these methods. Currently, the methods which have been shown to be consistent are implemented by exhaustive search through the space of all tree topologies, and hence such methods – even if accurate on a given input – are computationally infeasible except on very small data sets.

In a recent paper, Farach and Kannan gave a polynomial time algorithm which, with high probability, would return a tree and mutation probabilities on the edges such that the distance (under a metric they define) between the tree they obtain and the model tree is bounded. However, the analysis they give does not imply any guarantee about the probability of obtaining the topology of the model tree.

The topology of the model tree indicates the pattern of the evolutionary events, and thus is significant biologically. In this paper we present a polynomial time method which will obtain, with high probability, the topology of the model tree provided that the mutation probabilities lie in a range that depends upon the sequence length.

The main theorem we obtain is:

Theorem: *Under certain conditions on the edge mutation probabilities, we can with probability* $1-\epsilon$ *obtain the topology of the model tree that generated a set of* n *binary sequences of length* k *in* $O(n^4\alpha(n^4)+nk)$ *time, and* k *may grow sublinearly as a function of* n.

Keywords: evolutionary trees, Cavendar-Farris model of evolution

Address: Tandy Warnow
Department of Computer and Information Sciences
200 S. 33rd Street
University of Pennsylvania
Philadelphia PA 19104-6389
e-mail: tandy@central.cis.upenn.edu

A polynomial time algorithm for solving the Cavendar-Farris phylogeny problem

Péter L. Erdős
Mathematical Institute of the Hungarian Academy of Sciences
Budapest P.O.Box 127, Hungary-1364
E-mail: elp@math-inst.hu Michael A. Steel

Biomathematics Research Centre, University of Canterbury
Christchurch, New Zealand
E-mail: m.steel@math.canterbury.ac.nz

László A. Székely
Department of Computer Science Eötvös University
Budapest, Hungary-1088
E-mail: szekely@cs.elte.hu

Tandy J. Warnow
Department of Computer and Information Science
University of Pennsylvania
Philadelphia PA 19104-6389
E-mail: tandy@central.cis.upenn.edu

Evolutionary history inference is a fundamental and yet computationally difficult problem in biology. The Cavendar-Farris model of evolution is frequently assumed when the taxa (i.e. observed sequences) at the leaves of the tree are binary sequences, although this model can be easily generalized (and has been) to the general case, such as with four letter alphabets (for DNA and RNA sequences) or twenty letter alphabets (for amino acid sequences). Under this model, the model tree M is a fixed but unknown tree T, mutation probabilities $p(e)$ assigned to each edge $e \in E(T)$, a probability distribution over the sequences at the root, and a parameter k indicating the length of the sequences. In this model the positions evolve identically and independently (iid), and $0 < p(e) < .5$ for every edge $e \in E(T)$. Given the observed sequences which label the leaves, the problem is to infer the model tree.

One aspect of this problem that has been studied in detail is the *consistency* question: i.e., whether a given method can, with infinitely long sequences, determine the model tree, with results generally indicating that many methods are not consistent over sections of the parameter space, but that some methods (notably the maximum likelihood methods when based upon the assumptions of the model that generated the data) are consistent. There are serious limitations to these results. First, these results analyze the performance of methods on infinite data rather than on finite data, and hence do not provide any performance guarantees for real-world instances, and secondly, these studies have not considered the computational complexity of applying

one character, but also present algorithms to solve exactly or approximate OTR. In particular, we present a 2-approximation and a PTAS for OTR.

The motivation for studying the OTR problem is twofold. First, we will show that our algorithms for OTR can be used to solve the parsimony problem efficiently for very good data (that is, for data for which either the compatibility or parsimony scores are close to optimal). Because most biological datasets are analyzed with respect to this criterion, the major phylogenetic software packages (Phylip, PAUP, etc), use heuristics to find locally optimal trees, and only provably solve the parsimony problem for small (typically at most 14 or so species) input sets. Hence, new techniques for solving the parsimony problem may prove to help resolve some of the harder controversies in the field; at the very least, however, better algorithms for parsimony will help us analyze the performance of the existing techniques!

However, OTR is also of interest in its own right. Character data are derived from either biomolecular sequences or from morphological features, and a recurring issue in phylogenetic analysis is how to utilize characters which have values only on a proper subset of the species set[1]. The question that arises in this case is how to infer the evolutionary history of the subset of species when the characters that apply to the entire set suggest a tree that differs (potentially significantly) from the tree suggested by the enriched set of characters.

The results in this paper are as follows:

- We have introduced an interesting and realistic problem in phylogenetic tree construction called the Optimal Tree Refinement problem (OTR), and presented two exact algorithms, a 2-approximation algorithm, and a polynomial time approximation scheme for this problem.
- We have shown that the exact algorithms for the OTR problem can be used to solve the bounded imperfection parsimony (BIP) problem.

Each of these problems is relevant to practice for phylogeny construction.

Keywords: evolutionary trees, parsimony

Address: Tandy Warnow
Department of Computer and Information Sciences
200 S. 33rd Street
University of Pennsylvania
Philadelphia PA 19104-6389
e-mail: tandy@central.cis.upenn.edu

[1]The distinction between a species having an unknown state for a character, versus a character being inapplicable to a given species, can in some cases be critical. This was one of the fundamental methodological issues raised in the *Green Plant Phylogeny Coordination Grant Workshop,* held March 1996 in Baton Rouge, LA.

Bounded imperfection parsimony and optimal tree refinement

Mike Steel and Tandy Warnow

We introduce a new problem in evolutionary tree construction which we call the Optimal Tree Refinement (OTR) problem. We show that OTR is in general NP-hard, but provide exact algorithms, a 2-approximation, and a polynomial time approximation scheme (PTAS) for OTR. The OTR problem arises in several contexts, one of which is the problem of how to analyze characters which apply only to a subset of the species set. We formulate the problem of bounded imperfection parsimony, and show how OTR algorithms can be used to find the most parsimonious tree on data sets of bounded imperfection.

Evolutionary tree construction problems typically presume that the input species sets S is described by qualitative characters, where a qualitative character c is a function $c : S \rightarrow Z$ and Z denotes the set of integers. Given a set S of species described by a set C of characters, the output is a labelled tree T called a *phylogeny for* (S, C), in which the species in S label the leaves of the tree T, and each internal node is assigned a state for each character. Thus, the nodes in T are labelled with integer vectors of length $|C|$, so that these labels *extend* the domain of the characters in C to the set $V(T)$.

Two of the most frequently used criteria used to evaluate phylogenetic (evolutionary) trees are the parsimony and compatibility criteria. Let $cost(c, T)$ be the number of edges in T for which character c has different values at the endpoints. Note that $cost(c, T) \geq r_c - 1$, where r_c is the number of states that character c exhibits on S. When $cost(c, T) = r_c - 1$ then the character c is said to be *compatible* (or *convex*) on T. The *parsimony score* of T is $pscore_C(T) = \sum_{c \in C} cost(c, T)$. An equivalent definition of the parsimony score is $\sum_{e \in E} H(e)$, where $H(e)$ denotes the Hamming distance between the endpoints of the edge e, i.e. $H(e) = |\{i, 1 \leq i \leq |C| : v_i \neq w_i, e = (v, w)\}|$. Given (S, C), finding a phylogeny T for (S, C) of optimal parsimony score is the *parsimony problem*. The *compatibility score* of T is $cscore_C(T) = |\{c \in C : c \text{ is convex on } T\}|$. When the set C of characters is clear from the context, we may omit the subscript C. Finding trees which minimize the parsimony score or maximize the compatibility score are both NP-hard problems, even for binary (two state) characters. In this paper we present a new algorithm for the parsimony problem. One of the techniques we use for solving the parsimony problem is based upon a new problem which we pose, which we call the *Optimal Tree Refinement Problem*. A leaf-labelled tree T' *refines* the leaf-labelled tree T if T can be obtained from T' by contracting edges in T'. The OTR problem is as follows: we are given a tree T and a set of characters C, and we seek a refinement T' of T of optimal parsimony score with respect to the set C. We show that OTR is NP-hard, even for

Phylip, PAUP, and MacClade), algorithms and software to *evaluate* fixed leaf-labelled tree topologies for polymorphic data have explicitly required that the number of states be kept quite small because the evaluation requires time exponential in the number of states. This is the first algorithmic study of this problem to go beyond fixed topology problems for bounded number of states.

The major contribution of this paper is a methodology for inferring perfect phylogenies from monomorphic and polymorphic characters, and an application of this methodology to a highly debated data set – the Indo-European (IE) family of languages. Recently, Warnow and Ringe, a historical linguist at the University of Pennsylvania, showed that perfect phylogenies could be obtained for the IE family from properly selected and encoded linguistic characters. Their research (with postdoctoral researcher Ann Taylor) into the Indo-European family of languages revealed the extensive prevalence of polymorphism, and indicated that without an appropriate method for handling polymorphic characters, a definitive analysis of the history of that family would be difficult to obtain. Our methodology has been applied to the IE family, and shows surprisingly strong support for a number of decades long conjectures in the field, and potentially settles significant controversies such as the *Indo-Hittite hypothesis*, the *Italo-Celtic hypothesis*, the dual-allegiance of the Germanic subfamily, and the relationship between Greek and Armenian. Our methodology and preliminary results were presented at the Symposium on the Frontiers of Science at the National Academy of Sciences in November 1995, and will appear in an upcoming paper in the Proceedings of the National Academy of Sciences.

