignment [GRA03] just as they are required to justify optimality criteria used during tree search.

One-Step Parsimony Analysis

Several researchers have noted that performing phylogenetic analysis into two steps is not consistent with the goals of finding the most parsimonious solutions due to the interdependence of multiple alignment and tree estimation [PHI00, JAN02]. In fact, popular multiple alignment programs such as CLUSTAL [THO94] use a guide tree used to construct the alignment. Therefore, methods have been proposed to make a simultaneous estimation of the optimal sequence alignment and the optimal phylogenetic tree [SAN83]. A modern implementation of the one-step concept in POY [WHE05], termed direct optimization, allows unaligned sequence data to be analyzed without precomputing an alignment. In direct optimization, sequence data are aligned as various trees are built and their optimality is assessed. Thus for each tree considered in a search, various sets of homology statements for the sequence data are considered. One advantage of direct optimization is that the outgroup need not be designated by the investigator. POY allows for randomization of the outgroup taxon and thus adds rigor to the search for optimal trees and homology statements. In some implementations of character-based methods where prealignment is necessary the outgroup can be randomized by scripting a series of analyses, e.g., TNT [GOL03]. One important difference is that in molecular data a rigorous tree search with on a prealigned dataset with unordered characters should lead to the same tree length irrespective of outgroup choice; whereas in direct optimization the homology statements and hence tree length can be dependent on outgroup choice.

Several groups are developing algorithms for simultaneous estimation of alignment and phylogenetic trees. Methods for a one-step phylogenetic analysis have been developed using maximum likelihood [TKF92, FLE05] and parsimony optimality criteria [WHE96], as well as for Bayesian analysis [RED05].

Although the one-step approach has the appeal of using a unified and epistemologically consistent method of alignment and tree estimation, the time and space requirements for computation are considerable. This problem of tree-based alignment is known to be NP-complete [WAN94]. In this situation, genes that vary in length (as most noncoding and intronic containing genes do) present a huge number of possible hypothetical ancestral sequences even for a single binary tree. During a phylogenetic analysis many trees will be examined and compared. For s taxa and l nucleotides per taxon, the cost of computation per tree ranges from $(s-l)l^2$ to $(s-2)l^3$, depending on the heuristics applied. The memory requirements scale proportional to l^3. Fortunately, procedures such as the optimized diagonal transition algorithms described by Ukkonen[UKK85] abate the space and time dependence on l, the number of nucleotides [WHE05].

2.1.4 Tree Searches

Phylogenetic analysis under the parsimony criterion is based on an objective function (tree length). Tree length is used to evaluate the optimality of each phylogenetic tree considered. However, finding the optimal phylogenetic tree (among all possible topologies) is an NP-hard problem [FG82], that resembles the Steiner tree problem. The combinatorial optimization problem of phylogenetic analysis consists of finding the optimal solution from a very large number of possible trees. The number of possible trees increases dramatically with the number of organisms being analyzed [FEL78]. The number of possible (unrooted) phylogenetic trees (T) for a given set of organisms increases following

$$T = (2 \times s - 5)!!, \tag{2.1}$$

where s is the number of organisms (leaves) of the phylogenetic tree. Therefore, the number of possible phylogenetic trees is extremely large even for trees with moderate number of organisms.

As stated above, a phylogenetic analysis consists evaluating topologies in order to find the optimal solution [i.e., the tree(s) with the minimum length]. It is interesting to note that the computing time of an exhaustive evaluation of all possible trees for a fixed number of taxa will increase nearly linearly with the number of characters (e.g. length of DNA sequences) because the Fitch algorithm [FIT71] for evaluating the tree length of a particular topology works in polynomial time.

However, the excessively large number of possible phylogenetic trees of 20 or more organisms (see Table 2.1) makes exhaustive evaluation of all phylogenetic trees intractable.

Note on Multiple Optimal Trees. Frequently, in phylogenetic analysis based on an optimality criterion, there are multiple trees that score the same minimum

Table 2.1. Number of possible unrooted trees as a function of the number of organisms

Organisms	Trees
4	3
5	15
6	105
7	945
10	10^6
20	10^{20}
50	10^{74}
100	10^{182}
1,000	10^{2860}

Order of magnitude is given for the number of trees with more than 10 organisms

for tree length or likelihood. In the set of known optimal trees, the transformations may be differentially distributed and different organismal groups may be implied. Therefore, this set of known optimal trees must be considered equally valuable. These cases represent alternative hypotheses (i.e., phylogenetic trees) that are equally supported by the available data and can be summarized through a *consensus tree*. Several kinds of techniques for consensus estimation exist. The *strict consensus* tree is one of the most frequently used. A strict consensus calculation represents a tree that has all the edges shared by all the known optimal trees. See [SWO91] for further information on various consensus trees.

Exhaustive Searches

Two algorithms can be applied to perform exhaustive searches that evaluate (explicitly or implicitly) all possible phylogenetic trees in order to find the optimal tree. The first of these is exhaustive enumeration, which computes the optimality value (e.g., tree length) of every possible phylogenetic tree and select the tree (or trees) with the minimal value.

The second method is the branch and bound algorithm [HEN82] that implicitly evaluates all possible trees but avoids, in practice, computing all possible trees (see [SEA96]). In current phylogenetic software packages (e.g., [SWO02, GOL03]) this algorithm can be applied to data sets of up to 20 (or 25) organisms and guarantees to find the optimal trees (or trees) for a given phylogenetic data matrix.

Heuristic Searches

For analysis of data sets with larger number of organisms, the number of trees is prohibitively large for conducting an exhaustive search. In modern biology, interesting data sets consider hundreds to thousands organisms. Thus the problem of phylogenetic tree search is compute bound and must be approached through heuristic searches. In these tree searches, a large number of phylogenetic trees are evaluated and the best solution is kept as known estimate of minimum tree length.

Some early examples of heuristic tree searches include the algorithm to compute Wagner Trees [FAR70]. The Wagner algorithm creates a phylogenetic tree of three taxa and progressively adds organisms, attaching them to the branch that generates the minimal increase in tree length at that step. This stepwise procedure is conducted until the last organism is added to the tree. Although this procedure usually results in a tree that has a suboptimal tree length, in most cases this score is significantly better than that obtained with a random choice among all possible topologies. As various starting points are used for building Wagner trees, this aspect of phylogenetic analysis can be considered a type of Monte Carlo randomization.

Given one or many Wagner trees, the next standard heuristic refinement techniques that would be typically applied in tree search are known as branch swapping or hill-climbing procedures (see [SEA96]). This class of refinement procedures consists of performing minor rearrangements of branches in the starting tree. Each Wagner tree is modified by pruning a subtree and reattaching it to a different branch of the remaining tree. The tree length of the modified tree is then calculated. If the modified tree has a shorter tree length it is kept in a buffer of new candidate trees. Branch swapping is applied to all Wagner and candidate trees until the algorithm converges. When no further rearrangements can improve the current topology the branch swapping is finished.

Tree Search Strategies

The results of the branch-swapping algorithm depend on the quality of the starting point (i.e., Wagner tree). In many cases, the tree resulting from the application of branch swapping to a Wagner tree is a local optimum that cannot be further improved by swapping. Therefore, multiple replicates (100s to 1,000s) of independent Wagner trees followed by swapping are typically preformed. At the end of these stages of analysis, the best trees found in all the replicates are kept as a set representing topologies at the known minimum length.

Replication of Wagner builds plus swapping (or random-restart hill climbing) is the most widely used routine implemented in most software packages (e.g., [SWO02, GOL03]). One major drawback of Wagner builds plus swapping is that this procedure is subject to finding only local optima. Finding the globally optimal tree(s) for a dataset of >20 taxa is a NP-hard problem [FG82]. However, performing multiple replicates of this procedure can provide a relative degree of confidence if the minimum length tree(s) converge at the same tree length from numerous independent starting points [GOL99].

Replication of Wagner builds plus swapping is usually efficient for data sets smaller than a hundred taxa. Because of advances in automated DNA sequencing technology, the size of modern comparative data sets far exceed the limits for which these techniques are efficient analytical tools for phylogenetic analysis.

2.1.5 Computational Problems

Large phylogenetic problems are becoming increasingly common across the life sciences due to the prevalence of high throughput nucleotide sequencing technology. Large data sets are of interest to biologists because they provide a rich context of phenotypes and genotypes and permit worldwide and longitudinal sampling of genomes. These large phylogenetic problems will become increasingly common in the years ahead. Thus, phylogenetic methods suited to large datasets will have important consequences not only for the study of

organismal classification and evolution, but also for many aspects of public health (see Sect. 2.2). Furthermore, in a operational context, strong organismal sampling has been shown to correlate with improved performance of phylogenetic methods [HIL96, POE98, RAN98, ZWI02, HIL03].

The dauntingly high cost of computation of large-scale phylogenetic analysis stunted this line of research. However, in recent years two main lines of research have provided efficient tools to analyze large phylogenetic datasets: the development of new algorithms and the use of parallel computing.

New Algorithms

Several researchers have combined groups of algorithms into heuristic tree search strategies that have proven to be efficient for phylogenetic analyses of hundreds to thousands of organisms under the parsimony criterion. These heuristic search strategies are based on basic Monte Carlo and hill climbing techniques with the addition of other classes of algorithms including simulated annealing [GOL99], data perturbation [NIX99], divide-and-conquer [GOL99, ROS04], and genetic algorithms [MOI99, GOL99]. Similar search strategies that combine several layers of algorithms have been employed using other optimality criteria such as maximum likelihood [LEW98, SAL01, LEM02, BRA02].

