Pacific Symposium on Biocomputing 6:544-546 (2001)

Phylogenetics in the Post-Genomic Era

J.E. Steinbachs
The Field Museum
1400 S. Lake Shore Drive
Chicago, IL USA 60605-2496
stein@fmnh.org

P. Kearney
Department of Computer Science
University of Waterloo
Waterloo, ON, Canada N2L 3G1
pkearney@math.uwaterloo.ca

Ready or not, phylogenetics has moved into the post-genomic era of massive and complex data sets. Issues of immediate and emerging importance include large scale phylogenetic inference, whole genome phylogeny, incorporation of structure into phylogenetics, horizontal transfer, and phylogenetics in clinical studies. Papers presented in this session cover these, and other, topics.

Due to the breadth of genetic data now available for phylogenetic analysis, biologists are able to address fundamental evolutionary questions once out of reach. For example, the Green Plant Project [1] and the Ribosomal Database Project [2] contain evolutionary trees on the order of thousands of taxa. Current phylogenetic tools and conventional phylogenetic wisdom do not scale to such massive projects.

Bininda-Emonds *et al.* focus their efforts on the scaling of accuracy using large data sets, like those mentioned above. Employing a simple search algorithm (maximum parsimony without branch swapping), they determined that the number of characters required to estimate 80% of an evolutionary tree can, in fact, scale well as the number of taxa increases. Furthermore, they demonstrate that the scaling of accuracy varies significantly with the level of accuracy required. Most importantly, they point out that the usual strategy in molecular systematics – that of sequencing large numbers of homologous genes – might not always be the best approach to resolving the phylogeny of interest.

Besides having large numbers of sequences, data sets can be large in regards to the number of nucleotides, as found in whole genome data sets. In fact, Koonin [3] and Wooley [4] both point out that one of the most important

current research topics in computational biology involves the construction of species phylogenies from whole genome data. Due to the numerous nuclear and mitochondrial genomes now sequenced, the ability to construct whole genome phylogenies has arrived. Though in its infancy, this area of research promises to develop quickly. In this session, computer scientists and biologists will present, discuss and define ideas essential to the future of whole genome phylogeny.

Moret *et al.* introduce new computational techniques that increase the efficiency of breakpoint analysis of gene order data by two orders of magnitude. This result not only enables the analysis of larger collections of gene order data but it also enables the examination of the breakpoint analysis itself by permitting large scale studies.

Armed with complete mitochondrial genomes from seven *Drosophila* (fruit fly) species with a well-corroborated lineage, Steinbachs *et al.* demonstrate the efficiency of the different genes in recovering the assumed topology, using a variety of phylogenetic methods. Only some of their findings on gene performance compare with previous studies. Surprisingly, the most accurate method (a maximum likelihood model) fails to recover the well-supported topology for more than half of the genes.

Despite years of research, the area of phylogenetics has progressed little in establishing benchmarks and tools for evaluating phylogenetic methods. The development, application and assessment of phylogenetic methods would be enhanced by progress in these areas. Such development requires an ability to evaluate and compare new methods on benchmark data. The application of phylogenetic methods requires a knowledge of which methods work best on certain types of data. Steinbachs *et al.* offer their data as one potential benchmark data set.

Evans and Wareham present algorithms that incorporate secondary structure information into phylogenetic analyses. In particular, they compute distances on pairs of annotated RNA or protein sequences, align those pairs, and compute 3-median annotated sequences from triples of annotated sequences. Only a few research groups to date have begun to incorporate structure into phylogenetic methods. This paper should help to push deeper into this important field of inquiry.

Researchers realized some time ago that horizontal transfer of genes between organisms occurs. This phenomenon poses a problem for traditional phylogenetic methods based on bifurcation. Kim and Salisbury propose a method to determine the organismal phylogeny based on several molecular phylogenies, in the presence of limited lateral gene transfer events.

Finally, the work proposed by Ren *et al.* introduces phylogenetic analysis into the realities of molecular epidemiology. Data in this field of study

often arise from longitudinal sampling schemes. Until this study, no phylogenetic method has been able to accommodate this type of data. In their paper, Ren *et al.* describe an algorithm for constructing the phylogeny and inferring selection, using viral protein-coding DNA sequences collected from different years. This new method may prove to be useful for inspecting the change of selective pressure on the viral gene over time, as indicated by the nonsynonymous/synonymous substitution rate ratio. Finally, the authors also develop a codon-based model, useful for protein-coding genes, to calculate the rate of evolution. Given the large amounts of within-patient viral sequence data now available, the work presented here should have a lasting impact on phylogenetic analysis in molecular epidemiological studies.

We hope that the papers in this session help to stimulate discussion on phylogenetic methods in the post-genomic era. A fruitful interaction among the authors of the papers in these proceedings, along with other attendees of this conference, should facilitate the transfer of knowledge across the different disciplines interested in developing and applying these methods.

References

1. http://ucjeps.berkeley.edu/bryolab/greenplantpage.html
2. http://www.cme.msu.edu/RDP/html/index.html
3. E.V. Koonin, *Bioinformatics* **15**: 265–266 (1999)
4. J.C. Wooley, *J. Comp. Biol.* **6**: 459–474 (1999)