1-1-1994

Taking a Cat Map: Genome Analysis by Supercomputer

Jose V. Lopez
National Cancer Institute, joslo@nova.edu

Computational Biomedical Research

NATIONAL CANCER INSTITUTE

Frederick Biomedical Supercomputing Center
1994

Taking a Cat Map

Genome Analysis by Supercomputer

Jose V. Lopez, NCI/Frederick Cancer Research and Development Center

The cat, or rather its mitochondrial DNA (mtDNA), is almost out of the bag. This is because part of our research in the Laboratory of Viral Carcinogenesis focuses on DNA and protein sequences to infer evolutionary relationships between living, and ofttimes, endangered biota, which includes most species of cats. Sequencing by "walking" along the 17,000 base pairs (bp) of the complete mitochondrial genome of the domestic cat (*Felis catus*) embodies a basic research initiative similar to the Human Genome Project, albeit on a much smaller scale. The results will provide a bioinformatics resource for the design of polymerase chain reaction (PCR) primers or hybridization probes that can be used with other techniques to measure the uniqueness of feline species or subspecies, their geographical distribution and population structure, relative rates of evolution, and other genetical parameters that are central to the field of molecular evolution and population genetics. Mitochondrial DNA remains a pivotal component of these studies due to its compact size and accelerated rate of evolution, maternal transmission, and the resulting lack of recombination between different mtDNA genotypes.

Jose V. Lopez

Information can be seen as the memorization of an initially random ordering. For biological phenomenon, this ordering may be said to occur in the code of DNA. Entry into the computerized databases is one way of memorizing this natural order. The versatility of analytical programs run on the high-speed computers at the Frederick Biomedical Supercomputing Center (FBSC) enables researchers to have several perspectives of the DNA double helix. Of course, the four basic DNA nucleotides (A,C,G,T) comprise the one-dimensional array of macromolecules that make up the genetic code. Programs that allow the prediction of potential higher-order structures in nucleic acids or the translated proteins can provide invaluable spatial information for verifying the primary sequence alignments and assessing the presence of invariant regions due to evolutionary constraints on a molecule. The underlying strategy of comparative analysis can place the observed similarities and differences between various sequences in a meaningful biological context, and possibly reveal cryptic patterns of mutations that could point to important evolutionary trends.

Phylogenetic Inference

One of the first steps involved in the sequence and phylogenetic analysis of novel genes is their identification, often facilitated by alignment with previously characterized sequences. Several efficient computer programs can search expanding sequence databases to find significant matches or produce the alignments that will enable the conclusion that the two or more sequences under comparison are actually homologous (i.e., derived from a single common ancestral sequence). Proving homology is not trivial, and indeed constitutes the empirical denouement for systematists and evolutionary biologists. Homology should not always be assumed, since rapidly diverging gene sequences (like those in the mtDNA regulatory or promoter region) increase the probability that high similarity scores may be only due to a "convergence" of mutations, the antithesis to homology.

Clarification of the evolutionary history or "phylogeny" of the cat with its wild relatives in the *Felidae* family or *Carnivore* order affects various disciplines ranging from natural history of mammalian radiations to the epidemiology of pathogen-host interactions to endangered wildlife management. Molecular evolutionary studies have shown that following the pattern of mutations in mtDNA can, but not always, reflect the actual evolution of a species or higher taxonomic group. Agreement depends on various factors such as the degree of molecular polymorphism present in ancestral