The judicious application of various algorithms has provided efficient solutions for the analysis of datasets of several hundreds organisms [GOL99] in a single CPU [TEH03]. In particular, the successive combination hill-climbing, genetic, and simulated annealing algorithms of tree search have produced a drastic speed up in comparison to other strategies [GOL99]. Efficient implementations of these algorithms have become recently available in software packages [GOL03].

Parallel Computing

The need of phylogenies depicting the evolutionary relationships of datasets consisting of thousands of taxa has prompted the synergistic implementation of efficient heuristic tree search strategies and parallel computing hardware. An increasing number of researchers are developing software suited for parallel computing using Beowulf class clusters [STE00]. Beowulf clusters are simply arrays of commodity PCs and switches enabled by scalable, open source operating systems (e.g. LINUX) and message passing software (e.g. PVM or MPI). Although the advantages of parallel computing in phylogenetics and multiple alignment have been clear for some time [WHE94], the means to exploit this potential for research gain have not been broadly and economically available until the Beowulf concept was developed by the end of the 1990s.

Alignment and tree search problems are naturally suitable for parallel computing. Phylogenetic researchers quickly realized the opportunity presented by Beowulf computing [CER98, JAN01]. Finding an optimal phylogeny requires

the evaluation of the same objective function on a large number of alternative trees. Because many trees can be examined concurrently and independently, this has led several authors to implement phylogenetic tree searches in parallel [JON95, SNE00, CHA01, GOL02, BRA02, STA02]. These implementations use the parsimony and maximum likelihood optimality criteria as well as Bayesian analysis. Researchers have also used parallelism to speedup one-step phylogenetic analysis [JAN01] and multiple alignment[WHE94, LI03].

2.2 Applied Phylogenetics

Originally, phylogenetics was considered relevant only to taxonomic and evolutionary studies. However, the ability to identify conserved and divergent regions of genomes is becoming critical data for numerous disciplines in biology and medicine. These fields include vascular genomics [RUB03], ecology [SIL97], physiology [CAR94], pharmacology [SEA03], epidemiology [ROS02], developmental biology [WHI03], and forensics [BUD03]. Phylogenetics has even been used in successful criminal prosecution of a doctor who attempted to cause HIV infection in his former girlfriend via blood products taken from a HIV patient under his care [MET02].

Here we focus on cases in which phylogenetic analyses have helped researchers to understand the evolution and spread of infectious diseases. We provide exemplar cases in which phylogenetic analyses of viral genomes have been crucial to understand complex patterns of transmission among animal and human hosts: Severe Acute Respiratory Syndrome (SARS) [KSI03] and influenza [WEB92].

2.2.1 Phylogenetic Analysis in the Context of Emerging Infectious Disease Research

Emergent infectious diseases often evolve via zoonosis; shifts of an animal pathogen to human host. In fact, most category A pathogens and potential agents of bioterrorism and more than 75% of emergent diseases have zoonotic origins [TAY01, FRA02].

A typical set of tests for the hypothesis of animal host of a disease might be (1) experimentally exposing the candidate host animals with isolated viruses and ascertaining whether infection and viral shedding occurs [MAR04]; or (2) survey populations of animals with antibodies for exposure to the virus [GUA03]. These activities often provide model organisms for vaccine and drug development, data on seroprevalance, and sequence data for viruses isolated from various candidate hosts.

Phylogenetic analysis of genomes is a complement to laboratory and survey studies with a distinct advantage. With phylogenetics, the researcher is not restricted to testing a single hypothesis for a specific candidate host in each experiment. Provided with sequence data for a diverse set of candidate hosts,

a researcher performing a single phylogenetic analysis makes a vast number of comparisons, thus evaluating simultaneously many alternative hypotheses. These hypotheses include the evaluation of pathways of transmission among several hosts and the polarity of the transmission events. For example, experimentalists report that small carnivores in Chinese markets have been exposed to SARS-CoV [GUA03] and the virus can infect domestic cats [MAR04]. On their own these data do not necessarily reconstruct the history of the zoonotic and genomic events that underlie the SARS epidemic.

Furthermore, whether or not interspecies transmission is observed or enhanced under controlled laboratory conditions, phylogenetic research is distinct as it can address whether the genomic record has evidence to support a hypothesis for a particular transmission pathway. For example, if phylogenetic analysis reveals multiple independent events of human to avian transmission of influenza viruses without intermediate hosts such as swine that provides a strong argument to reevaluate the hypothesis that pigs serve as "mixing vessels" for avian and human viruses leading to influenza epidemics [SCH90].

In many cases, host shifts occur via recombination between two ancestral pathogen genomes to produce a chimeric descendent. Epidemics can occur when, subsequent to recombination, a lineage of pathogens establishes itself in a new population of hosts, vectors, or reservoir species that can amplify and distribute the pathogen [MOR95]. A host shift can require key mutations and rearrangement of the pathogen genome to infect cells of new hosts followed by adaptation to novel regulatory machinery. Phylogenetics can reconstruct genomic changes at the level of each nucleotide and unravel parental and descendent strains in recombination mediated host shifts [WAN05].

2.3 Evolution of Influenza

Influenza is a widespread respiratory disease caused by an RNA virus (Orthomyxoviridae). The influenza virus has been traditionally divided in three major types: A, B, and C. Influenza viruses of type A are known from many strains that infect both mammal and avian hosts, whereas the other two type are primarily known from humans. Influenza A is characterized by antigenic subtypes (see Sect. 2.3).

Influenza is interesting from both epidemiological and evolutionary points of view due to the interplay between genetic changes in the viral population and the immune system of hosts [EAR02]. There are two basic hypotheses on how influenza A viruses escape the immune response in host population to cause epidemics: (1) antigenic drift, meaning that random point mutations produces novel influenza strains that succeed and persist if they can infect and spread among hosts; (2) antigenic shift, meaning that genes derived from two or more influenza strains reassort thus creating a novel descendent genome with a constellation of genes that can infect and spread among hosts. In both

scenarios zoonosis is often involved. In case of antigenic shift the ancestry of only a fraction of the influenza genes may be zoonotic.

Two major classes of influenza epidemics are recognized in humans: seasonal outbreaks and large-scale epidemics known as pandemics [WEB92]. Seasonal influenza is a significant public health concern causing 36,000 deaths and 200,000 hospitalizations in the United States in an average year [GER05]. Elderly and children account for many of these severe cases of seasonal influenza. Much of the population has partial immunity to seasonal influenza strains that are typically descendents of strains circulating in previous years. Pandemics are often caused by infection, replication, and transmission among the human population with influenza strains of zoonotic origin to which few people have prior immunity. Pandemics are rare but can affect the entire human population, irrespective of an individual's predisposition to respiratory diseases. In fact, the 1918 pandemic disproportionately affected young adults [TAU06], suggesting that older adults may have had some immunity.

There have been three major influenza A pandemics, 1918 (H1N1), 1957 (H2N2), and 1968 (H3N2). The pandemic of 1918 is estimated to have killed tens to a hundred million people worldwide and 675,000 in the United States [TAU01]. The Asian flu pandemic of 1957 and the Hong Kong flu pandemic of 1968 were less severe, but caused tens of thousands of deaths in the United States [HHS04]

All of these pandemic strains are thought to have originated in wild birds [WEB92]. The 1957 and 1968 strains are believed to be the results of antigenic shift. However, recent studies suggest that the H1N1 influenza virus that caused the pandemic of 1918 was entirely of avian origin rather than a human-avian reassortant [TAU05]. Other researchers have countered that the 1918 H1N1 strains had a more commonly accepted route to infection of human populations by reassortment in mammals [GIB06; ANT06].

Pandemics can theoretically occur with any strain of influenza. Most influenza infections since 1968 have been attributed to influenza A H3N2 or H1N1 strains. However, there have been several recent reports of novel human infections from avian strains of influenza with subtypes thought to occur rarely in humans. Several cases of human infection of viruses of subtype H7 of avian origin have recently occurred in Canada [TWE04] and the Netherlands [KOO04]. Avian influenza of antigenic subtype H5 and H7 viruses can be found as low or high pathogenic forms depending on the severity of the illness they cause in poultry. Thus far, influenza H9 virus has only been identified as strains with low pathogenicity [LIN00].

Alarmingly, highly pathogenic strains of influenza A with an H5N1 subtype have spread rapidly among various species of birds in China, Southeast Asia, Russia, India, the Middle East, Africa, Eastern, and Western Europe [WHO07a]. These H5N1 influenza A strains share common ancestry with the outbreak of H5N1 that lead to a massive chicken cull and six human deaths in Hong Kong in 1997 [LI04]. Between 2003 and September 10, 2007, there have been 328 cases and 200 deaths among humans [WHO07b]. There are several

instances of H5N1 infection of felids and swine in Asia. There is scant evidence of human-to-human transmission in Thailand [UNG05] and Indonesia [YAN 2007]. If lethality to human cases of H5N1 drops, the virus might spread rapidly and without being detected.

Many predict an upcoming avian influenza pandemic of devastating human and economic costs. In the United States alone, it is projected that 15–35% of the population will be affected and the costs could range from 71.6 to 166.5 billion United States (US) dollars [GER05]. Although vaccine production can in theory be modified to include H5N1 strains [DUT05], the genomes of interest are moving targets. It remains unknown whether the descendents of the contemporary H5N1 virus will achieve efficient human-to-human transmission and if this will occur via incremental mutations or a more punctuated reassortment mediated change. Thus phylogenetics is a key technology to track the evolution of H5N1 and compare those changes to genomic and zoonotic events that underlie pandemics.

