Special Topics in Computational Biology

Lecture #2: Phylogeny

Bud Mishra
Professor of Computer Science and Mathematics
2 ¦ 1 ¦ 2001

Bio-Diversity

Life is ubiquitous and old. (3.7 billion years old!)
Living organisms on the Earth have diversified and adapted to almost every environment.
All living organisms can replicate, and the replicator molecule is DNA.
The information stored in DNA is converted into products used to build similar cellular machinery.
Comparative study of the DNA can shed light on its function in the cell and the process of evolution.

Tree of Life

All living organisms are divided into five kingdoms:
- **Protista,**
- **Fungi,**
- **Monera (bacteria),**
- **Plantae,** and
- **Animalia**.

A different scheme:
- **Prokaryotae** (bacteria, etc.)
  - **Bacteria**
  - **Archea**
- **Eukaryotae** (animals, plants, fungi, and protists).

No one of these groups is ancestral to the others.
A fourth group of biological entities, the **viruses**, are not organisms…

Human Evolution

Two Models:
- **Multiregional Model**
- **Out of Africa Model**
  - Evolution of a tree of hominids originating in Africa. Left Africa about 1 million years ago. Two waves of migration are speculated.

African human population has the most diversity.
Australopithecus (3.5million years old), *Homo habilis* (2 million yrs), *Home erectus* (1 million yrs), *Homo sapiens* (60,000-100,000 yrs)
- **Cro Magnon Man** (Our immediate H. sapien ancestor)
- **Neanderthal Man** (Became extinct ~30,000 yrs ago.)

Two distinct species; supported by **DNA amplification and sequence alignment** (S. Paabo)

Mitochondria and Phylogeny

Mitochondrial DNA (mtDNA): Extra-nuclear DNA, transmitted through maternal lineage. Mitochondria are inherited in a growing mammalian zygote only from the egg.
16.5 Kb, contains genes: coding for 13 proteins, 22 tRNA genes, 2 rRNA

genes.
mtDNA has a pointwise mutation substitution rate 10 times faster than nuclear DNA.
Phylogeny based on human mtDNA can give us molecular (hence accurate?) information about human evolution.

African Eve

Statistical analysis of mtDNA extracted from placental tissue of 147 women of different races and regions. (Cann, Stoneking, & Wilson, 87).
Phylogenetic tree (assuming a constant molecular clock) was constructed by Wilson.
A single rooted tree with the root being closest to the modern African woman.
Conclusion: Modern man emerged from Africa 200,000 years ago. Race differences arose 50,000 years ago. "Mitochondrial Eve Hypothesis"

Mitochondrial Eve's Africanness

A simple reordering of the data could result in 100 distinct trees al at most 2 steps away---all supporting non-African hypothesis. (Templeton)
Assuming a non-constant molecular clock results in a least universal common ancestor (Luca) 10^5 to 10^6 years old.
In general, mathematical descriptions and algorithms that may lead to "historically correct phylogenetic tree" remain to be developed.

Taxon

Taxon (Taxonomical Unit): *is an entity whose similarity (or dissimilarity) can be numerically measured*. E.g., Species, Populations, Genera, Amino Acid Sequences, Nucleotide Sequences, Languages.
Phylogeny is an organization of the taxons in a rooted tree, with distances assigned to the edges in a such manner that the "tree-distance" between a pair of taxons equals the numerical value measuring their dissimilarity.
The dissimilarity and the edge lengths of the phylogenic trees can be related to the rate of evolution (perhaps determined by a molecular clock).

Comparing a Pair of Taxons

Discrete Characters: Each taxon possesses a collection of characters and each character can be in one of finite number of states. One can describe an *n* taxons with*m* characters by an *n£m* matrix over the state space. **Character State Matrix**.
Comparative Numerical Data: A *distance* is assigned between every pair of taxons. One can describe the distances between n taxons by an *n£n* matrix over $\mathbf{R}_{+.}$ **Distance Matrix.**

Examples

Tree

A tree is an undirected connected acyclic graph.
Nodes of a tree is either: an **exterior** node (**leaf** node) or an **interior** node. Leaves have degree=1 and interior nodes have degree > 1.
A graph is edge-weighted, if each edge is associated with a real number.
A phylogenic tree is a (positively edge-weighted) tree with each of its leaves labeled by a taxon or a set of taxons.

Characterizing a Phylogenic Tree

Topology: The connection among the nodes. This is also referred to as the branching pattern.
Distance, or Edge-Weights: Distance between an interior node (ancestral taxon) and a leaf node (present-day taxon) is an estimate of the time taken to evolve from one to the other.
If the tree is rooted the ancestry relation is implied by the tree topology.

Often, not enough information in the data exists to determine the root. Only an unrooted tree is constructed.