The paper has the following structure. We discuss the causes of polymorphism in Linguistics and Biology, and define the problem of inferring trees from polymorphic characters in these two domains. We show that a perfect phylogeny is an appropriate objective when working with linguistic data as well as some biological data. We present two algorithms, one graph theoretic and one combinatorial, for the problem of inferring perfect phylogenies from polymorphic data. We then present a methodology for inferring perfect phylogenies from data which combine monomorphic and polymorphic data. We present our analysis of the Indo-European data studied by Warnow, Ringe, and Taylor. We also consider the problem of inferring evolutionary trees from polymorphic data when a perfect phylogeny is an unlikely outcome (as is the case for biomolecular data), and present algorithms and hardness results.

Keywords: evolutionary trees, linguistics, polymorphism

Address: Tandy Warnow
Department of Computer and Information Sciences
200 S. 33rd Street
University of Pennsylvania
Philadelphia PA 19104-6389
e-mail: tandy@central.cis.upenn.edu

Inferring evolutionary trees from polymorphic characters, and an analysis of the Indo-European family of languages

Maria Bonet, Cindy Phillips, Tandy Warnow, and Shibu Yooseph

Most phylogenetics literature and construction methods based upon characters presume monomorphism (one state per character per species), and yet polymorphism (multiple states per character per species) is well documented in both Biology and Historical Linguistics. In this paper we consider the problem of inferring evolutionary trees for polymorphic characters. We show efficient algorithms for the construction of perfect phylogenies from polymorphic data. These methods have been used to help construct the evolutionary tree proposed by Warnow, Ringe, and Taylor for the Indo-European family of languages, which was presented by invitation at the National Academy of Sciences in November 1995.

Determining the evolutionary history of a set S of objects (taxa or species) is a problem with applications in a number of domains such as Biology, Comparative Linguistics, and Literature. Primary data used to compare different taxa (whether biological species, populations, or languages) can be described using *characters*, where a character is a function $c : S \rightarrow Z$, where Z denotes the integers and thus represents the set of possible *states* of c. In this paper we consider tree construction when characters are permitted to have more than one state on a given object. We call this the *polymorphism problem*. A character which is permitted to have more than one state on a given object will be called a *polymorphic character*, and one which can have only one state for every object is referred to as a *monomorphic* character.

Polymorphism is well-documented in both the molecular genetics and comparative linguistics domains. For example, the population geneticist Masatoshi Nei writes: *The study of protein polymorphism has indicated that the extent of genetic variation in natural populations is enormous. However, the total amount of genetic variation cannot be known unless it is studied at the DNA level. The study of DNA polymorphism is still in its infancy, but the results so far obtained indicate that the extent of DNA polymorphism is far greater than that of protein polymorphism. Polymorphism also arises in the compa rison of different languages. The Indo-Europeanist Donald Ringe writes:* In choosing lexical characters we try to work with basic meanings (semantic slots), choosing from each language the word that most usually expresses each basic meaning. Languages typically have one word for each basic semantic slot, but instances of two (or even more) words apparently filling the same basic slot are not rare.

Thus, polymorphic data is a reality when working with evolutionary tree construction for both linguistic analysis and biological taxa, and methods appropriate for such construction must be devised. In the phylogenetics literature and programs (such as

expression *purely* in terms of morphisms. Thus, the idea slowly grew that perhaps one should take *morphisms as atomic* instead of elements as is traditionally done when we use Set Theory. This is the Category Theoretic view of Mathematical structures. (For a beautiful introduction to Category Theory, which moreover assumes only high school level mathematics, see the book **Conceptual Mathematics** by *F.W.Lawvere and S.Schanuel*, Buffalo University Press)
What is important to us in this abstract is that, this theory has all the features listed as desirable earlier.
It provides as objective — in the precise sense of presentation invariance — a view of structures as possible. It has mechanisms — Functorial semantics — which enables us to say that an object has structure X relative to an universe which is already Y-structured. Moreover, it has a method of measuring the amount of structure, in the sense of having natural hierarchy of structures definable within the theory. It has natural relation between structures — Tripleability is one example — which are Non-transitive. Roughly, Tripleability is a kind of 'Graphical Reconstructibility'. Functional correlate aspect will be treated in the next section.

5. Towards Reasonable Hierarchies

There could be many reasons why systems are organized in a Hierarchical manner; the often cited reason for Artificial Systems is complexity management. For Biological systems it is could be independence of mutations, and so on. Here we are concerned with artificial systems, in particular, those occurring in Computer Science. In this context it is possible to replace 'complexity' with the more precise measure of 'complexity of provability of system properties'. We would like to say that systems occurring in computer science are hierarchical to enable their builders to reason about them efficiently. I call such hierarchies *Reasonable Hierarchies*. Building systems whose properties could be proved efficiently is one of the main themes of Computer Science research. My current research involves describing system specifications using the category theoretic method of *Sketches*, finding suitable non-transitive relations between such descriptions, relating them to the complexity of verifying propertirs of systems and eventually develop tools to build Reasonable Hierarchies.

Keywords: Non-transitive relations, category theory, sketches, reasonable hierarchies

Address:
P.S.Subramanian
Theoretical Computer Science
Tata Institute of Fundamental Research
Homi Bhabha Road
Bombay- 400 005, INDIA
e-mail: cayley@tifrvax.tifr.res.in

It is interesting to speculate on what more features should be added to this list.

3. The Importance of Being Non-transitive

Relations on a set are classified according to some properties- *Reflexive*- every element is related to itself; *Symmetric*- if a is related to b then b is also related to a; *Transitive*- if a is related to b and b related to c then a is also related to c.

A Non-transitive relation is one in which the last property fails to hold. The most basic example is of course the *immediate successor of* relation between Numbers: Note however, that the *greater than* relation is transitive. Sometimes, the occurrence of non-transitive relation surprises us- as in relativity theory which replaces simultaneity- which is transitive- with concurrency which is not. The surprise element in M.Escher's drawings is also due to non-transitivity.

What has all this to do with Hierarchy?

It is the hidden presence of some non-transitive relation which induces Hierarchy.

Supposing, we have a space of system oragnisations achieving certain goals and moreover in a logical sense the same goal could have been achieved either with *flat* hierachies or *tall* hierarchies, but only *tall* ones are observed to occur.

A mathematical explanation of this could involve some suitable cost function defined over the space of hierachies and a proof that the *tall* ones minimise (or maximise) this function. Now any function induces an equivalence relation on it's domain of definition- two elements (in this case hierarchies) belong to the same equivalence class iff they map to the same value. Now our explanation has to separate *flat* and *tall* hierarchies- i.e we do not want these two types occurring in the same class. Now comes the crucial point- if all meaningful relations we can define on the nodes of the hierarchy were transitive, there is nothing to prevent 'telescopic collapse' within the same class. This will contradict our main requirement of separation. Therefore, there must exist some non-transitive relation witnessing the separation.

Once we are convinced of this arguement, we can say that the study and classification of Hierarchies can be conducted mathematically by suitable identifications of cost spaces, minimisation or maximisation requirements and identification of the associated non-transitive relations.

Is there a surprise element of non-transitivity in this context? Possibly, it is the *apparent* failure of 'reductionism'. Computer Scientists will see an it apparent failure of compositionality principle in this situation.

4. Category Theory as a suitable formalism

Modern mathematics studies *Structures* . How are these structures presented? It has been customary since the advent of Set Theory to present structures as sets with additional relations imposed on their elements. Then, to compare structures, one considers *homomorphisms*- they are functions in the sense of set theory which preserve the additional relations. Thus, homomorphisms in this setup are *non-atomic* entities- they have some intricate internal detail. It was realised around 1945 that many important properties of structures and constructions on them admit of natural

Towards reasonable hierarchies

P.S.Subramanian

Summary
Reasonable Hierarchies are those about which we can *reason* efficiently. We consider this concept in the context of developing mathematical formalisms for a *Science of the Artificial*, in the sense of Prof. H.A.Simon. One of the important desideratum of such formalisms is identified to be the natural occurrance of *Non-transitive relations* within the mathematical framework, together with appropriate functional correlates of these relations. In this talk we identify a formalism based on Category Theory, possible Non-transitive relations within such a formalism and the corresponding functional correlates.

1. Introduction

In an eloquent booklet published in 1969, Prof. H.A.Simon had raised the problem of building a *Science of the Artificial*. An important aspect of this problem is the search for a mathematical basis for the notion of Hierarchy. In this abstract, I describe some of my efforts towards this goal. The organisation of the abstract is as follows. In the second section I list some desirable properties of a mathematical theory which is to serve as a formal basis for the concept of Hierarchy. One of these properties is the natural occurrance of *non-transitive* relations within the theory. In the third section, I supply motivations and arguments for viewing the presence of such relations as not merely desirable, but even necessary. In the fourth section, I identify the Mathematical theory of Categories as a suitable formalism and justify the choice of this theory in terms of the properties identified earlier. In the concluding part, I briefly indicate how this formalism can be used to define the concept of a Reasonable Hierarchy, and it's importance to some central problems in Computer Science.