Subtypes of Influenza Type A

The viruses of influenza type A are classified as various subtypes that represent differences in the antigenic reaction of two key glycoproteins: hemagglutinin (HA) and neuraminidase (NA). These proteins reside on the surface of the virion. These proteins play key roles in recognition and infection of susceptible hosts (HA) and viral replication (NA). These surface proteins are primary antigens recognized by the host immune system [WEB92].

The subtypes of influenza A are labeled according to the reaction of standard monoclonal antibodies to these HA and NA proteins provided by the US Centers for Disease Control to laboratories participating in the World Health Organization's (WHO) surveillance program [HHSb].

Although this number will soon expand, there are currently 16 different antigenic subtypes recognized for HA (labeled from H1 to H16) and 9 different antigenic subtypes of NA (from N1 to N9). Thus, a subtype of influenza virus type A is labeled with the number associated with HA and NA proteins (e.g., the most common subtype found in humans H3N2).

Since 1948, influenza viruses have been the focus of a coordinated surveillance program organized by the WHO [WHO05]. The hemagglutinin gene (HA) is the major target of the influenza surveillance. This program helps track predominant strains to inform the development of new vaccines. Influenza viruses are sampled worldwide through the National Influenza Centers located in 54 countries [WHO05]. Many of the viral isolates sampled by these programs are sequenced for the hemagglutinin gene, although there has been an increasing interest in sampling complete influenza genomes [GHE05, OBE06].

An extensive record of hemagglutinin sequences of the influenza viruses type A isolated since 1902 are publicly available. These data provide a unique set of challenges and opportunities for phylogenetics. The geographically wide

and temporally long sampling of viral isolates provides an unprecedented opportunity to study evolutionary patterns underlying the spread and host range of an infectious disease. However, as described earlier, large datasets present an enormous search space of possible evolutionary scenarios to be evaluated.

2.3.1 Phylogenetic Approaches to Influenza Type A

The availability of nucleotide sequences of influenza viruses has triggered numerous research groups to attempt reconstruction of the phylogenetic history of these viruses (e.g., [BUS99, YUA02, FER03, BUS04]). These groups draw on data from currently circulating strains as well as from historically important strains gathered from archival tissue samples. Examples of archival tissues that have provided date of interest to the 1918 epidemic include lung biopsies of deceased soldiers, victims frozen in Alaskan permafrost[TAU97], and waterfowl collected for the Smithsonian in 1916–1917 [FAN02].

Phylogenetic analysis of seasonal influenza sequence data has been used to classify nucleotide substitution mutations. In many codons of the HA gene mutations that produce a change in protein sequence are more frequent than those that do not [BU99]. This finding indicates that selective pressures imposed by the immune system of the hosts can drive the evolution of some codons of HA. Thus an evolutionary perspective can illuminate functional studies of infectious disease [EAR02].

2.3.2 Large Scale Phylogenetic Analysis of HA

As noted, phylogenetics have been widely used to understand history of influenza epidemics, host shifts, as well as evolutionary interactions with the hosts immune system (see Sect. 2.3.1). However, most phylogenetic analyses of influenza thus far have used only fractions of the dataset of influenza nucleotide sequences in the public domain. The sequences in the public domain are largely HA, but recently whole genomes have been produced. The Institute for Genomic Research (TIGR) is rapidly sequencing and releasing into the public domain thousands of influenza genomes under the Microbial Sequencing Center (MSC) program sponsored by the National Institute of Allergy and Infectious Disease (NIAID) [GHE05]. St. Jude Children's Research Hospital in Memphis has contributed a significant increase in the number of avian influenza genomes sequences [OBE06].

Most existing phylogenetic analyses of influenza have focused on the phylogenetic relationships of particular subgroups of influenza type A, such as the H5N1 subtype (e.g., [LI04]) or the H3N2 subtype (e.g., [BUS99]). These analysis have provided useful information but have depicted a disjoint picture of the evolution of the major lineages of influenza. In contrast, other studies have attempted broader subtype scope; however, they included a single viral isolate as an exemplar of each subtype [SUZ02]. This study failed to include an extensive sampling of strains. Poor strain sampling can have a negative

impact on the performance of phylogenetic methods (see Sect. 2.1.5) and does not test whether the subtypes are natural groups (i.e. monophyletic). A very recent study has used whole genomes of 136 isolates drawn from a variety of avian influenza subtypes [OBE06].

Here we show results of a comprehensive phylogenetic analysis based on hemagglutinin DNA sequences of 2,359 viral isolates. These sequences include representatives of the 16 different subtypes of the hemagglutinin protein of influenza type A, recorded worldwide by the World Health Organization surveillance program. The analyzed viruses were isolated as early as 1902, from tissues of patients who died during the 1918 Spanish flu epidemic, to recently sequenced isolates from the 2004 seasonal flu and H5N1 outbreak. The analyzed DNA sequences also implies a broad range of host organisms, including multiple species of wild and domestic birds, humans, swine, horses, felids, and whales.

An inclusive phylogenetic analysis with a large number of taxa require the use of efficient tree search strategies (see Sect. 2.1.5) and the use of multiple computers dedicated to the phylogenetic analysis. The cost of computation is tied primarily to the number of strains, not nucleotides. Thus the inclusion of whole genomes does not contribute significantly to the compute bound nature of phylogenetic analysis. However, the inclusion of whole genome data does increase memory demands.

This 2,359 isolate dataset was analyzed with a parallelization of the tree search strategy implemented in a recently developed software for parsimony analysis [GOL03]. The results of this analysis are used here to illustrate two new uses, longitudinal analyses of patterns of zoonotic transmission and assessment of surveillance quality.

Relationships of HA Subtypes

Our results on the relationships of HA subtypes shown in Fig. (2.6) has similarities with the results of Suzuki and Nei [SUZ02], including the clades ((H8 H12) H9), ((H15 H7) H10), ((H4 H14) H3), and (((H2 H5) H1) H6). However, the position of H13 and H11 differ in our trees due to our inclusion of H16. Moreover, the relationship of these clades to one another differs in our assessments. Our tree has a staircase shape with ((H8 H12) H9) basal most, whereas Suzuki and Nei's [SUZ02] tree has a symmetrical shape with no clear basal group.

Host Shifts in HA

Influenza A viruses from wild aquatic birds have been identified as the source of influenza viruses isolated from birds of the order Galliformes (e.g., turkeys, grouse, quails, pheasants, domestic chickens, and their ancestral stock the jungle fowl) [WEB92]. Direct human infection by avian strains of influenza A is considered rare [LIP04]. After the discovery of receptors for both avian

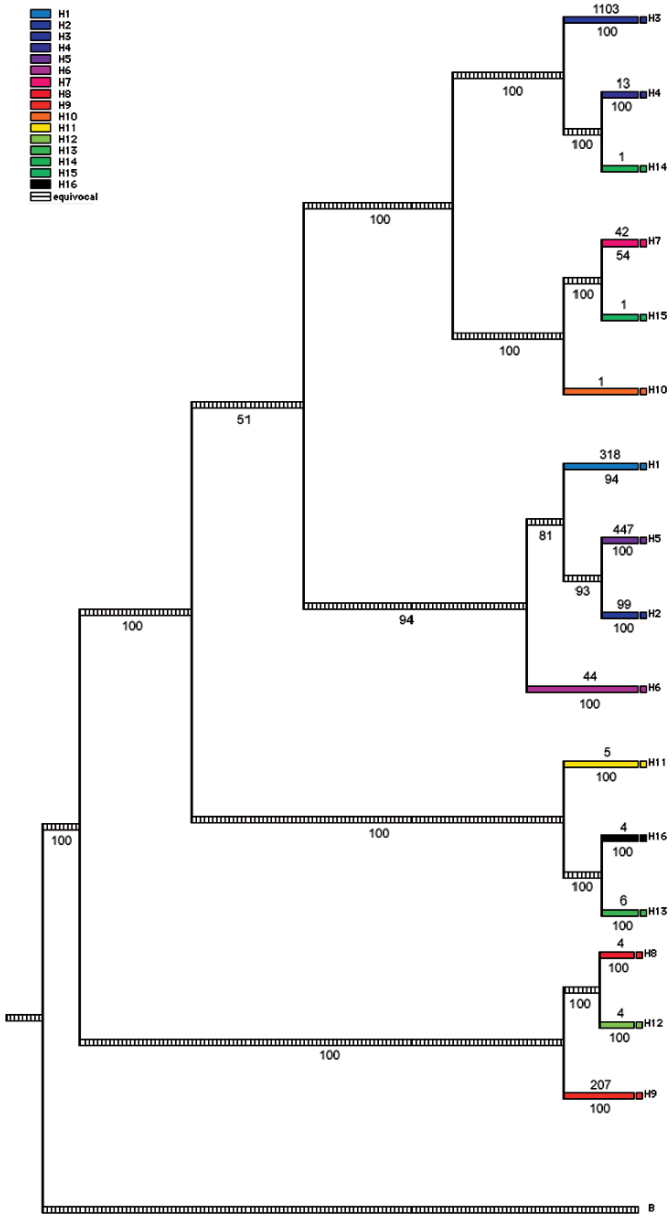

Fig. 2.6. Phylogeny of hemagglutinin (HA) sequences representing 2,358 isolates of influenza A, with a single sequence of influenza B as outgroup. To summarize the source tree we have condensed each subtype clade into a single branch. The numbers of isolates included in the full tree are presented as the numerals above each branch. The numerals below each branch represent jackknife support values (0 worst to 100 best). Sequence and character data was drawn from Genbank (www.ncbi.nlm.nih.gov) and the Influenza Sequence Database (www.flu.lanl.gov)

and mammalian strains of influenza in the trachea of pigs, it has been hypothesized that domestic swine act as intermediate hosts in which human and avian viruses can recombine [SCH90]. This mechanistic hypothesis of viral transmission is widespread. However, as discussed above, a number of events of suspected direct transmission of avian influenza viruses to humans have been reported [LIP04, UNG05].

