Apostolico *et al. Algorithms for Molecular Biology* 2013, **8**:7
http://www.almob.org/content/8/1/7

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Ultrametric networks: a new tool for phylogenetic analysis

Alberto Apostolico[1], Matteo Comin[2*], Andres Dress[3] and Laxmi Parida[4]

Abstract

Background: The large majority of optimization problems related to the inference of distance-based trees used in phylogenetic analysis and classification is known to be intractable. One noted exception is found within the realm of ultrametric distances. The introduction of ultrametric trees in phylogeny was inspired by a model of evolution driven by the postulate of a molecular clock, now dismissed, whereby phylogeny could be represented by a weighted tree in which the sum of the weights of the edges separating any given leaf from the root is the same for all leaves. Both, molecular clocks and rooted ultrametric trees, fell out of fashion as credible representations of evolutionary change. At the same time, ultrametric dendrograms have shown good potential for purposes of classification in so far as they have proven to provide good approximations for additive trees. Most of these approximations are still intractable, but the problem of finding the nearest ultrametric distance matrix to a given distance matrix with respect to the L_∞ distance has been long known to be solvable in polynomial time, the solution being incarnated in any minimum spanning tree for the weighted graph subtending to the matrix.

Results: This paper expands this *subdominant ultrametric* perspective by studying ultrametric *networks*, consisting of the collection of *all* edges involved in some minimum spanning tree. It is shown that, for a graph with n vertices, the construction of such a network can be carried out by a simple algorithm in optimal time $O(n^2)$ which is faster by a factor of n than the direct adaptation of the classical $O(n^3)$ paradigm by Warshall for computing the transitive closure of a graph. This algorithm, called UltraNet, will be shown to be easily adapted to compute relaxed networks and to support the introduction of artificial points to reduce the maximum distance between vertices in a pair. Finally, a few experiments will be discussed to demonstrate the applicability of subdominant ultrametric networks.

Availability: http://www.dei.unipd.it/~ciompin/main/Ultranet/Ultranet.html

Keywords: Phylogenetic network, Ultrametric distance, STR data analysis

Background

As is well known, most optimization problems related to the inference of distance-based trees used in phylogenetic analysis and classification are intractable (see [1,2] for a pertinent discussion). One notable exception is found within the realm of ultrametric distances (cf. [3]). The introduction of such distances in phylogeny was inspired by a model of evolution, now largely abandoned, driven by the postulate of a molecular clock whereby the amount of phylogenetic change observable between any two extant species is directly related to the amount of time that elapsed since their last common ancestor roamed this planet, implying that phylogenetic distances could simply be represented by a weighted tree in which the sum of the weights of the edges separating any given leaf from the root is the same for all leaves.

Both molecular clocks and rooted ultrametric trees fell out of fashion as credible representations of evolutionary change. At the same time, a rooted dated tree is still the "object of desire" in taxonomy and Tree-of-Life research, and ultrametric dendrograms have shown good potential for purposes of classification in so far as they have proven to provide good approximations for additive trees. While finding the "best" such approximation is, in most cases, still intractable, the problem

*Correspondence: comin@dei.unipd.it
[2]Department of Information Engineering, University of Padova, Via Gradenigo 6/A, 35131 Padova, Italy
Full list of author information is available at the end of the article

BioMed Central

© 2013 Apostolico et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

of finding an ultrametric distance matrix that is closest to a given distance matrix with respect to the L_∞ distance has long been known to be solvable in polynomial time, its solution being incarnated in any minimum spanning tree for the weighted graph subtending to the matrix.

Applications of minimum spanning trees in connection with problems of population classification and genetics are as old as any other of their numerous applications. An application to taxonomic problems related to species interrelationship dates back to [4]. And as early as 1964, Edwards and Cavalli Sforza [5] used MSTs to approximate evolutionary trees reconstructed from gene frequencies in blood groups from fifteen contemporary human populations.

Most approximation problems arising in this context fall within the framework of the following

Closest Metric Problem: *Given a set $\mathcal{M}$ of metrics C defined on a set V, an $|V| \times |V|-$matrix M, and a distance function $D : (M', M'') \to \mathbb{R}_{\geq 0}$ defined on the set $\mathbb{R}^{|V|\times|V|}$ of all $|V| \times |V| -$ matrices, find a metric $C \in \mathcal{M}$ with minimum distance to M relative to D.*

The basic facts are summarized in Table 1.