2. Desirable features of a Mathematical Theory of Hierarchy

Before we embark on the search for a suitable mathematical formalism for the concept of Hierarchy, let us ponder over what features such a formalism *should* possess. This will help us to focus the search. Here is partial list of features; I regard the first three as self evident. More is said about the last two in the later sections.

A. The theory must have at it's disposal a way of talking about the elusive notion of structure in as *objective* and representation independent way as possible.

B. The theory must be able to deal with the notion of *relative* structure.

C. The theory must have (at least) a semi-quantitative way of talking about the *amount* of structure.

D. There must exist *natural* relations between structures which are *non-transitive*.

E. There must be *functional correlates* for such relations.

give conditions under which the two models may be distinguished if the full matrix is examined, (iii) that the two models can, in principle, be distinguished when there are at least four monophyletic groups of species. In particular, with a view to a possible test of the covarion hypothesis (against a rates-across-sites model) we construct a distance measure which is (tree) additive under certain versions of the covarion model (satisfying a certain separability condition) but which, in general, will not be additive under a rates-across-sites model. Such a measure may also be useful for reconstructing the tree on the monophyletic groups when the covarion model applies.

Keywords: Trees, tree metrics, isolation index, nucleotide substitution, covarion model, Markov processes, moment generating function.

Address: Biomathematics Research Centre
University of Canterbury
Private Bag
Christchurch, New Zealand
e-mail: m.steel@math.canterbury.ac.nz

Dissimilarity maps and substitution models - some new results

Mike Steel

In part one we describe some new results on reconstructing trees from distance measures, based on Peter Buneman's pioneering (1971) paper. In part two we analyse a covarion-style model, of the type suggested by Walter Fitch in 1971 that offers a plausible explaination for why sequence sites "appear" to evolve at different rates.

Part One: [Joint work with Vincent Moulton]
Trees with positively-weighted edges induce a natural metric on any subset of vertices, however not every metric is representable in this way. A problem arising in areas of classification, particularly in evolutionary biology, is how to approximate an arbitrary distance function by such a tree metric, and thereby estimate the underlying tree that generated the data. Such transformations, from distances to tree metrics (and thereby to edge-weighted trees) should have some basic properties such as continuity, so that a small change in the input data does not result in a drastically different tree, but this is lacking in several popular methods, for example (as pointed out by Buneman) in methods that attempt to find a closest fit tree metric, and (as we show) in the popular *neighbor joining* method. However known continuous transformations, such as Buneman's original construction, often produce ininteresting (unresolved) trees, leading Buneman to suggest that perhaps this was "the price paid for continuity". One way to extract more information (and continuously!) from distances is Bandelt and Dress' elegant *split decomposition* theory. Based on a modification of Buneman's construction, this continuous map produces a much larger number of splits than Buneman's map, though these splits generally do not form a tree. Here we suggest an alternative modification to the Buneman construction that always leads to trees, and which are, in general, more resolved than those obtained via Buneman's construction. Yet we can achieve this goal without sacrificing continuity, and still in polynomial time. This suggests the possibility of finding other such maps.

Part Two: [Joint work with Chris Tuffley]
A "covarion" model for nucleotide substitution which allows sites to turn "on" and "off" with time was proposed 25 years ago by Fitch and Markowitz. It has been argued that evidence supports such models over later, alternative models which postulate a static distribution of rates across sites. However, in contrast to these latter well-studied models, little is known about the analytic properties of the former model. Here we analyse a covarion-style model and show (i) how to obtain the evolutionary distance between two species from the expected proportion of sites where two species differ (ii) that the covarion model cannot be distinguished from a suitably chosen rates-across-sites model on pairs of taxa if only the trace of the joint probability matrix is considered (i.e. the probability that the two taxa are in the same state) and

Testing fit of the "introns-early" and the "introns-late" theories to the aldehyde dehydrogenase genes

Andrey Rzhetsky (1), Lily C. Hsu (2), Cheng Chang (2), and Akira Yoshida (2)

(1) Columbia University, 630 West 168th Street, BB 16-1611, New York, NY 10032

(2) Department of Biochemical Genetics, Beckman Research Institute of the City of Hope, Duarte, CA 91010

Whether or not the nuclear introns predate the divergence of bacteria and eukaryotes is in the center of argument where proponents of two competing theories, the "introns-early" and the "introns-late", argue that the same data confirms one theory while rejecting the other. In this study we compared the "goodness-of-fit" of the two theories using a probabilistic model of exon/intron evolution and new genomic sequences of non-allelic genes encoding human aldehyde dehydrogenases (ALDH). We first reconstructed a tree for ALDH sequences from bacteria and eukaryotes which indicated that the common ancestor of eukaryotes and bacteria had at least four different ALDH genes. Then, assuming that the tree is correct, we computed likelihoods of obtaining the present-day ALDH sequences under assumptions of the two competing theories. The result was that on the grounds of its own assumptions each theory fitted the same ALDH data significantly better than the rival theory. However, the model corresponding to the "introns-early" theory assumes possibility of "intron slippage" and the maximum likelihood estimates of the slippage rates were too high to be consistent with the previously reported correlation between boundaries of the ancient protein modules and ends of the ancient exons. We also argue that the molecular mechanisms proposed for explaining intron slippage are incapable of providing high slippage rates and are incompatible with the observed intron distribution in the higher eukaryotes. Therefore, we claim that the assumption of intron slippage leads to an internal contradiction in the argument supporting the "introns-early" theory and conclude that ALDH data support the "introns-late" theory.

Address: Andrey Rzhetsky
Columbia Genome Center
Columbia University,
630 West 168th Street -BB 16-1611,
New York, NY 10032
e-mail: andrey@genome2.cpmc.columbia.edu

Phylogeny numbers

Fred S. Roberts and Li Sheng

Motivated by problems of phylogenetic tree reconstruction, we introduce notions of phylogeny graph and phylogeny number. These notions are analogous to notions of competition graph and competition number that arise from problems of ecology. Given an acyclic digraph $D = (V, A)$, define its phylogeny graph $G = P(D)$ by taking the same vertex set as D and letting $xy \in E(G)$ if and only if $(x, y) \in A$ or $(y, x) \in A$ or $(x, a), (y, a) \in A$ for some vertex $a \in V$. Given a graph $G = (V, E)$, we shall call the acyclic digraph D a phylogeny digraph for G if G is an induced subgraph of $P(D)$ and D has no arcs from vertices outside of G to G. The phylogeny number $p(G)$ is defined to be the smallest r such that G has a phylogeny digraph D with $|V(D)| - |V(G)| = r$. In this paper, we obtain results about phylogeny number analogous to a number of well-known results about competition number. In particular, we show that the computation of $p(G)$ is NP-complete and we calculate the phylogeny number for triangulated graphs, line graphs, and graphs with zero or one triangle.

Address: F.S.Roberts
Math Department (Hill Center)
Rutgers Universuty
New Brunswick, NJ 08903
e-mail: froberts@dimacs.rutgers.edu

of the two input trees. All $o(n^2)$ dynamic programming algorithms for the *MAST* problem must find a sparse subset of significant local computations to perform. In Farach and Thorup (FOCS94), this was done implicitly by a fairly intricate recursive scheme specifically designed to give an optimal algorithm for the case of unbounded degrees. As mentioned above, for binary trees their recursion takes $O(nc^{\sqrt{\log n}})$ time. Our algorithm is a simple dynamic program tailored to binary trees. A key feature of our algorithm is that we can identify all the local computations initially before the dynamic program starts. We are thus able to radically simplify the method for actually computing the needed values, and therefore we have an algorithm which is both more efficient and much simpler. It solves the *MAST* problem for binary trees by performing $O(n \log^2 n)$ local computations in $O(n \log^3 n)$ total time. The main idea is to combine a decomposition of the trees into paths with dynamic programming algorithms for two other optimization problems: a variant of the Maximum Weight Common Subsequence Problem (which, because of the interpretation in terms of bipartite graphs used in this paper, we call the Maximum Non-Crossing Matching problem) and a problem that we call the Maximum Crossing problem.

Recently, Cole and Hariharan reduced the running time of the algorithm to $O(n \log n)$. They provided two improvements each of which gains a $\log n$ factor. The first improvement is to reduce the number of interesting pairs to $O(n \log n)$ and the second (and more difficult), to reduce the average processing time for each interesting pair to $O(1)$. The SW approach is naturally generalized to k trees, but there are $\Theta(n^k)$ k-tuples of subtrees, so the size of the dynamic program is exponential in k. It seems that none of our sparsification ideas for the SW algorithm for two trees can avoid exponentiallity in k. The Amir-Keselman bound of $O(kn^{d+1} + n^{2d})$ is achieved by computing local *MAST* values for certain k-tuples of subtrees. Their sparsification is implicit in their recursive formulation of the problem. Our improvement to $O(kn^3 + n^d)$ is also based on a dynamic program, but a key is that we are **not** computing *MAST* values of k-tuples of subtrees. We are still considering such k-tuples, but, in contrast to all other efficient algorithms for *MAST* problems, we *compute a different function on the k-tuples*, allowing for our substantial savings.