Hypotheses on the relative frequency of host shifts can be made on a phylogenetic tree through the optimization [FIT71] of a character with states representing various hosts of the viral isolates under consideration (see Sect. 2.2.1). We performed this analysis on our tree of 2,359 HA sequences and found that most of the internal nodes close to the root are optimized as having an avian origin (Fig. 2.7). Thus the results of this analysis are consistent with the hypothesis of an avian origin of all influenza type A viruses [WEB92]. These results also show that most major lineages of influenza A that infect domestic birds originated in aquatic birds. This is compatible with the hypothesis that wild aquatic birds as the natural reservoir of influenza viruses of type A [WEB92].

However, the pattern of host shifts resulting from our study of 2,359 HA sequences seems to be much more complex than previously thought [GAM90, LIP04]. For instance, in many cases, after the spread of influenza type A viruses into domestic bird and mammal populations (including humans), some derived lineages are later spread again to aquatic birds. Furthermore, the results indicate that direct shifts from avian to human hosts have occurred 18–27 times independently in different lineages (without observed intermediate hosts). It must be noted that the possibility of an intermediate host in avian-to-human transmission events cannot be completely rejected. It is possible that an intermediate host existed in nature but it was not sampled by the surveillance program and therefore not included in the analysis. However, based on the available evidence, it seems that host shifts from birds to humans have been frequent in the evolutionary history of influenza type A. Moreover, avian-to-human shifts are more common than swine to human shifts in the history of influenza.

Multiple direct avian-to-human shifts appear to occur in the case of the putative pandemic strains of influenza A (subtype H5N1) that have spread across Eurasia since 1997 [WHO05]. In addition to being highly pathogenic, these H5N1 strains have independently infected other hosts such as felids and pigs in several instances.

Predictive Power of Phylogenetics Analysis

Phylogenetics is practiced by most as a historical science; however, several researchers noted that aspects of the tree shape may be used in predicting future genetic lineages of influenza against which it is important to design vaccines [GRE04]. Notable among these assertions are the studies in the shape of influenza A phylogeny as viewed through the hemagglutinin (HA) gene

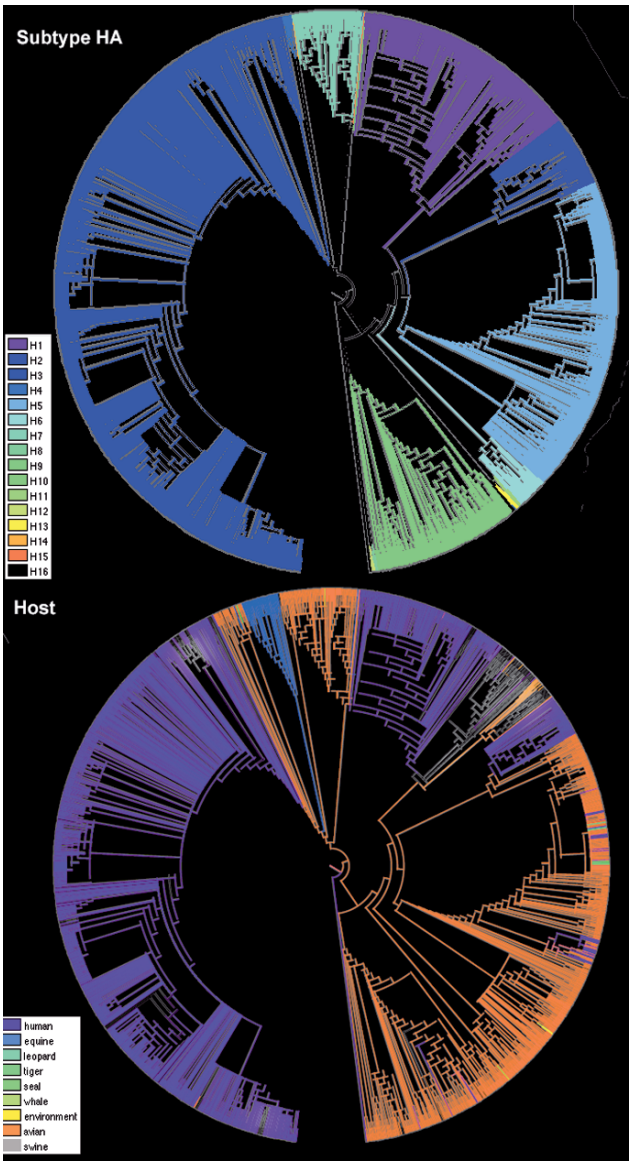

Fig. 2.7. Two character optimizations on the for hemagglutinin (HA) sequences representing 2,358 isolates of influenza A, with an influenza B outgroup at the root. The top tree has an optimization of the character "HA antigenic subtype". The lower tree depicts optimization of the character "host". Character data was drawn from Genbank (www.ncbi.nlm.nih.gov) and the Influenza Sequence Database (www.flu.lanl.gov). Optimizations and tree graphics were made with Mesquite (www.mesquiteproject.org). For better visualization contact the authors for files in scalable pdf format

[BUS99, FER02]. The HA gene codes for a surface glycoprotein of the virion responsible for binding to sialic acid on host cell surface receptors. At a genomic level, lineages of influenza are constantly changing due to mutation that occurs at high rates in RNA viruses. Extinction of evolutionary lineages of viruses to which hosts have become immune or when susceptible hosts are in short supply is common [GRE04]. This process of constant replacement of influenza lineages produces a characteristic coniferous shape to a phylogeny reconstructed from HA sequences[BUS99]. The "conifer" metaphor refers to the hypothesis that influenza HA is constantly changing but there is limited diversity at any time [FER02]. Thus an influenza HA tree appears to be formed by addition of strains to the apex of the tree's trunk that contains the contemporary "infectious" viruses rather than more basal presumably "extinct" lineages to which hosts are immune. Other groups of researchers have used the assumption that there is limited influenza A diversity at any one time to downplay the utility of phylogenetic approaches [PLO02]. As an alternative to phylogenetics, which they consider difficult, these groups make predictions based on size of various clusters of related isolates, termed "swarms" [PLO02]. Several groups, whether using trees or swarms, have identified putatively dominant strains of influenza to predict the genetic makeup of future viral populations [PLO02] [BUS99].

If these assumptions were never violated, the diversity of a previous year's flu season could be assessed, forthcoming strains predicted, and thus used to inform vaccine design. In practice, the CDC uses a mixture of viral strains comprised of H1N1 and H3N2 of influenza A and an influenza B virus. For example, the 2005–2006 vaccine was based on A/New Caledonia/20/99 (H1N1), A/California/7/2004 (H3N2), and B/Shanghai/361/2002 viruses [PAL06]. Notably the H5N1 strain (or any of the other avian strains with potential to infect humans) is currently not considered in the vaccine that is seasonally administered to civilians in the United States.

The ability to predict influenza viral strains that will affect human and animal populations is important. However, prediction methods and experimental designs that are relevant to those methods are in their infancy. Current surveillance programs are focused on detection of antigenically novel strains. As such, surveillance programs are not designed as ecological experiments to quantitatively measure strain-specific incidence and cluster size. Furthermore, the current sample of influenza diversity may be biased by partial genomic sequencing, differences in effort within various geographic and political boundaries, focus on certain subtypes of interest, and differential efforts over time due to variable public concern. Recent papers using whole genome data have indicated that the conifer like growth assumption of HA-based phylogenies that has been central to predictive models of H3N2 seasonal influenza [BUS99, FER02] may be violated. Full genome analysis of H3N2 has shown that there are multiple co-circulating lineages; some of which may be overlooked by vaccine designs [HOL05, GHE05]. Similarly, our large scale-analysis of 2,359 HA sequences depicts that many subtypes and lineages within

subtypes of influenza are circulating and being exchanged among human and animal populations at any one time Fig. 2.7.

Viral Surveillance Quality: A Phylogenetic Perspective

In addition to providing hypotheses on the relationships of a group of organisms, phylogenetic trees imply a temporal order of the successive internal nodes (i.e., the time at which a single evolutionary lineage splits producing two independent descendent lineages). Minimal estimates on the date at which these evolutionary splits occur can be obtained through the analysis of the time at which the descendant organisms (leaves) are known to occur. These estimates can be computed with the implementation of an irreversible Sankoff character in which the cost of transformation between two character states represents the amount of time elapsed between the time of appearance of two terminal taxa [PN01].