Character States

Some Assumptions:

The characters are inherited independently from one another.
Observed states of a character have evolved from one “original state” of the nearest common ancestor of a taxon.
Convergence or parallel evolution are rare. That is the same state of a character rarely evolve in two independent manners.
Reversal of a character to an ancestral state is rare.

Classifying Characters

Characters:

Unordered / Qualitative Character: All state transitions are possible.
Ordered / Cladistic Character: Specific rules regarding state transition are assumed.

Linear Ordering
Partial Ordering (with a derivation tree).

Perfect Phylogeny

A **phylogenic tree** *T* is called **perfect**, if for each state *s* of each character *c*, the set of nodes

{ *u* = node : the state at *u* with respect to *c* = *s* }

forms a subtree of *T*.

Perfect Phylogeny Problem:

Given: A set O with n taxons, a se C of m characters, each character having at most r states.

Decide: If O admits a perfect phylogeny.

A set of defining characters are **compatible**, if a set of objects defined by a character set matrix admits a perfect phylogeny.

Binary Character Set

Each character has two states = {0, 1}

If a character is ordered then 0 ! 1 (0=ancestral and 1=derived), or converse.

For binary characters (ordered or unordered), perfect phylogeny problem can be solved efficiently

Poly time, for *n* taxons and *m* characters, Time = *O(nm).*

A two phase algorithm:

Perfect Phylogeny Decision Problem

Perfect Phylogeny Reconstruction Problem

Compatibility Condition

T = Perfect Phylogeny for *M* iff

(8 c_i = character)(9 e = tree-edge) label(e) = {c_i, 0! 1}

root(T) = (0, 0, 0, …, 0)

A path from root to a taxon *t* is labeled (c_{i1}, c_{i2}, …, c_{ij})

) *t* has 1's in positions i_1, i_2, …, i_j.

Perfect Phylogeny Condition

M = *n* £ *m* Character State Matrix, *j* 2 {1..*m*}

O_j = {*i* = taxon : M_{ij} = 1}

O_j^c= {*i* = taxon : M_{ij} = 0}

Key Lemma

Lemma: *A binary matrix M admits a perfect phylogeny iff*

(8 *i*, *j* 2 {1, *m*}) O_i Å O_j = ; or O_i µ O_j or O_i ¶ O_j

Proof: ()) T_i = subtree containing O_i, T_j = subtree containing O_j. r_i = root(T_i) and r_j = root(T_j)

r_i is neither an ancestor nor descendant of r_j) O_i Å O_j= ;

r_i is a descendant of r_j) O_i µ O_j

r_i is an ancestor of r_j) O_i ¶ O_j

(() By induction, Base case m=1 is trivial. Induction case, m=k+1:
T_k = Tree for k characters. O_{k+1} is contained in a subtree with minimal # taxons rooted at ***r***.
r must be a leaf node. Either an edge needs to be labeled or the subtree rooted at ***r*** has to be split. ¤

Simple Algorithm based on the Lemma

Compare every pair of columns for the intersection and inclusion properties. Total of $O(m^2)$ pairs, each comparison can be done in $O(n)$ time.
Total Time Complexity = $O(nm^2)$
Can be improved to $O(nm)$ time.

Improved Decision Algorithm

Algorithm
First radix sort columns of M based on the number of 1's in each column.
for each L_{ij} **do** L_{ij} := 0;
for i := 1 **to** n **do**
    k := -1;
    for j := 1 **to** m **do**
        if M_{ij} = 1 **then** {L_{ij} := k; k := j}
for each column of j of L **do**
    if 9 i,l L_{ij} L_{lj} **and** both nonzero **then**
        return False
return True. ¤

Example
Reconstruction Algorithm
Two Characters

An n £ 2 Character State Matrix with arbitrary number of states admits a perfect phylogeny iff its corresponding *state intersection graph* (SIG) is acyclic.
The SIG, G = (V, E) has at most 2n vetices and $O(n)$ edges.
Acyclicity can be tested in time $O(|V|+|E|) = O(n)$ time.
For two character taxons with arbitrary number of states the perfect phylogeny problem has an efficient solution.

Compatibility Criteria

Allow reversal and convergence properties in the models of evolution.
Parsimony Criteria: Minimize the occurrences of reversal and convergence events in the reconstructed phylogeny tree.
Dollo Parsimony Criterion: Minimize reversal while forbidding convergence.
Camin-Sokal Parsimony Criterion: Minimize convergence while forbidding reversal.
Compatibility Criteria: Exclude minimal number of characters under consideration so that the reconstructed phylogeny tree is perfect and does not admit any occurrence of reversal or convergence.

Computational Infeasibility

Perfect Phylogeny Problem for arbitrary (> 2) number of unordered characters and arbitrary (> 2) number of states is NP-complete.

Optimal Phylogeny Problem under compatibility criteria is NP-complete.

Optimal Phylogeny Problem either under Dollo or Camin-Sokal parsimony criteria is NP-complete.