Address: T. Przytycka
Dept. of Computer Science
University of Maryland
A.V. Williams Bldg.,
College Park, MD 20742
e-mail: przytyck@cs.umbc.edu

Sparse dynamic programming techniques for maximum agreement subtree problem

Teresa M. Przytycka

The Maximum Agreement Subtree problem arises in biology as a measure of consistency between two evolutionary trees. An *evolutionary tree* for a species set A is a rooted tree in which the leaves are uniquely labeled by the species in A, and the internal nodes represent ancestors. The set of leaf labels of a tree T is denoted $L(T)$. The *size*, n, of a tree is defined to be the number of leaves. Internal nodes of an evelutionary tree are assumed to have at least two children.

Given a leaf labeled tree T on set L, and given $L' \subseteq L$, then the *topological restriction of T to L'*, written $T|L'$, is the tree with vertex set $\{\mathtt{lca}^T(a,b) | \, (a,b) \in L' \times L'\}$, where the arcs are defined such that for all $(a,b) \in L' \times L'$, $\mathtt{lca}^{T|L'}(a,b) = \mathtt{lca}^T(a,b)$. Here $\mathtt{lca}^T(x,y)$ is the least common ancestor of nodes x and y in T, for any tree T. Such a tree is uniquely determined by L'.

Given two leaf labeled trees T_0, T_1 on the same set L, the *Maximum Agreement Subtree* problem (*MAST*) is to compute a maximum cardinality subset L' of L such that $T_0|L'$ and $T_1|L'$ are isomorphic. (Isomorphism of leaf labeled trees is assumed to preserve the labels.) The unrooted version of the problem, (*UMAST*), is also biologically important.

Following a heuristic algorithm of Finden and Gordon 85 and an $O(n^{(\frac{1}{2}+\epsilon)\log n})$ time algorithm of Kubicka *et al.* 92 the first polynomial time algorithm for computed *MAST* for the *MAST* and *UMAST* problem was given by Steel and Warnow 93. Their algorithms are dynamic programming alorithms run in $O(n^2)$ time on bounded degree unrooted trees and in $O(n^{4.5} \log n)$ time on unbounded degree unrooted trees. This algorithm computes the local solution to the *MAST* problem for every pair of rooted subtrees of the two input trees. However, as was first observed by Faracha and Thorup FOCS94, instances of dynamic progamming that corresponds to the *MAST* problem are *sparse* thus to compute *MAST* using dynamic programming technique it suffices to compute the agreement between significantly smaller number of subtrees. Since then several subqadratic algorithms for *MAST* and *UMAST* problems have been proposed.

In practice, evolutionary trees have small degree, and most frequently they are simply binary trees. Thus the *MAST* problem for binary trees is of particular interest. In this paper, we give an $O(n \log^3 n)$ time algorithm for the Maximum Agreement Subtree problem for binary trees. As with the SW algorithm, our algorithm uses dynamic programming. The quadratic cost of the SW algorithm follows from the fact that it computes the local solution to the *MAST* problem for every pair of rooted subtrees

Median weak hierarchies

R. McMorris and R. C. Powers

A weak hierarchy on a set S is a collection H of non-empty subsets of S such that for any $A, B, C \in H$, $A \cap B \cap C \in \{A \cap B, A \cap C, B \cap C\}$. Weak hierarchies were introduced into the cluster analysis literature in 1989 as a generalization of hierarchies so that overlapping clusters are allowed. Let $\mathcal{W}$ be the set of all weak hierarchies on S and d the symmetric difference metric on $\mathcal{W}$. A median for $H_1, ..., H_k \in W$ is a weak hierarchy H such that $\sum_{i=1}^{k} d(H, H_i)$ is minimum. The median function on $\mathcal{W}$ is the function that returns the set of all medians for any given k-tuple of weak hierarchies. In this paper we give a characterization of the median function on $\mathcal{W}$ building on previous work involving characterizations of the median functions on the set of all hierarchies and other ordered sets.

Keywords: median, weak hierarchy

Address: R.C. Powers
Department of Mathematics
Natural Science Building
University of Louisville
Louisville, KY 40292
e-mail: rcpowe01@ulkyvx.louisville.edu

Complexity issues in hierarchical optimization

Panos M. Pardalos

In recent years hierarchical decision processes received an increasing interest in optimization within the area of multi-level programming. Although many problems of practical interest can be formulated as multi-level programming problems, most of these problems are NP-hard from the complexity point of view. We are going to discuss recent results regarding the complexity of various types of multi-level problems and the importance of these results in future algorithmic developments.

Address: Panos Pardalos
Center for Applied Optimization
Industrial and Systems Engineering Department
303 Weil Hall, University of Florida
Gainesville, FL 32611-6595
e-mail: pardalos@ufl.edu

Reconciling incongruent gene and species trees

Roderic Page

Incongruent gene and species trees can be reconciled by embedding the gene tree in the species tree. This embedding incurs a cost in terms of postulated gene duplications and losses. One way of visulazing this embedding is to construct a reconciled tree, which is a generalization of n-trees to m-sets. If the leaves of the gene tree are labeled with the name of the organism from which the gene was obtained, then the gene tree is a subtree of the reconciled tree. The number of branches pruned from the reconciled tree to yield the gene tree corresponds to the number of gene losses incurred by embedding the gene tree into the species tree. A limitation of reconciled trees is that they exclude horizontal transfer. Incorporating this process adds an extra level of complexity. In particular, horizontal transfer establishes a temporal link between edges in the species tree. Typically evolutionary trees are n-trees and hence lack temporal information (unlike, for example, dendrograms), which increases the difficulty of ensuring that sets of two or more horizontal transfers are compatible. Progress twowards solving this problem is outlined.

Address: R.D.M. Page
Division of Environmental & Evolutionary Biology, IBLS
University of Glasgow
Glasgow G12 8QQ
Scotland, U.K.
e-mail: r.page@bio.gla.ac.uk

event) from parent X to offspring Y that no intermediate form is possible. That is, if Z is intermediate between X and Y then Z must be identical to X or Y. So the changes from generation to generation proceed as continuously as can be. Now one may ask, How different can the X and Y be in terms of complexity? How big an "evolutionary jump" is possible in one such step? The answer is: the complexity of Y can (at most) slightly exceed twice that of X! This is shown by proving an upper bound and then explicitly constructing appropriate X and Y that realize it.

This result due to Rhodes (circa 1970, unpublished) leads to the conclusion that even with strong assumptions on continuity during evolution, a large jump in the complexity hierarchy is possible. It also provides intuition about the manner in which such a jump may occur. In particular, the offspring Y can more than double the complexity of X in some cases where X is not using its resources / structure / organization effectively. In the examples constructed, essentially non-interacting components of X are suddenly combined to achieve a new grade of hierarchical organization in Y. This suggests that such evolutionary transitions in complexity may rely on "fragmentation" or lack of coordination between existing components for their genesis. After the jump, another such jump from the resulting Y is impossible since Y's structure is too well-integrated to permit another advance. (If Y survives, generations later its progeny might once again be ripe for another jump in complexity.) Duplication of genes or "gene-nets" (which does not change the complexity as defined above) and the subsequent divergence of the copies in evolution might provide just this type of fragmentation as a source of raw material for evolutionary jumps.

As L. Buss has observed, the Modern Synthesis in evolutionary theory has not generated a theory of ontogeny, relating the origin of higher level organizational grades of units of evolution to genetics and development (as it presupposes the existence of the individual on which natural selection acts). Meanwhile, the molecular revolution in biology is making it possible to understand not only gene frequencies, but also epistatic interactions (including hierarchical ones) between parts of the genome, for instance, homeobox control of vertebrate development.

References

L. Buss, The Evolution of Individuality, Princeton, 1987.
J. Tyler Bonner, The Evolution of Complexity, Princeton, 1988.
C.R. Marshall & J. Wm. Schopf, eds., Evolution and the Molecular Revolution, Jones and Bartlett, 1996.

Address: (corresponding author)
C. Nehaniv University of Aizu
Aizu-Wakamatsu City, 965-80 Japan
e-mail: nehaniv@u-aizu.ac.jp

Krohn-Rhodes theory, hierarchies, and evolution

Chrystopher L. Nehaniv & John L. Rhodes

Biologists generally agree that evolution via natural selection need not entail "progress", yet it clear that there have been increases in complexity throughout the evolution of life on earth. New types of structure arose during its major transitions from minimal self-replicators to prokaryotes and via symbiogenesis to eukaryotic cells to colonies of clones to multicellular life and higher organizational grades. This complexity increase has evidently been accompanied by a corresponding increase in the number of levels of hierarchy in biological systems. Hierarchical structuring of previously independent units entails chains of dependencies, communication and feedback in processes within an organism. Measuring and explaining this observed complexity of biological systems has remained an open problem.

A very rough index of complexity is the number of base-pairs of DNA [or alternatively, coding DNA]. J. Tyler Bonner has suggested that the number of cell types in an organism provides a better measure of its complexity. Another intuition is to count the number of levels in the (ill-understood) hierarchy going from say the gene to cell to tissues to organs to organism and possibly beyond, or, alternatively, to define complexity as the maximal length of chains of dependency within the organism. The latter idea admits a natural axiomatization from the viewpoint of Krohn-Rhodes algebraic theory of automata, if one models organisms as finite-state automata. Indeed, then there is a rigorous definition of complexity that corresponds to the intuitive idea of number of levels of hierarchy or longest essential chain of dependence.