Influenza A viral sequences are named with the host, locale, and year in which each isolate was sampled by the surveillance program. Several methods exist to measure the correlation between the temporal dates of sampled organisms and the relative order they show in the phylogenetic tree. Here we adapt the Manhattan Stratigraphic Metric (MSM*) to influenza surveillance. The MSM* was originally developed to assess the quality of the fossil record [PN01] (Table 2.2). However, the MSM* is simply a quantitative measure of how well the available data reflects the diversification pattern of the taxa present in the optimal phylogenetic trees and is thus of general utility.

An extensive sampling of sequences, such as the one gathered for the study of 2,359 isolates, is critical to comparatively assess quality of surveillance in various regions, among various strains, and over periods of time. Our results show that this correlation between branching pattern and dates of viral isolation is good in that it significantly differs from a random expectation. This is true over the entire tree as well as when some individual lineages are measured. However, the relative quality of surveillance differs markedly between lineages.

Table 2.2. Results for Manhattan Stratigraphic Metric (MSM*) applied to various subtype clades isolated in the past 30 years

Strain	MSM*
H5N2	0.12
H9N2	0.36
H1N1	0.41
H3N2	0.49
H5N1	0.53

A score close to 1 in the MSM* reflects good surveillance and values close to 0 imply poor surveillance.

One example of differential surveillance quality occurs in two closely related groups of avian influenza of H5 hemagglutinin subtype. One group in this example contains the highly pathogenic H5N1 strains that currently circulate in Eurasia, the Middle East, and Africa. This large clade has been the focus of intense surveillance since the discovery of widespread infection among wild and domestic birds and some avian-to-human transmission[YUA02, UNG05]. The H5N1 viral isolates form a sister clade to H5N2 known from domestic and wild bird in the Americas (H5N2). The number of available hemagglutinin sequences of H5N2 comprise less than one fifth of the number of HA sequences for H5N1. This in itself represents a measure of the surveillance intensity devoted to these two groups of avian influenza. However, even if the number of sequences is normalized at 100 sequences to perform the MSM test, the surveillance quality of the H5N1 clade is far superior to the H5N2 clade.

We can also use visualization techniques to assess surveillance quality. Typically branches of a phylogenetic tree are scaled used to depict the number of mutations or other character changes assigned to each branch. However, we have adapted this use of branch scaling to reflect the number of years that have passed between sampling of related isolates rather than mutations or characters. Compare Fig. 2.8 which has short branch lengths reflecting good surveillance quality with Fig. 2.9 which has long branch lengths implying poor surveillance quality.

Cases in which there is poor correlation between the date of sampling of a given isolate and its inferred date of origin would indicate that the surveillance program is failing to closely monitor the persistence of diverse lineages of influenza (2.9).

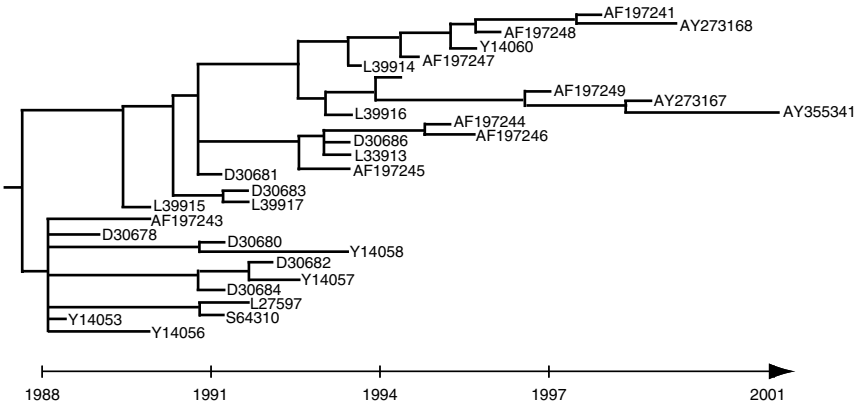

Fig. 2.8. Cases of efficient surveillance in which the sampling of sequences through time reflects the diversification patterns of the phylogenetic tree based on the hemagglutinin sequences. For a full tree contact the authors for files in scalable formats such as pdf and nexus

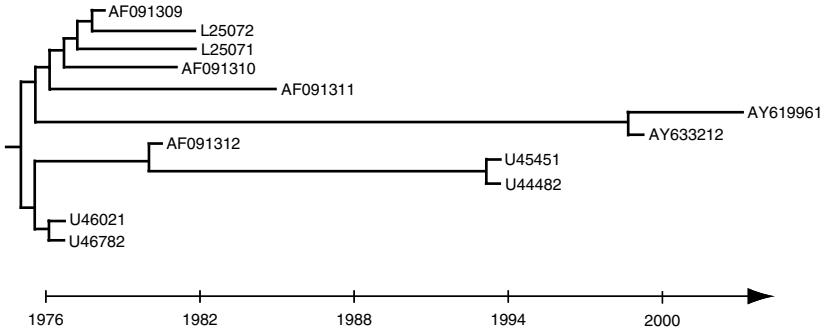

Fig. 2.9. Cases of poor surveillance in which viral lineages have gone undetected for many years since its evolutionary origins. In this case, sequences AY619961 and AY633212 were detected around the year 2000 but their evolutionary origin dates back to the late 1970s. The available sampling of isolates through time does not reflect the diversification patterns of the phylogenetic tree. For a full tree contact the authors for files in scalable formats such as pdf and nexus

2.4 Evolution of SARS Coronaviruses

The Zoontic Origins of Coronaviruses Associated with SARS

Severe Acute Respiratory Syndrome (SARS) is a novel human illness caused by a previously unrecognized coronavirus (CoV) termed SARS-CoV [MAR03] [ROT03]. SARS-CoV may be of zoonotic origin. However as of today, there remain conflicting reports on the zoonotic origins of SARS CoV. Guan et al., 2003 [GUA03] report that SARS CoV originated in small carnivores whereas Li et al., 2005 [LI05] and Lau et al., (2005) [LAU05] counter that SARS CoV originated in bats.

No matter the type of phylogenetic perspective they may espouse, most virologists produce the same basic data by surveying putative host animals and patients with antibodies, then isolating and sequencing partial or whole genomes of various viruses detected in hosts. Molecular phylogenetic analyses of the nucleotide or inferred amino acid sequence data from various viral isolates can then be used to reconstruct the history of the transmission events the virus among hosts. The fundamental belief associated with this research program is that the branching pattern of the phylogeny will reveal a temporal series of transformations when character of interest such as the host is optimized on the viral phylogeny.

Most virology researchers rely on distance methods. The most popular distance method among virologists is neighbor-joining (NJ) [SAI87]. Distance methods require a precomputed multiple alignment of DNA or amino acid sequences drawn from homologous genes of the viral strains of interest. Then in NJ, the most grossly similar pair of isolates (as represented by sequences) are clustered. The clustered pair is then considered as a single taxon and the

next most similar pair of taxa is to cluster until only two taxa remain and are joined. In distance methods no outgroups are proposed and no assumptions of ancestral character states are considered. As a result, polarity of transformations can only be inferred as from dissimilar to similar. Distance methods output a single unrooted, star-shaped graph.

Nevertheless preparing figures, some investigators who use distance methods choose to impart directionality by selecting an edge of the graph to serve as a root of the tree. The choice of root is crucial in depicting the polarity of host shifts and depicting clades. The rooting step has been executed variably by researchers comparing sequence data from CoVs isolated from humans with sequence data CoVs isolated from small carnivores. In the case of Guan et al. ([GUA03] see their Figs. 2 and S2 of their supplemental materials) and the Chinese SARS Molecular Epidemiology Consortium ([CSARS04] their supplemental Fig. S7) these researchers simply force the root position on their drawings such that they represent SARS-CoV isolates from small carnivore hosts as ancestral. In other drawings, no outgroup is designated ([CSARS04] their Fig. 2) or a human SARS-CoV outgroup is used and the animal SARS-CoV isolates are omitted from the tree ([CSARS04] their Fig. S6 of the supplemental materials). In the case of Song et al., trees are presented as unrooted ([SON05] their Fig. 1a), rooted on a clade comprised of two SARS-CoV isolates, one from human and the other from a carnivor ([SON05] their Fig. 1b), and in a regression analysis a date for a common ancestor of SARS-CoV isolated from humans is calculated using a human basal group ([SON05] their Fig. 3).

In papers comparing sequence data from SARS-Like CoV recently isolated from bats to that from humans and small carnivores Lau et al.,[LAU05] do not root their trees (their Fig. 2) and Li et al., [LI05] and force the root position on their drawing such that one of the bat sequences is ancestral (their Fig. S4 of the supplemental material). Thus, although all these studies employ distance methods, the researchers use various, often facultative means to infer the animal origins of SARS-CoV.

2.4.1 Importance of Outgroups

Other methods of phylogenetic analysis focus on characters, states, edit costs for changes among states, outgroup assumptions, and polarity of change among states – rather than gross similarity in the case of distance methods. Characters can be polymorphisms recognized in columns of aligned nucleotides or amino acids from sequences of interest or phenotypic states such as host, date of isolation, or antigenic subtype. Another feature of most character based methods that differs from NJ is that character based methods examine many randomly generated trees (each representing an evolutionary hypothesis of character transformations and organismal relationships). Thus the concepts of optimality and hypothesis testing are tightly associated with cladistic and maximum likelihood inference. Optimal trees represent more

defensible hypotheses. Moreover, character based methods of molecular phylogenetics rely on explicit choices of outgroup to make assumptions about ancestral character states and thus polarize transformations of phenotypes and genotypes that can be reconstructed from data. In order to make an explicit assumption of ancestral character states the investigator designates at least one taxon as the outgroup. The outgroup method originated in cladistics [WAT81] and has become central to the phylogenetic inference [NIX94]. If chosen carefully, the outgroup estimates baseline character states in sequence and phenotypes such that transformations (such as host shifts) can be reliably inferred. To illustrate the choice of outgroup taxon or taxa and clarify the relationships of the organisms, character based trees are often rooted from the outgroup and ingroup.