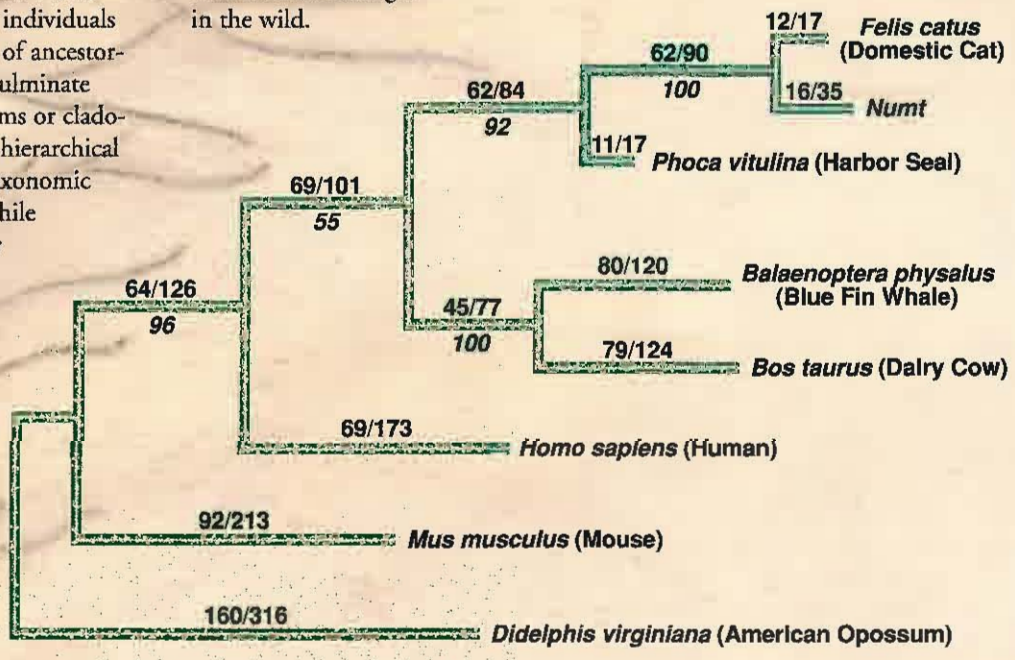
populations and the determination of appropriate evolutionary time scales for the taxa and genes under investigation. Eventually, laboratories must confront the controversial topic of molecular clocks, which is based on the premise that the quantity of observed mutations (e.g., base substitution, deletions or insertions) in various molecules is linearly correlated to the time since they had diverged from common ancestral molecules. Biology, however, almost never conforms to synthetic models, no matter how elegant or logical, for in this regard most molecular clocks behave idiosyncratically, rather than being universally applicable for every gene or taxon. The first clocks were in fact derived from amino acid data, which was highly dependent on the conservation of tertiary structure.

In the same context, the success of phylogenetic reconstructions depends on the essential neutrality of sequences. With adherence to this assumption, mutations will accumulate stochastically, rather than being influenced by external factors such as natural selection or population size, and follow a Poisson distribution. Studying past (unverifiable) events combined with accounting for all of the complex parameters involved in evolutionary analyses justifies the common usage of phylogenetic "inference" to describe the practice.

Since much of the phenotypic diversity between organisms ultimately stems from the genetic code, molecular evolutionists evaluate the informative changes between divergent DNA or amino acid sequences from different organisms or individuals to derive the most accurate phylogeny of ancestor-descendant relationships. The efforts culminate with tree-like diagrams (also phenograms or cladograms), meant to convey a branching, hierarchical ordering between extant operational taxonomic units (OTUs) at the tips of the tree, while internal branch points or "nodes" closer to the trunk of the tree represent older or possibly extinct taxa. OTUs that consistently group together are considered clades, such as all of the cats. Phylogenetic trees should be viewed as discrete scientific hypotheses, formulated to interpret available biological evidence (molecular sequences, fossils, bone measurements etc.) and infer possible historical relationships and taxonomic classification. They may be verified or rejected upon application of alternative chronometers and more definitive data.

The various algorithms designed to construct phylogenies are usually based on specific evolutionary models and philosophies (e.g., most parsimonious evolution, minimum genetic distance, statistical criterion). A similar result for cat mtDNA sequences was found with DNAML, a statistical routine that recreates all possible trees and the associated probabilities of occurring with a specific data set (number of taxa and characters) and model of evolution (i.e., mode of base replacement). This program epitomizes the reliance of biologists on supercomputers to carry out the enormous number of iterations and computations involved in phylogenetic algorithms. To illustrate this, calculations show that 8.87x1023 different branching trees can be created with only 20 tip OTUs, while the time required to compute all possible alternative trees increases exponentially (cube of the number of species). We have submitted data sets consisting of about 370 nucleotides from the mt 12S rRNA gene and more than 35 exotic cat species to DNAML, and discovered that between 28-160 hours of "wall" time was necessary to complete the run on the Cray Y-MP.

Inevitably, through the alliance of biotechnology and bioinformatics, we aim to continue adding sequences to the databases and thereby arrive at a better understanding of evolutionary processes in the cat genome and its endangered relatives. This, in turn, can improve models for human disease and/or aid conservation strategies in the wild.



Phylogenetic tree produced with total 16S large subunit rRNA gene sequences. The topology was created using maximum parsimony criteria, but also reproduced with different (distance) methods, such as Neighbor-Joining. The total tree length is 1615 steps with a consistency index of 0.755. Numbers above the branches designate the ratio of total homoplasies (uninformative parallelisms, convergent mutations etc.)/total changes on that branch. Bootstrap percentages in support of each node derived from at least 100 replications are shown beneath the branches in italics.