The Krohn-Rhodes theorem states that any finite automaton can be emulated by a hierarchical cascade of basic component automata (permutation automata and combinatorial automata). Roughly speaking, the least number of cascaded levels over all such possible ways of building a given automaton is its complexity. Thus, finite automata form a [properly increasing, infinite] complexity hierarchy: the automata at level n are those requiring n or fewer levels to construct them. This notion of complexity as the number of hierarchically combined levels of basic components leads to interesting complexity hierarchies in various areas of mathematics (group theory, semigroup theory, language theory, linear algebra, algebraic topology and other areas) in addition to finite automata theory.

Let each organism in a species or population be modelled as a finite-state automaton, describing its states and their transitions upon stimulus. (This is a much more elaborate representation than the gene and genotype frequencies used by evolutionary geneticists.) Hence each organism (or unit of selection) has an associated complexity. Suppose that between generations, changes are absent or as small as possible. (There are no "hopeful monsters".) That is, we assume in a single generation (or speciation

Reconstruction of ancient molecular phylogeny

Ilya Muchnik, Roderic Guigo, and Temple F. Smith

Under the assumption that differences among genes trees can be explained by gene duplications and consequent losses, we have developed a method to obtain the global phylogeny minimizing the total number of postulated duplications and losses. The method is formulated within a graph theory representation of the biological concepts of gene duplication and loss. We have used the method to infer the most likely phylogenetic relationship among 16 major higher eukaryotic taxa from the sequences of 53 different genes. Only five independent genome duplication events need to be postulated in order to explain the inconsistences among these trees.

Address: Ilya Muchnik
RUTCOR, Rutgers University
P.O. Box 5062
New Brunswick, NJ 08903-5062
e-mail: muchnik@rutcor.rutgers.edu

Linear embedding of finite binary hierarchies

Boris Mirkin

The discrete binary hierarchy (DBH) is a nested set of subsets (clusters) of a finite n-element set, such that any nonsingleton cluster is split in exactly two smaller clusters. It appears that any DBH one-to-one corresponds to an orthonormal basis of the n-dimensional Euclidean space. The properties of these bases (and of the groups of linear transformations between the bases) may be a mathematical framework to be developed and applied to a bunch of problems: clustering, multiresolution image/signal processing, ordering, and indexing/quering in large data bases.

Clustering applications are based on a DBH-based analogue of the singular-value-decomposition of data matrices; they lead both to some methods and new insights in some hard practical problems in clustering (as standardization of the data or interpretation of the results). The approach can be employed in analyzing the multialignment problems for nucleotide sequences. Multiresolution image processing applications are based on a particular kind of "homogeneous" DBH (to be referred to as wavelet hierarchies) which allow the development of a rather complete discrete analogue to a very recent topic in many applications, so called "wavelet and multiresolution theory". Due to its discrete nature, the DBH multiresolution theory is expected to imply more effective methods for image/signal processing due to relaxing the "dimensionality/continuity" restriction of the classical theory. Ordering applications are based on analysis of particular kinds of the hierarchies related to orders or weak orders (caterpillar hierarchies). Data-base applications are connected with a preliminarily designed DBH and DBH-derived weights of importance of the data base descriptors.

Address: Boris Mirkin
DIMACS
Rutgers University
P.O.Box 1179
Piscataway NJ 08855-1179
e-mail: mirkin@dimacs.rutgers.edu

Learning algorithms generating multigranular hierarchies

Alex Meystel

Address: Alex Meystel
NIST
Bldg. 220, Rm. B124
Gaithersburg, MD 20899-0001
e-mail: meystel@cme.nist.gov

Address: Iven Van Mechelen
Department of Psychology
University of Leuven
Tiensestraat 102
B-3000 Leuven Belgium
e-mail: Iven.VanMechelen@psy.kuleuven.ac.be

A family of hierarchical decomposition models for binary two–way two–mode data

Iven Van Mechelen and Seymour Rosenberg

A family of models for binary two–way two–mode data, that is, data that take the form of a binary $m \times n$ matrix D is proposed. Examples of such data include person by intelligence task success/failure data in psychology and species by site presence/absence data in biology.

In the data analysis, the $m \times n$ matrix D is approximated, in the least squares sense, by an $m \times n$ matrix M, M being the model matrix of a hierarchical decomposition model. The decomposition of M is defined in terms of a decomposition function f of an $m \times r$ component matrix S and an $n \times r$ component matrix P, r being the rank of the decomposition model. The component matrices S and P are assumed to take values in an ordered set with k elements, $k \geq 2$. In substantive applications S and P may be interpreted in terms of latent characteristics (e.g., requisites, needs, strategies) underlying the data.

A natural hierarchy can be defined on the row elements of M in terms of the quasi–order induced by the inclusion relation on their sets of associated column elements in M; similarly a natural hierarchy can be defined on the column elements of M in terms of the inclusion relation on their sets of associated row elements. A decomposition of M into component matrices S and P is said to be hierarchical if the natural hierarchies of the row and column elements of M are either directly or inversely represented by the natural hierarchies that can be defined on the row elements of S and P.

A special case of hierarchical decomposition models are the decomposition models with component matrices that take binary values described by De Boeck and Rosenberg (1988) and by Van Mechelen, De Boeck, and Rosenberg (1995) under the name hierarchical classes models. Those models can be given a comprehensive graphical representation, which is illustrated with a hypothetical example.

A subclass of hierarchical decomposition models based on so–called homogenous binary decomposition rules is defined. It is shown that for each value of k two and only two types of models (each of which having four variants) exist. This further implies that, under the homogeneity restriction, the set of binary hierarchical decomposition models described by De Boeck, Rosenberg and Van Mechelen is exhaustive.

Keywords: clustering, partial order, Boolean matrix decomposition, hierarchical classes models

Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees

Vladimir Makarenkov and Bruno Leclerc

The circular orders associated with planar drawings of an X-tree (or phylogenetic tree) have been studied by several authors, especially for the encoding of an X-tree by $2n$-3 lengths of paths between leaves (n being the cardinality of X). It is shown here that circular orders are the same as those obtained from the table of a tree metric by a construction due to Yushmanov (1984). More, this construction applies to any dissimilarity, tree metric or not. Several fast algorithms (of complexity $O(n^2)$) are derived from these results: for the determination of a Yushmanov order; for the reconstruction of the valued X-tree representing a tree metric; for the recognition of a tree metric; and for the adjustment of a tree metric to a given metric; the last one proceeds by successive local least squares approximations. It leads to a fitting procedure which, tested on some examples, gives interesting results.

Address: Bruno Leclerc
CAMS-EHESS, 54 bd Raspail
F-75270 Paris cedex 06
FRANCE
e-mail: leclerc@ehess.fr

Internal validation of phylogenies reconstructed from distance matrices: An application of the jackknife

Francois-Joseph Lapointe

The jackknife strict consensus (Lanyon, 1985) has been used to detect topological incompatibilities in phylogenetic trees derived from distance data. I will here extend this approach to account for branch lengths, as well as topological relationships, when comparing jackknife pseudoreplicates. The average consensus procedure (Lapointe and Cucumel, in press) is used to derive a tree reflecting the agreement among the jackknife phylogenies. Combining the average tree with a minimum and maximum consensus provides information about jackknife tree variability; the range consensus between the minimum and the maximum consensus represents the topological agreement among them. I will demonstrate the effect of the number of pseudoreplicates on the resulting consensus trees by considering single and multiple deletions. The jackknife extensions that I propose will be applied to both hypothetical and DNA-hybridization distance data. I will show that consideration of branch lengths reveals greater stability of the jackknife phylogeny than would be demonstrated by a strict consensus based on topology alone. The new MAJACK computer package designed for matrix jackknifing will be introduced.

References

Lanyon, S. M. 1985. Detecting internal inconsistencies in distance data. Systematic Zoology, 34: 397-403.

Lapointe, F.-J., and G. Cucumel. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of leaves. Systematic Biology, in press.

Address: F.-J. Lapointe
Departement de sciences biologiques
Universite de Montreal
C.P. 6128, succursale centre-ville
Montreal, Quebec, H3C 3J7
Canada
email: lapoinf@ere.umontreal.ca

Estimation of missing cells in distance matrices prior to phylogenetic reconstruction purposes: The additive procedure outperforms the ultrametric one

Pierre-Alexandre Landry and Francois-Joseph Lapointe

The use of DNA hybridization or comparative serology in evolutionary studies often leads to lacunose (i.e. incomplete) distance matrices not directly amenable to phylogenetic reconstruction. In previous papers, De Soete (1984) and Lapointe & Kirsch (1995) have shown that missing cells can be estimated using the ultrametric property of distances, and that reliable trees can be derived from such filled matrices. In this work, we are introducing a novel way to estimate missing distances, based on the four-point condition of additive matrices. This condition is a generalization of the ultrametric property, and therefore was suspected to provide better estimations of distances that were not meeting the ultrametric inequality.

A simulation study, based on DNA hybridization data, was designed to assess whether the additive procedure is superior to the ultrametric one in recovering distances randomly deleted from complete distance matrices. Our results clearly indicate that the topologies and branch lengths of the trees derived from matrices which were estimated additively are better recovered than for matrices estimated ultrametrically; the original distances are also better recovered with the additive procedure. Except in the case of small matrices with many missing cells for which both methods perform equally, the additive is generally superior to the ultrametric method for estimating missing cells in distance matrices prior to phylogenetic reconstruction.