Just as in distance methods, in most character-based methods, sequence data is aligned before the phylogenetic analysis. Novel implementations, termed direct optimization, allow unaligned sequence data to be analyzed without precomputing an alignment, Wheeler [WHE96]. In direct optimization, sequence data are aligned as various trees are built and their optimality is assessed (using maximum likelihood and cladistic optimality criteria as specified by the investigator). Thus for each tree a specific alignment that is optimal for that tree is constructed. One additional advantage of direct optimization is that the outgroup need not be designated by the investigator but rather randomized during the search for optimal trees and alignments. In some implementations of character based methods where prealignment is necessary, the outgroup can be randomized by scripting a series of analyses. Outgroup randomization enables analyses of taxa where previous knowledge of ingroup/outgroup relationships is lacking or is among the hypotheses the investigator wants to test via tree search.

2.5 Discussion

Large-scale phylogenetic analyses are particularly useful to study global problems of infectious disease. However, phylogenetic analysis of large number of organisms and whole genomes is an extremely challenging computational problem. Recent advances in heuristic tree search algorithms, alignment methods, and parallel computing strategies have been successful. These advances have pushed upward the limits of taxon sampling considered tractable.

Large data sets analysis is interesting not only because it presents interesting computational challenges, moreover large dataset analysis is leading to new knowledge about natural phenomena. For example, in the recent past, researchers working on small datasets argued that influenza had limited diversity at any one time and that this should allow us to predict which strains are important for vaccine design. On the contrary, with large datasets, we find that there are multiple co-circulating lineages at any one time. Thus, large datasets and means to analyze them are important for future vaccine design.

Character-based approaches to phylogenetics provide a wide variety of tools that can be used to better understand the evolutionary processes underlying the spread of infectious diseases. We have discussed here only some of the wide array of applications that phylogenetics and outgroup criteria can have on genomic studies of infectious disease. We look forward to the continued synergistic development of technological and scientific means to better understand infectious and zoonotic disease.

Acknowledgements

Mr. Farhat Habib M.S. was instrumental to build and maintain cluster computers. Rebecca Allen and Jiarui Lian helped to organize phenotype and genetic data. The authors have no competing interests, financial or otherwise. DP acknowledges the National Science Foundations (NSF) support of the Mathematical Biosciences Institute (MBI) and the MBI. DP and DJ acknowledge the support of the Department of Biomedical Informatics and the Ohio State University Medical Center. DJ acknowledges that this material is based upon work supported by, or in part by, the US Army Research Laboratory and the US Army Research Office under contract/grant number W911NF-05-1-0271 and NSF 0531763.

References

[ANT06]. J. Antonovics, M.E. Hood, and C.H. Baker, Molecular virology: was the 1918 flu avian in origin? Arising from: J.K. Taubenberger et al. Nature **437**, 889–893, Nature **440** (2006) E9.

[ARI343]. Aristotle, Historia Animalium, 343 BC.

[BRA02]. M.J. Brauer, M.T. Holder, L.A. Dries, D.J. Zwickl, P.O. Lewis, and D.M. Hillis, Genetic Algorithms and Parallel Processing in Maximum-Likelihood Phylogeny Inference, Mol. Biol. Evol. **19** (2002) 1717–1726.

[BU99]. R.M. Bush, W.M. Fitch, C.A. Bender, and N.J. Cox, Positive Selection on the H3 Hemagglutinin Gene of Human Influenza Virus A., Mol. Biol. Evol **16** (1999) 1457–1465.

[BUD03]. B. Budowle, M.W. Allard, M.R. Wilson, and R. Chakraborty, Forensics and mitochondrial DNA: Applications, Debates, and Foundations, Annu. Rev. Genomics Hum. Genet. **4** (2003) 119–141.

[BUS04]. R.M. Bush, Influenza as a model system for studying the crossspecies transfer and evolution of the SARS coronavirus, Phil. Trans. R. Soc. Lond. B. **359** (2004) 1067–1073.

[BUS99]. R.M. Bush, C.A. Bender, K. Subbarao, N.J. Cox, and W.M. Fitch, Predicting the Evolution of Human Influenza A., Science **286** (1999) 1921–1925.

[CAR94]. J.P. Carulli, D.M. Chen, W.S. Stark, and D.L. Hartl, D. Phylogeny and physiology of Drosophila opsins., J. Mol. Evol. **38** (1994) 25–62.

[CER98]. C. Ceron, J. Dopazo, E. Zapata, J. Carazo, and O. Trelles, Parallel implementation of DNAml program on message-passing architectures, Parallel Computing **24** (1998) 701–716.

[CHA00]. B. Chang, and M. Donoghue, Recreating ancestral proteins, Trends Ecol. Evol. **15** (2004) 109–114.

[CHA01]. M.A Charleston, Hitch-Hiking: A Parallel Heuristic Search Strategy, Applied to the Phylogeny Problem, J. Comput. Biol. **8** (2001) 79–91.

[CSARS04]. The Chinese SARS Molecular Epidemiology Consortium: Molecular Evolution of the SARS Coronavirus During the Course of the SARS Epidemic in China, Science **303** (2004) 1666–1669.

[DUT05]. G. Dutton, Preparing for a Potential Pandemic. Therapeutic and Vaccine Manufacturers Working to Combat the Avian Flu, Genetic Engineering News **25** (2005) 15:1.

[EAR02]. D. Earn, J. Dushoff, and S. Levin, Ecology and evolution of the flu, T.R.E.E **17** (2002) 334–340.

[FAN02]. T. Fanning, R. Slemons, A. Reid, T. Janczewski, J. Dean, and J. Taubenberger, 1917 Avian influenza Virus Sequences Suggest that the 1918 Pandemic Virus Did Not Acquire Its Hemagglutinin Directly from Birds, Journal of Virology **76** (2002) 7860–7862.

[FAR70]. J.S. Farris, Methods for Computing Wagner trees, Syst. Zool. **19** (1970) 83–92.

[FAR83]. J.S. Farris, The logical basis of phylogenetic analysis. In: Platnick, N.I., Funk, V.A. (eds), Advances in Cladistics, Columbia University Press, New York, 1983.

[FAR96]. J.S. Farris, V.A. Albert, M. Kallersjo, D. Lipscomb, and A.G. Kluge, Parsimony Jackknifing Outperforms Neighbor-Joining, Cladistics **12** (1996) 99–124.

[FEL73]. J. Felsenstein, Maximum Likelihood and Minimum-Step Methods for Estimating Trees from Data on Discrete Characters, Syst. Zool. **22** (1973) 240–249.

[FEL78]. J. Felsenstein, The Number of Evolutionary Trees, Systematic Zoology **27** (1978) 27–33.

[FEL81]. J. Felsenstein, Evolutionary trees from dna sequences: A maximum likelihood approach, J. Mol. Evol. **17** (1981) 368–376.

[FER02]. N.M. Ferguson and R. Anderson, Predicting evolutionary change in the influenza A virus, Nat Med. **8** (2002) 562–3.

[FER03]. N.M. Ferguson, A.P. Galvani, and R.M. Bush, Ecological and immunological determinants of influenza evolution, Nature **422** (2003) 428–433.

[FG82]. L.R. Foulds and R.L. Graham, The Steiner problem in phylogeny is NP-complete, Adv. Appl. Math. **3** (1982) 43–49.

[FIT71]. W.M. Fitch, Towards defining the course of evolution: Minimum change for a specific tree topology, Syst. Zool. **20** (1971) 406–416.

[FIT97]. W.M. Fitch, R.M. Bush, C.A. Bender, and N.J. Cox, Long term trends in the evolution of H(3) HA1 human influenza type A, Proc. Natl. Acad. Sci. USA **94** (1997) 7712–7718.

[FIT83]. W.M. Fitch and T. Smith, Optimal sequence alignments, Proc. Natl. Acad. Sci. USA **80** (1983) 1382–1386.

[FLE05]. R. Fleissner, D. Metzler, and A.V. Haeseler, Simultaneous Statistical Alignment and Phylogeny Reconstruction, Syst. Biol. **54** (2005) 548–561.

[FRA02]. D. Franz, The potential bioweaponization of zoonotic diseases. pp. 15–17 in The Emergence of Zoonotic Diseases: Understanding the Impact on Animal and Human Health. T. Burroughs, S. Knobler, and J. Lederberg, eds., Forum on Emerging Infections Board on Global Health (BGH), Institute of Medicine (IOM). National Academy of Sciences Press, Washington D.C., USA, 2002.

[GAM90]. M. Gammelin, A. Altmller, U. Reinhardt, J. Mandler, V.R. Harley, P.J. Hudson, W.M. Fitch, and C. Scholtissek, Phylogenetic analysis of nucleoproteins suggests that human influenza A viruses emerged from a 19th century avian ancestor, Mol. Biol. Evol. **7** (1990) 194–200.

[GER05]. J. Gerberding, Pandemic Planning and Preparedness, http://www.cdc.gov/Washington/testimony/in05262005.htm (2005)

[GHE05]. E. Ghedin, N. Sengamalay, M. Shumway, J. Zaborsky, T. Feldblyum and 14 others, Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution, Nature **437** (2005) 1162–1166.