References

DE SOETE, G. 1984. Additive-tree representations of incomplete dissimilarity data. Quality and Quantity 18:387-393.

LAPOINTE, F.-J. and J. A. W. KIRSCH. 1995. Estimating phylogenies from lacunose distance matrices, with special reference to DNA hybridization data. Mol. Biol. Evol. 12:266-284.

Address: Pierre-Alexandre Landry
Dept. des Sciences Biologiques
Universite de Montreal
C.P. 6128, succursale centre ville
Montreal, Quebec
H3C 3J7 CANADA
e-mail: landryp@ERE.UMontreal.CA

Agreement metric for trees revisited

Ewa Kubicka, Grzegorz Kubicki, and F. R. McMorris

Given two trees T_1 and T_2, an agreement subtree is a tree that can be obtained from T_1, as well as from T_2, by pruning the fewest number of leaves from the two trees. In earlier work we investigated two metrics on trees that are based on the concept of agreement trees. In this paper we study the geodesic properties of these metrics, introduce a new related metric, and compare it with the previous ones.

Keywords: agreement subtree, metric

Address: F. R. McMorris
Department of Mathematics
University of Louisville
Louisville, KY 40292
e-mail: frmcmo01@homer.louiville.edu
Grzegorz Kubicki
Department of Mathematics
University of Louisville
Louisville, KY 40292
e-mail: gmkubi01@homer.louiville.edu

where $A_{k-1} \in \Omega_{k-1}$, $A_k \in \Omega_k$, and the maximization is over all $A_{k-1} \subset A_k$; the increment in merit, $M(A_{k-1}, A_k)$, when placing the class $A_k - A_{k-1}$ at the k^{th} order position obtained from one of the set functions given above: $F(\cdot), I(\cdot), J(\cdot)$, or $L(\cdot)$. Beginning with the direct computation of $\mathcal{F}(A_1)$ for $A_1 \in \Omega_1$, and proceeding with the recursive generation of $\mathcal{F}(A_k)$ for $k \geq 2$, an optimal ordered partition is finally identified by $\mathcal{F}(A_M)$ for $S = A_M$; an ordered partition attaining this value can be constructed by working backwards through the recursive steps. In addition, optimal ordered partitions into 2 through $M-1$ classes are identified by $\mathcal{F}(A_k)$ for $A_k = S \in \Omega_k, 2 \leq k \leq M-1$.

Several data analysis examples are given based on a computer implementation for the recursive optimization strategy just summarized. We also discuss how object sets larger than those that might be feasible using the current computer implementation (i.e., when object set sizes are greater than, say, 20) might still be analyzed with heuristic adaptations of the basic recursive strategy (e.g., through restrictions to subsets of a larger object set, and to the incorporation of object classes as the basic entities through which an optimal ordered partition is to be constructed). Finally, several connections are noted between our approach for obtaining optimal ordered partitions and the construction of medians within a collection of relations defined between the objects in S.

Keywords: classification, combinatorial optimization, dynamic programming, ordered partitions.

Address: Lawrence Hubert
Department of Psychology
The University of Illinois
603 East Daniel Street
Champaign, Illinois 61820
e-mail: lhubert@uiuc.edu

and for $2 \leq k \leq M-1$,

$$I(S_k) = (1/n_k)\left(\sum_{O_{i'} \in S_1 \cup \cdots \cup S_{k-1}, O_{k'} \in S_k} p_{k'i'} - \sum_{O_{j'} \in S_{k+1} \cup \cdots \cup S_M, O_{k'} \in S_k} p_{k'j'}\right).$$

A maximization of this index solves a least-squares problem of finding a collection of optimal coordinates, $x_1^* \leq \ldots \leq x_M^*$ (where $\sum_{k=1}^M n_k x_k^* = 0$), that minimizes

$$\sum_{k \leq k'} \sum_{O_{i_k} \in S_k^*, O_{j_{k'}} \in S_{k'}^*} (p_{i_k j_{k'}} - \mid x_{k'}^* - x_k^* \mid)^2,$$

where $x_k^* = (1/n) I(S_k^*), 1 \leq k \leq M$, and $S_1^* \prec \ldots \prec S_M^*$ defines the optimal ordered partition.

For a skew-symmetric matrix $\mathbf{T}$ we consider the two indices given in (c) and (d) below:

(c) $\sum_{k=1}^M J(S_k)$, where $J(S_1) \equiv 0$, and for $2 \leq k \leq M$,

$$J(S_k) = \sum_{O_{i'} \in S_1 \cup \cdots \cup S_{k-1}, O_{k'} \in S_k} t_{i'k'}.$$

This merit index is merely the sum of proximities in $\mathbf{T}$ from objects in S_k to $S_{k'}$ for all $k < k'$.

(d) $\sum_{k=1}^M n_k [L(S_k)]^2$, where for $1 \leq k \leq M$,

$$L(S_k) = (1/n_k) \sum_{O_{k'} \in S_k} \sum_{O_{i'} \in S - S_k} t_{k'i'}.$$

The maximization of this latter index of merit would produce an optimal unordered partition, $S_1, \ldots, S_M$, and the least-squares task of finding a collection of optimal coordinates $x_1, \ldots, x_M$ (not necessarily ordered) to minimize

$$\sum_{k,k'} \sum_{O_{i_k} \in S_k, O_{j_{k'}} \in S_{k'}} (t_{i_k j_{k'}} - (x_k - x_{k'}))^2,$$

can be solved by letting $x_k = (1/n) L(S_k)$. Thus, a final ordering of the classes according to their increasing coordinate values identifies an optimal ordered partition $S_1^* \prec \ldots \prec S_M^*$.

Given $\mathbf{P}$ or $\mathbf{T}$ and one of the chosen measures of merit, the optimization task of locating an optimal ordered partition that maximizes merit can proceed through a recursive (dynamic programming) strategy. Explicitly, define a collection of M sets of entities, $\Omega_1, \ldots, \Omega_M$, where Ω_k contains *all* $2^n - 1$ nonempty subsets of the n object subscripts; let $\mathcal{F}(A_k)$ for $A_k \in \Omega_k$ be the optimal value for placing k classes at the first k order positions where A_k is the union of these k classes, and define recursively $\mathcal{F}(A_k) = \max[\mathcal{F}(A_{k-1}) + M(A_{k-1}, A_k)]$,

The construction of globally optimal ordered partitions

Lawrence Hubert, Phipps Arabie, and Jacqueline Meulman

The classification task considered is the construction of an (optimal) ordered partition for a set of n objects, $S = \{O_1, \ldots, O_n\}$, defined by a collection of M mutually exclusive and exhaustive subsets of S, denoted $S_1, S_2, \ldots, S_M$, for which an order is imposed on the placement of the classes, $S_1 \prec S_2 \prec \ldots \prec S_M$, but not on their constituent objects. The data available to guide this search are assumed to be in the form of either (i) an $n \times n$ symmetric proximity matrix $\mathbf{P} = \{\mathbf{p_{ij}}\}$, where p_{ij} $(= p_{ji} \geq 0)$ is a dissimilarity for the objects O_i and O_j in which larger values indicate more dissimilar objects, or (ii) an $n \times n$ skew-symmetric proximity matrix $\mathbf{T} = \{\mathbf{t_{ij}}\}$, where t_{ij} $(= -t_{ji}$, and $t_{ii} = 0)$ characterizes a dominance relation for the objects O_i and O_j (i.e., when $t_{ij} > 0$, O_i dominates O_j with the degree of dominance given by t_{ij}). The identification of an optimal ordered partition for S will be guided by the maximization of some index of merit that measures how well a given ordered partition reflects the data in $\mathbf{P}$ or $\mathbf{T}$.

When data are in the form of either a symmetric matrix $\mathbf{P}$ or a skew-symmetric matrix $\mathbf{T}$, there are two general types of indices of merit that can be used to characterize the adequacy of a specific ordered partition in reflecting the pattern of proximities in $\mathbf{P}$ or $\mathbf{T}$. One is defined directly by how well an ordering in the sizes of proximities is mirrored by the ordering of the M classes, and the second is obtained indirectly through a set of (optimally) estimated coordinates for the M classes and how differences in the latter can reconstruct the proximities in a least-squares sense.

Explicitly, for a symmetric matrix $\mathbf{P}$ we consider the two indices in (a) and (b) below:

(a) $\sum_{k=1}^{M} F(S_k)$, where $F(S_1) = F(S_M) \equiv 0$, and for $2 \leq k \leq M-1, F(S_k) =$

$$\sum_{O_{i'} \in S_1 \cup \cdots \cup S_{k-1}, O_{k'} \in S_k, O_{j'} \in S_{k+1} \cup \cdots \cup S_M} \{f(p_{i'k'}, p_{i'j'}) + f(p_{k'j'}, p_{i'j'})\},$$

for $f(\cdot, \cdot)$ defining a function of how the order relation present for its two arguments counts in the total merit measure. Letting x and y denote two proximities, we explicitly consider two specifications for $f(x, y)$: $f(x, y) = \mid y - x \mid sign(y - x)$ and $f(x, y) = sign(y - x)$, where $sign(y - x) = +1$ if $y - x > 0$; $= 0$ if $y - x = 0$; $= -1$ if $y - x < 0$. The use of these are referred to respectively as weighted and unweighted gradient measures of merit for the ordered partition, $S_1 \prec \ldots \prec S_M$.

(b) $\sum_{k=1}^{M} n_k [I(S_k)]^2$, where

$$I(S_1) = -(1/n_1) \sum_{O_{k'} \in S_1, O_{j'} \in S - S_1} p_{k'j'}, \quad I(S_M) = (1/n_M) \sum_{O_{k'} \in S_M, O_{i'} \in S - S_M} p_{k'i'},$$

Address: Jody Hey
Dept. Biological Sciences
Nelson Biological Laboratories
Rutgers University
P.O. Box 1059
Piscataway, NJ 08855-1059
e-mail: hey@mbcl.rutgers.edu

The fortunate failure of hierarchy in phylogenetic and genealogical patterns

Jody Hey

Bifurcating networks, or trees, are often used as models of ancestral relationships among biological entities. The ancestor-descendant relationships among species are phylogenetic relationships, and they are often modeled with trees (species tree). Similarly the relationships among DNA sequences are genealogical relationships, and they too are often modeled with trees (i.e. gene trees). Both applications have major failings. The use of a tree model for phylogeny overlooks the fact that divergence of taxa does not happen instantaneously. A related problem is that estimated gene trees are often used as an approximation of a species tree. However, even if a bifurcating model holds for a gene tree, there is no necessary correspondence between the topology of a gene tree and the topology of a species tree for the taxa from which the genes were drawn. This is the well known gene-tree-species-tree problem. However it is only a problem for students of phylogenetic pattern. For inquiries on the forces that generate taxonomic diversity, it is a bonanza. In general an absence of topological correspondence between species and gene tree provides insight on the biological causes of the conflicting patterns. Furthermore these causes are necessarily part of the process of taxonomic divergence (i.e. speciation) and so of great interest.

In general, a genealogical relationships are expected to fit a tree model better than are phylogenetic relationships. For example, true gene trees are necessarily bifurcating while this need not be the case for a true species tree. However, an entirely different failing of the bifurcating tree model arises for many sets of DNA sequences, and that is recombination. If a sample of DNAs has a history of recombination, then different portions of the DNA sequences will have different gene tree histories, and there will be no overall bifurcating history for the entire sequence. Recombination creates complex interconnected networks of historical relationships. However, this failing of hierarchy can also be a boon for research on phylogenetic processes. This is because of the properties of the variance of apparent phylogenetic relationship. If a gene tree model holds well for a set of DNAs, then that model can be used for the analysis, but the gene history may not be representative (just by chance) of the history of the organisms from which it came. If there is recombination, then gene tree models do not fit, but the variance is reduced. Other methods can be used to study history, and there is a better chance that the genealogical relationships are more representative of the species from which they came.

Nesticity

Pierre Hansen, Dominique de Werra and Michael Gerber

The talk is devoted to a study of the smallest number of families of nested sets into which a given set of sets (or hypergraph edge set) can be decomposed. We also consider partially defined hypergraphs.

Address: Pierre Hansen
GERAD
École des Hautes Études Commerciales
5255 Avenue Decelles
Montréal (Québec), H3T 1V6
CANADA
e-mail: pierreh@crt.umontreal.ca

Address: Alain Guénoche
LIM-CNRS, 163 Av. de Luminy
13288 Marseille Cedex 9
France
e-mail: guenoche@lim.univ-mrs.fr

Ordinal approach in tree reconstruction

Alain Guénoche

Let D be a dissimilarity on X and O_x be the weak order associated with $x \in X$ that corresponds to the increasing distance order from x. Two weak orders O_x and O_y disagree about $\{z,t\}$ when $D(x,z) < D(x,t)$ and $D(y,t) < D(y,z)$ and there is a semi-agreement when $D(x,z) = D(x,t)$ but $D(y,z) \neq D(y,t)$. The distance $\Delta(O_x, O_y)$ is equal to the number of pairs $\{z,t\}$ for which these weak orders disagree, plus half the number of semi-agreements. Δ is the symmetric difference distance between weak order relations. The order distance associated with D is defined as $D_o(x,y) = \Delta(O_x, O_y)$. The aim of this paper is to explain how to use order distances in clustering, because:

Theorem [Bonnot, Guénoche & Perrier 1995] If D is a tree metric that can be represented by A as a set of edges, its order distance D_o is also a tree metric that can be represented by some $A' \subseteq A$.

In other words, the support tree of a tree metric D and its order distance have compatible topologies. Some edges may disappear, but if D is an ultrametric distance, they are all retained.

Consequently, one can use order distance to give more stability to tree reconstruction methods starting from a given dissimilarity D. Because if we slightly modify the value $D(x,y)$, we do not necessarily alter weak orders O_x and O_y. One can think that if D is close to a tree distance, a tree topology computed from D_o could be more accurate than the one computed from D.

We also introduce a method to compute a tree topology starting from a preordonnance (that is a weak order relation on X pairs). In that case, one can define quadruples of X-elements that are representable on a tree (and those that are not), and derive to an ordinal scoring methods.

References

J.P. Barthélemy, A. Guénoche (1991) Trees and proximity representation, J. Wiley & Sons, New York, 238 p.

A. Guénoche (1994) Order distance associated with a hierarchy, Conference "Combinatorics in Behavioral Sciences", Irvine, August 1994, Journal of Classification, to appear.

F. Bonnot, A. Guénoche, X. Perrier (1995) Properties of an order distance associated to a tree distance, Proceedings of "Ordinal and Symbolic Data Analysis", E. Diday et al. (Eds.), Springer Verlag, 1996, pp. 252-261.

we propose to choose among these formulae the one which induces the minimum variance. To compute this variance, we use a simplified Poisson model of sampling noise, which is well suited to deal with biological sequences. Within this model, the variances and co-variances of distances between species or between clusters of species may be represented using a valued tree having the same topology as the true tree. At each step, the algorithm not only reduce the distance matrix using the minimum variance reduction, but also the variance-covariance matrix. Its complexity remains the same and it still has the property of necessarily finding the correct tree with additive data. But, its ability in recovering the correct topology is very significantly improved, up to about 20% when the true tree has very heterogeneous branch lengths (e.g., tree CC in Kumar 1996). When the true tree has homogeneous branch lengths (e.g., tree BB in Kumar 1996), this new algorithm is theoretically and practically equivalent to NJ.

Keywords: Tree distance, agglomerative algorithm, biological sequences

Address: Olivier Gascuel
Departement d'Informatique Fondamentale
LIRMM
161 rue Ada
34392 Montpellier France
e-mail: gascuel@lirmm.fr

currently
GERAD, Ecole des HEC,
5255 Av. Decelles
Montreal H3T 1V6 Canada
e-mail: olivierg@crt.umontreal.ca

Improving the NJ algorithm

Olivier Gascuel

The neighbor-joining (NJ) algorithm of Saitou and Nei (1987) is one of the most popular methods to reconstruct phylogenetic trees from a matrix of pairwise distances. This algorithm follows an agglomerative scheme, firstly proposed in the context of mathematical psychology by Sattath and Tversky (1977). Agglomerative algorithms iteratively pick a pair of taxa, create a new node which represents the cluster of these taxa, and compute a new distance matrix of reduced size where both taxa are replaced by this node. The process is repeated until the number of taxa is equal to three. To agglomerate pairs of nodes, NJ follows the minimum evolution (ME) principle which consists in choosing the tree with the smallest sum of branch length. Rzhetsky and Nei (1993) have shown that this principle is well founded when the lengths are obtained by the ordinary least-squares method, and when an unbiased estimate of evolutionary distances is used. However, due to its greedy, agglomerative approach, NJ does not usually find the ME tree, but only a "short" tree which is generally similar to the ME tree (Saitou and Imanishi 1989). This does not preclude for good performance since the true tree itself is usually close but not identical to the ME tree, and numerous computer simulations (Saitou and Nei 1987; Nei 1991; Kuhner and Felsenstein 1994) have shown a high relative efficiency of the NJ method in recovering the true topology. Moreover, the NJ algorithm, as formulated by Studier and Keppler (1988), is efficient from a computational point of view and has an $O(n^3)$ time complexity, where n is the number of taxa.

Several attempts have been done to improve the NJ algorithm and to design methods able to find trees very close or identical to the ME tree. Saitou and Imanishi (1989) proposed an exhaustive search method which applies when the number of taxa is small ($n < 10$). Rzhetsky and Nei (1993) designed various strategies of searching for the ME tree in the neighborhood of the NJ tree by conducting local rearrangements. These authors also suggested to generate alternative topologies by using a bootstrap procedure (Rzhetsky and Nei 1994). Finally, Kumar (1996) designed efficient heuristics for searching the tree space in a more or less exhaustive manner. All these methods most often find trees which are shorter than the NJ tree but, unfortunately, computer simulations (Saitou and Imanishi 1989; Kumar 1996) indicate that the ability in recovering the true topology is not increased, and that NJ may hardly be outstripped.