[GIB06]. M. Gibbs and A. Gibbs, Was the 1918 pandemic caused by a bird flu? Arising from: J.K. Taubenberger et al. Nature **437**, 889–893, Nature (2005) 440:E8.

[GOL02]. P.A. Goloboff, W.C. Wheeler, and D. Pol, Parallel searches of large datasets, Cladistics **19** (2002) 151.

[GOL03]. P.A. Goloboff, S.J. Farris, and K.C. Nixon, TNT: Tree Analysis Using New Technologies, Software package distributed by the authors and available at: http://www.zmuc.dk/public/phylogeny/TNT (2003)

[GOL99]. P.A. Goloboff, Analyzing Large Datasets in Reasonable Times: Solutions for Composite Optima, Cladistics **15** (1999) 415–428.

[GRA03]. T. Grant and A. Kluge, Data exploration in phylogenetic inference: Scientific, heuristic, or neither, Cladistics **19** (2003) 379–418.

[GRE04]. B. Grenfell, O. Pybus, J. Gog, J. Wood, J. Daly, J. Mumford, and E. Holmes, Unifying the epidemiological and evolutionary dynamics of pathogens, Science **303** (2004) 327–332.

[GUA03]. Y. Guan, B. Zheng, Y. He, X.L. Liu, Z.X. Zhuang, and 13 others, Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China, Science **302** (2003) 276–278.

[HEN66]. W. Hennig, Phylogenetic Systematics, University of Illinois Press, Urbana, 1966.

[HEN82]. M.D. Hendy and D. Penny, Branch and bound algorithms to determine minimal evolutionary trees, Math. Biosc. **59** (1982) 277–290.

[HHS04]. National Vaccine Program Office, Pandemics and Pandemic Scares in the 20th Century, http://www.hhs.gov/nvpo/pandemics/flu3.htm (2004)

[HHSb]. HHS Pandemic Influenza Plan, Part 2 Public Health Guidance for State and Local Partners, http://www.hhs.gov/pandemicflu/plan/pdf/S02.pdf (Date)

[HIL03]. D.M. Hillis, D.D. Pollock, J.A. McGuire, and D.J. Zwickl, Is Sparse Taxon Sampling a Problem for Phylogenetic Inference?, Syst. Biol. **52** (2003) 124–126.

[HIL96]. D.M. Hillis, Inferring Complex Phylogenies, Nature **369** (1996) 130–131.

[HOL05]. E. Holmes, E. Ghedin, N. Miller, J. Taylor, Y. Bao, and 6 others, Whole-Genome Analysis of Human Influenza A Virus Reveals Multiple Persistent Lineages and Reassortment among Recent H3N2 Viruses, PLoS Biol. **3** (2005) 1–11.

[HUE02]. J. Huelsenbeck, B. Larget, R. Miller, and F. Ronquist, Potential applications and pitfalls of Bayesian inference of phylogeny, Syst. Biol. **51** (2002) 673–688.

[JAN01]. D. Janies and W.C. Wheeler, Efficiency of parallel direct optimization. Cladistics, **17** (2001) S71–S82.

[JAN02]. D. Janies and W.C. Wheeler, Theory and practice of parallel direct optimization. pp. 115–124 in R. Desalle, G. Giribet and W. Wheeler eds. Molecular Systematics and Evolution: Theory and Practice, Birkhuser Verlag, Basel Switzerland, 2002.

[JON03]. R. Johnston, Integrating Methodologists into Teams of Substantive Experts, Studies in Intelligence **47** (2003) No. 1.

[JON95]. J.A. Jones, K.A. Yelick, Parallelizing the Phylogeny Problem, Proc. 1995 ACM/IEEE Conf. Supercomp., **25** (1995)

[KOO04]. M. Koopmans, B. Wilbrink, M. Conyn, G. Natrop, H. van der Nat, H. Vennema, A. Meijer, J. van Steenbergen, R. Fouchier, A. Osterhaus, and A. Bosman, Transmission of H7N7 avian influenza A virus to human beings during a large outbreak in commercial poultry farms in the Netherlands, The Lancet **363** (2004) 587–593s.

[KSI03]. T. Ksiazek, D. Erdman, C. Goldsmith, S. Zaki, T. Peret and 22 others, A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome, The New England Journal of Medicine **348** (2003) 1953–1966.

[LAU05]. S. Lau, P. Woo, K. Li, Y. Huang, H. Tsoi, and 5 others, Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats, Proc. Natl. Acad. Sci. USA **102** (2005) 14040–14045.

[LAV01]. G. Laver and E. Garman, The Origin and Control of Pandemic Influenza, Science **293** (2001) 1776–1777.

[LEM02]. A.R. Lemmon and M.C. Milinkovitch, The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation, Proc. Natl. Acad. Sci. USA **99** (2002) 10516–10521.

[LEW98]. P.O. Lewis, A Genetic Algorithm for Maximum-Likelihood Phylogeny Inference Using Nucleotide Sequence Data, Mol. Biol. Evol. **15** (1998) 277–283.

[LIN00]. Y. Lin, M. Shaw, Y. Gregory, K. Cameron, W. Lim, and 8 others, Avian-to-human transmission of H9N2 subtype influenza A viruses: Relationship between H9N2 and H5N1 human isolates, Proc. Natl. Acad. Sci. USA **97** (2000) 9654–9658.

[LI00]. S. Li, D.K. Pearl, and H. Doss, Phylogenetic Tree Construction using Markov Chain Monte Carlo, J. Am. Stat. Assoc. **95** (2000) 493–508.

[LI03]. K. Li, ClustalW-MPI: a parallel implementation of Clustal-W, based on MPI Bioinformatics **19** (2003) 1585–1586.

[LI04]. K. Li, Y. Guan, J. Wang, G. Smith, K. Xu, and 17 others, Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia, Nature **430** (2004) 209–213.

[LI05]. W. Li, Z. Shi, M. Yu, W. Ren, C. Smith, and 12 others, Bats are natural reservoirs of SARS-like coronaviruses, Science **310** (2005) 676–679.

[LIP04]. Lipatov et al, Influenza Emergence and Control, J. Virol. (2004) 8951–8959.

[MAR03]. M.A. Marra, S.J. Jones, C.R. Astell, R.A Holt RA, and 47 others, The Genome Sequence of the SARS-Associated Coronavirus, Science **300** (2003) 1399–404.

[MAR04]. B.E. Martina, B.L. Haagmans, T. Kuiken, R.A. Fouchier, G,F. Rimmelzwaan and 5 others, SARS infection of cats and ferrets, Nature **425** (2003) 915.

[MET02]. M. Metzker, D. Mindell, X. Liu, R. Ptak, R. Gibbs, and D. Hillis, Molecular evidence of HIV-1 transmission in a criminal case, Proc. Natl. Acad. Sci. USA **99** (2002) 14292–14297.

[MOI99]. A. Moilanen, Searching for most parsimonious trees with simulated evolutionary optimization, Cladistics **15** (1999) 39–50.

[MOR97]. D. Morrison and J. Ellis, Some effects of nucleotide sequence alignment on phylogeny estimation, Molecular Biology and Evolution **14** (1997) 428–441.

[MOR95]. S. Morse, Factors in the Emergence of Infectious Diseases, Emerging Infectious Diseases **1** (1995) 7–15.

[NIX94]. K.C. Nixon and J.M. Carpenter, On outgroups, Cladistics **9** (1994) 413–426.

[NIX99]. K.C. Nixon, The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis, Cladistics **15** (1999) 407–414.

[OBE06]. E. Obenauer, J. Denson, P. Mehta, X. Su, S. Mukatira, and 12 others, Large-Scale Sequence Analysis of Avian Influenza Isolates, Science, published online January 26 http://www.sciencemag.org/cgi/content/full/1121586/DC1 (2006)

[PAL06]. P. Palese, Making Better Influenza Virus Vaccines?, Emerging Infectious Diseases **12** (2006) 61–65.

[PHI00]. A. Phillips, D. Janies, and W.C. Wheeler, Multiple sequence alignment in phylogenetic analysis, Mol. Phylogenet. Evol. **16** (2000) 317–330.

[PLO02]. J. Plotkin, J. Dushoff, and S. Levin, Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus, Proc. Natl. Acad. Sci. USA **99** (2002) 6263–68.

[PN01]. D. Pol and M.A. Norell, Comments on the Manhattan Stratigraphic Measure, Cladistics **17** (2001) 285–289.

[POE98]. S. Poe, The Effect of Taxonomic Sampling on Accuracy of Phylogeny Estimation, Test Case of a Known Phylogeny, Mol. Biol. Evol. **15** (1998)1086–1090.

[RAN96]. B. Rannala and Z. Yang, Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference, J. Mol. Evol. **43** (1996) 304–311.

[RAN98]. B. Rannala, J.P. Huelsenbeck, Z. Yang, and R. Nielsen, Taxon Sampling and the Accuracy of Large Phylogenies, Syst. Biol. **47** (1998) 702–710.

[RED05]. B.D. Redelings and M.A. Suchard, Joint Bayesian estimation of alignment and phylogeny, Syst. Biol. **54** (2005) 401–418.

[RIC97]. K. Rice and T. Warnow, Parsimony is hard to beat, Computing and combinatorics, (Shanghai, 1997): 124–133 (1997)

[ROS02]. R.S. Ross, S. Viazov, and M. Roggendorf, Phylogenetic analysis indicates transmission of hepatitis C virus from an infected orthopedic surgeon to a patient, J. Med. Virol. **66** (2002) 4617.