We propose another line of approach. Instead of trying to find trees shorter than NJ trees, we reconsider the mere principle of the NJ algorithm. We show that the formula used by NJ to reduce the distance matrix at each step, may be generalized to a parametric class of admissible formulae which conserve the property to necessarily find the correct tree when data are additive (see also Bandelt and Dress, 1986). Then,

Evolution and additive metrics

Martin Farach

Address: Martin Farach
DIMACS Rutgers
P.O.Box 1179
Piscataway NJ 08855-1179
e-mail: farach@cs.rutgers.edu

On the equivalence of two tree mapping measures

Oliver Eulenstein

Tree-mapping is a method of deducing the evolutionary history of a set of genes, given two seemingly contradictory phylogenetic trees for the genes and for the species in which the genes were found. The method is based on the postulation of duplication events which then allow to resolve the contradiction between the phylogenetic trees. To quantify the difficulty of this reconciliation two biologically meaningful measures have been introduced in the literature. Mirkin, Muchnik, and Smith conjecture that the two measures are in fact equivalent. The talk will introduce the problem of comparing trees under the assumption of duplication events and motivate the measures in use. A short overview of the proof of the Mirkin-Muchnik-Smith conjecture will be given.

Address: Oliver Eulenstein
University of Bonn,
Institute for Computer Science - FGL,
Roemerstrasse 164,
D-53117 Bonn,
Germany
e-mail: Oliver.Eulenstein@gmd.de

Ancestor-descendant relations and incompatible data: motivation for research in discrete math

George Estabrook

The species in a collection of biological species may have evolved in such a way that each species has a unique species as its immediate ancestor and that the entire collection has a single most recent common ancestor. Some or all of the species ancestral to those in the collection may be missing from the collection. The concept "immediate ancestor" can be construed as a relation on the collection of species extended to include any missing ancestors (the AD relation).

Comparative biological data for such a collection are partitions of the collection into classes considered equivalent with respect to some basis for comparison. A task of interest to comparative biologists is to reconstruct the unknown AD relation from comparative data. In the mid 1960's some well known biologists, including Wilson, Camin, Sokal, and Hennig began to realize that often comparative data were somehow contradictory, so that this reconstruction was confounded.

In 1968, Estabrook formulated this problem in discrete math terms. A partition accurately reflects the AD relation if its classes are all convex on the haase diagram of the AD relation. Two partitions are contradictory if there does not exists any AD-like relation on whose haase diagram all classes of both are convex. Approaches to resolving these contradictions have motivated related research in discrete math since that time. This talk reviews progress and present challenges.

Address: George Estabrook
Department of Biology
University of Michigan
Ann Arbor, MI 48109
e-mail: gfred@umich.edu

Weak hierarchies

Andreas Dress

Weak hierarchies have been introduced a few years ago as a particularly efficient and comparatively stable way to associate a useful family of ¡¡global¿¿ clusters to given ¡¡local¿¿ similarity data. To achieve usefulness, efficiency and stability simultaneously, overlapping clusters were accepted up to a certain, exactly specified degree. Since then, many new insights into the combinatorics governing overlapping clustering procedures and the various forms of admissible ¡¡local¿¿ input data have been obtained.

In the lecture, an introduction into the theory of weak hierarchies will be given, complemented by a discusion of various recent combinatorial results relating to this subject.

Address: Andreas Dress
Uni Bielefeld / FSPM - Strukturbildungsprozesse
Postfach 10 01 31
33501 Bielefeld
Germany
e-mail: dress@Mathematik.Uni-Bielefeld.DE

Multiple trees: Fitting two or more tree structures to dissimilarity data.

J. Douglas Carroll

Tree structures are often used in psychology to represent certain types of conceptual or cognitive stimuli, based on proximity data deriving from, e.g., subjective judgments (say on a ten point scale) of dissimilarity of all pairs of stimuli, or behavioral measures of similarity (e.g., probability of confusion on a "same-different" or stimulus recognition task). A hierarchical tree structure can be characterized as describing the stimuli in terms of a nested set of discrete features, which in turn can be viewed as binary valued (all-or-none) dimensions or attributes.

For certain classes of stimuli a single hierarchical tree may not suffice to describe the conceptual structure of the stimuli. For example, a farmer who also has biological training may utilize two different tree structures—one based on the phylogenetic relationships of the animals, and a second based on their functional aspects in a farm setting. In some cases three or more hierarchical trees may underlie the conceptual structure of a class of stimuli (for at least some judges).

This reasoning leads to a representation of certain proximity data (assumed henceforth to be dissimilarities) in terms of *multiple* tree structures. In the case of hierarchically organized trees, the dissimilarities can be modeled as a sum of two or more ultrametrics, each associated with a separate hierarchical tree.

An alternative tree metric that has been used to model psychological data is the so-called "four-point," "additive" or "path length" metric, in which distance between two terminal nodes is the length of the unique path connecting these two nodes in the tree. This leads to unrooted "free trees," which are not necessarily hierarchically organized. Another form of multiple tree representation involves modeling dissimilarities via a sum of two or more path length or additive metrics associated with such unrooted nonhierarchical trees.

Some approaches to fitting multiple trees to dissimilarity data, and an application to some kinship data of Rosenberg and Kim, derived from a "similarity sorting" task, will be described.

Address: J. Douglas Carroll
Rutgers University, Faculty of Management,
81 New Street, Newark, NJ 07102-1895, USA
e-mail: dcarroll@rci.rutgers.edu

CFT trees. Our main result is $\Omega(\sqrt{\frac{n}{k}})$ lower bound on the expected distance between the original tree and a tree reconstructed by any algorithm.

Our result implies that, when k increases, the $O(\frac{1}{\sqrt{k}})$ ratio of convergence to the true tree shown by the algorithm of Farach and Kannan [1] is optimal. It also improves the previously known lower bound with respect to n.

If the topology of the tree is known and we need only to find out the probabilities corresponding to the edges, we have an algorithm which finds a tree within $O(\sqrt{n}\ln n/\sqrt{k})$ distance from the true tree. In this case, our lower bound is tight up to logarithmic factor.

References

[1] M. Farach and S. Kannan, *Efficient algorithms for inverting evolution.* Proceedings of the 28th ACM Symposium on the Theory of Computing, pp. 230-236, ACM, 1996.

Keywords: phylogeny, maximum likelihood methods

Address: Andris Ambainis
Institute of Mathematics and Computer Science
University of Latvia
Raina bulv. 29, Riga, Latvia
e-mail: ambainis@cclu.lv

On Maximum-likelihood reconstruction of evolutionary trees

Andris Ambainis

One of the problems considered in computational biology is the reconstruction of an evolutionary tree (phylogeny) for a set of species.

Maximum Likelihood Estimation (MLE) methods consider the evolution as a random process and search for the most likely process, given the data.

Maximum likelihood methods start with mapping the data about species (for example, DNA sequence information) to a finite set of properties called *characters*. Each of characters has finitely many possible values called *states*. Then, each species is described as a vector of states (or, equivalently, a function from the set of characters to the set of states).

In this paper, we assume that characters have two states (0 and 1). Further, k denotes the number of characters and n denotes the number of species.

As the description of the evolutionary processes, we use Cavender-Farris trees (CFT). A Cavender-Farris tree is a rooted tree with n leaves where the root has probability $P_r \in [0,1]$ each edge has probability $P_e \in (0, 0.5)$. It generates vectors from $\{0,1\}^n$ as follows. The root gets labeled 0 with probability P_r and 1 otherwise. Further, each labelled node broadcasts the bit with which it is labelled to all its children. If a bit b is broadcast down an edge $e = (u, v)$, then v gets labelled $1 - b$ with probability P_e and b with probability $1 - P_e$. The vector formed by the values of bits at the leaves is the output of the tree.

As input data, we have k vectors $x_1, \ldots, x_k \in \{0,1\}^n$ that describe k species. Our task is to find the CFT tree which has generated these vectors.

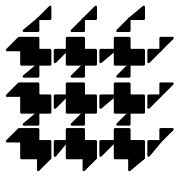
The amount of available data (species and characters which describe species) is finite. If we reconstruct the tree from finite amount of data, it is likely that our result will be only approximately correct. We define the distance between two trees as the distance between probability distributions generated by these trees and look for algorithms which minimize the distance between the original tree and the tree reconstructed by the algorithm.

Farach and Kannan [1] constructed an algorithm which produces an evolutionary tree within $O(n\sqrt{\ln n}/\alpha\sqrt{k})$ distance from the tree that has generated the analysed set of species, with high probability. They also gave a proof that no algorithm can construct a CFT that is within $o(1/k)$ distance from the original tree.

In this paper, we improve the lower bound on the algorithms for reconstruction of

DIMACS

Center for Discrete Mathematics & Theoretical Computer Science
A National Science Foundation Science and Technology Center



Workshop on Mathematical Hierarchies and Biology

November 13 - 15, 1996

ABSTRACTS

For information on visiting, email to `fellows@dimacs.rutgers.edu`

DIMACS Center • Rutgers University • P.O. Box 1179 • Piscataway, NJ 08855-1179 • USA
TEL: 908-445-5928 • FAX: 908-445-5932 • EMAIL: `center@dimacs.rutgers.edu`
WWW: `http://dimacs.rutgers.edu/` • TELNET: `telnet info.rutgers.edu 90`

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs, Bellcore, and Bell Laboratories