[ROS04]. U. Roshan, T. Warnow, B.M.E. Moret, and T.L. Williams, Rec-IDCM3: A Fast Algorithmic Technique for Reconstructing Large Phylogenetic Trees, Proc. IEEE Comp. Syst. Bioinf. Conf. (2004)

[ROT03]. P.A. Rota, M.S Oberste, S.S. Monroe, W.A. Nix, R. Campagnoli, and 30 others, Characterization of a novel coronavirus associated with severe acute respiratory syndrome, Science **300** (2003) 1394–9.

[RUB03]. E.M. Rubin and A. Tall, Perspectives for vascular genomics, Nature **407** (2004) 265–269.

[SAI87]. N. Saitou and M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. **4** (1987) 406–425.

[SAL01]. L.A. Salter and D.K. Pearl, Stochastic Search Strategy for Estimation of Maximum Likelihood Phylogenetic Trees, Syst. Biol. **50** (2001) 7–17.

[SAN83]. D. Sankoff and R. Cedergren, Simultaneous comparison of three or more sequences related by a tree in D. Sankoff and J. B. Kruskal eds. Time Warps, String Edits, and Macromolecules: the Theory and Practise of Sequence Comparison. Addison-Wesley, Reading, MA. (1983) 253–264.

[SCH90]. C. Scholtissek, Pigs as the mixing vessel for the creation of new pandemic influenza A viruses, Med. Principles Practice **2** (1990) 65–71.

[SEA03]. D. Searls, Pharmacophylogenomics: Genes, evolution and drug targets, Nature Reviews Drug Discovery **2** (2003) 613–623.

[SEA96]. D.L Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis, Phylogenetic inference. In: Hillis, D.M., Moritz, C. Mable, B.K. (eds) Molecular Systematics, second edition. Sinauer Associates, Sunderland, 1996.

[SIL97]. J. Silvertown, M. Franco, and J.L. Harper, Plant Life Histories Ecology, Phylogeny and Evolution, Cambridge University Press, Cambridge, 1997.

[SNE00]. Q. Snell, M. Whiting, M. Clement, and D. McLaughlin, Parallel Phylogenetic Inference, Proc. 2000 ACM/IEEE Conf. Supercomp., **35** (2000)

[SNE73]. P.H.A Sneath and R.R. Sokal, Numerical taxonomy The principles and practice of numerical classification, W. H. Freeman, San Francisco. xv + 573 p. (1973)

[SON05]. H. Song, C. Tu, G. Zhang, S. Wang, K. Zheng, and 21 others, Crosshost evolution of severe acute respiratory syndrome coronavirus in palm civet and human, Proc. Natl. Acad. Sci. USA **102** (2005) 2430–2435.

[STA02]. A. Stamatakis, T. Ludwig, H. Meier, and M.J. Wolf, Accelerating Parallel Maximum Likelihood-based Phylogenetic Tree Calculations using Subtree Equality Vectors, Proc. 15 IEEE/ACM Supercomp. Conf.(2002)

[STE00]. T. Sterling, J. Salmon, D. Becker, and D. Savarese, How to Build a Beowulf. A Guide to the Implementation and Application of PC Clusters, MIT press, 2000.

[SUZ02]. Y. Suzuki and M. Nei, Origin and Evolution of Influenza Virus Hemagglutinin Genes, Mol. Biol. Evol. **19** (2003) 501–509.

[SWO02]. D.L. Swofford, Paup: Phylogenetic Analysis using Parsimony (and other methods), Sinauer Associates, Sunderland, 2002.

[SWO91]. D.L. Swofford, When are phylogency estimates from molecular and morphological data incongruent? In: Miyamoto, M.M., Cracraft, J. (eds) Phylogenetic analysis of DNA sequences, Oxford Univ. Press, Oxford, 1991.

[TAU01]. J.K. Taubenberger, A.H. Ried, T.A. Janczeqski, and T.G. Fanning, Integrating historical, clinical and molecular genetic data in order to explain the origin and virulence of the 1918 Spanish influenza virus, Philos. Trans. R. Soc. Lond. B. **356** (2001) 1829–1839.

[TAU97]. J.K. Taubenberger, A. Reid, A.E. Frafft, K.E. Bijwaard, and T. Fanning, Initial Genetic Characterization of the 1918 Spanish Influenza Virus, Science **275** (1997) 1793–1796.

[TAU05]. J.K. Taubenberger, A. Reid, R. Lourens, R. Wang, G. Jin, and T. Fanning, Characterization of the 1918 influenza virus polymerase genes, Nature **437** (2005) 889–893.

[TAU06]. J.K. Taubenberger and D. Morens, 1918 influenza: the mother of all pandemics, Emerg Infect Dis. **12** (2006) 15–22.

[TAY01]. L.H. Taylor, S. M. Latham, and M. E. Woolhouse, Risk factors for human disease emergence, Philos. Trans. R. Soc. Lond. B. Biol. Sci. **356** (2001) 983–989.

[TEH03]. A. Tehler, D.P. Little, J.S. Farris, The full-length phylogenetic tree from 1551 ribosomal sequences of chitinous fungi, Fungi. Mycol. Res. **107** (2003) 901–916.

[THO94]. J.D. Thompson, D.G. Higgins, and T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice, Nucl. Acids Res. **22** (1994) 4673–4680.

[THR04]. J. Thornton, Resurrecting ancient genes, Experimental analysis of extinct molecules, Nat. Rev. Genet. **5** (2004) 366–375.

[TKF92]. J.L. Thorne, H. Kishino, and J. Felsenstein, Inching toward reality: An improved likelihood model of sequence evolution, J. Mol. Evol. **34** (1992) 3–16.

[TWE04]. S. Tweed, D. Skowronski, S. David, A. Larder, M. Petric, and 10 others, Human illness from avian influenza H7N3, British Columbia.

Emerg Infect Dis. http://www.cdc.gov/ncidod/EID/vol10no12/04-0961.htm (2004)

[UNG05]. K. Ungchusak, P. Auewarakul, S. Dowell, R. Kitphati, P. Wattana, and 10 others, Probable Person-to-Person Transmission of Avian Influenza A (H5N1), N Engl J Med **352** (2005) 333–340.

[UKK85]. E. Ukkonen, Algorithms for approximate string matching, Information and Control Archive **64** (1985) 100–118.

[WAN94]. L. Wang and T. Jiang, On the complexity of multiple sequence alignment, J. Comput. Biol. **4** (1994) 337–348.

[WAN05]. Q. Wang, M. Han, J. Funk, G. Bowman, D. Janies, and L. Saif, Genetic Diversity and Recombination of Porcine Sapoviruses. Journal of Clinical Microbiology **43** (2005) 5963–5972.

[WAT81]. L. Watrous and Q. Wheeler, The outgroup comparison method of character analysis, Syst. Zool. **30** (1981) 1–11.

[WEB92]. R.G. Webster, W.J. Bean, O.T. Gorman, T.M. Chambers, and Y. Kawaoka, Evolution and ecology of influenza A viruses, Microbiol. Rev. **56** (1992) 152–179.

[WHE94]. W. Wheeler and D. Gladstein, MALIGN: A multiple sequence alignment program, J. Hered. **85** (1994) 417–418.

[WHE95]. W. Wheeler, Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data, Systematic Biology **44** (1995) 321–331.

[WHE90]. W.C. Wheeler, Nucleic acid sequence phylogeny and random outgroups, Cladistics **6** (1990) 363–367.

[WHE04]. W.C. Wheeler, D. Janies, and J. DeLaet, DNA sequence alignment and parallel processing, McGraw-Hill Yearbook of Science and Technolog, (2004)

[WHE05]. W.C. Wheeler, A. Varon, D. Gladstein, and J. DeLaet, POY. Phylogeny program for optimization of nucleic acids and other data, American Museum of Natural History, http://research.amnh.org/scicomp/projects/poy.php (2005)

[WHE96]. W.C. Wheeler, Optimization Alignment: the end of multiple sequence alignment in phylogenetics?, Cladistics **12** (1996) 1–9.

[WHI03]. K.P. White, Functional genomics and the study of development, variation and evolution, Nature Genetics **2** (2003) 528–537.

[WHO05]. Report on Global Surveillance of Epidemic-prone Infectious Diseases - Influenza (2000)

[WHO07a]. H5N1 avian influenza: timeline of major events; 11 September (2007)

[WHO07b]. WHO. Cumulative Number of Confirmed Human Cases of Avian Influenza A/(H5N1) Reported to WHO; 10 September (2007)

[YAN07]. Y. Yang, M.E. Halloran, J. Sugimoto, Jr I.M Longini, Detecting human-to-human transmission of avian influenza A (H5N1), Emerg Infect Dis. http://www.cdc.gov/EID/content/13/9/1348.htm (2007).

[YUA02]. Y. Guan, J.S.M. Peiris, A.S. Lipatov, T.M. Ellis, K.C. Dyrting, S. Krauss, L.J. Zhang, R.G. Webster, and K.F. Shortridge, Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR, Proc. Natl. Acad. Sci. USA **99** (2002) 8950–8955.

[ZWI02]. D.J. Zwickl and D.M. Hillis, Increased Taxon Sampling Greatly Reduces Phylogenetic Error, Syst. Biol. **51** (2002) 588–598.