Subdominant ultrametrics have been traditionally applied to many problems of physics and optimization theory [9-12]. More recently, implications of this theory in the analysis of financial markets, stock exchange, and evolutionary biology have attracted new interest in the topic.

Phylogenetic networks are increasingly featured in modeling of molecular evolution, as evidence of reticulate events such as hybridization, horizontal gene transfer and recombination becomes more prominent. Traditionally, the use of binary data and, in particular, the notion of splits gave rise to a number of alternative models. In the literature, several definitions of networks have been proposed to model parallel events. Popular examples are consensus networks [13], reticulate networks, recombination networks, median networks [14], Neighbour Nets [15], QNets [16] etc. In order to control the degree of connectivity of a network, each model optimizes an objective function; examples are Bayesian methods, maximum likelihood methods, and maximum parsimony [17,18], calculated that the number of equally parsimonious trees for a data set of just 56 haplotypes exceeded one billion.

Table 1 Basic facts for the closest metric problem

$C \setminus D_M$	L_1	L_2	L_∞
Additive	NP-Hard[+]	NP-Hard	NP-Hard[*]
Ultrametric	NP-Hard[+]	NP-Hard	P[§]

[+]no approximation is known;
[*]a 3-approximation exists, due to Agarwal et al. [1];
[§]due to Gower and Ross [3] (see also ([6], p.158), ([7], p.134), [2,8]).

This estimate was computed through resort of the notion of Minimum Spanning Network. In a different context, they proposed a counting procedure based on the Prim's algorithm that is analogous to the work presented in this paper.

Anther popular framework is the statistical parsimony analysis [19]. Hart and Sunday [20] found empirically that subnetworks, as implemented in the TCS program [21], coincided significantly with taxonomy names. The TCS program calculates the maximum number of mutational steps constituting a parsimonious connection between two haplotypes with the probability of 95%. Although Hart and Sunday's [20] results suggest that statistical parsimony analysis could be used in practice to differentiate species, this methodology is not mathematically well-founded.

In this paper, we extend the approach based on the construction of subdominant ultrametric trees by studying ultrametric *networks*, consisting of the collection of all edges involved in some minimum spanning tree. This can be viewed as a network of kinship between the extant sequences that embodies the *least-resistant paths* in terms of *bottlenecks*, where a bottleneck is simply the worst possible transition between two intermediate states. We show that, for a graph with n vertices, the construction of such a network can be carried out by a simple algorithm in optimal time $O(n^2)$, which is faster by a factor of n than the more straightforward $O(n^3)$ closure performed by the classical Floyd-Warshall paradigm. We show that our algorithm can easily be adapted to compute relaxed networks and to support the introduction of artificial points when it is desirable to reduce maximum distance between vertices. Finally, we discuss a few experiments demonstrating the applicability of this method.

The ultrametric network

We study the following, rather abstract, conceptual frame work: We start with a finite set V representing the sequences and an arbitrary weighting

$$W : \binom{V}{2} \to \mathbb{R}_{>0} : \{v, u\} \to W(v, u)$$

that associates a positive weight $W(v, u)$ to every 2-subset $\{v, u\} \in \binom{V}{2}$ of V that we imagine to be deduced, in one way or the other, from the given sequences, and to represent, for every $\{v, u\} \in \binom{V}{2}$, the observed *degree of dissimilarity* between u and v.

It is well known (cf. [3]) and easy to see (cf. [8] for a review) that there exists a unique largest *ultrametric* defined on V and denoted by, say, W^* that is *dominated* by W, i.e., the (necessarily unique and symmetric) largest map from $V \times V$ into $\mathbb{R}$ for which

(i) $W^*(v,v) = 0$ and $W^*(v,u) \leq \max(W^*(v,w), W^*(u,w))$ holds for all u, v, w in V, and
(ii) $W^*(v,u) \leq W(v,u)$ for all $u, v \in V$.

Actually, as the supremum

$$\sup \mathcal{D} : V \times V \to \mathbb{R} : (u,v) \mapsto \sup(D(u,v) : D \in \mathcal{D})$$

of any collection $\mathcal{D}$ of ultrametrics defined on V that is bounded from above, is an ultrametric, too, and W^* is just the supremum of the set, $\mathcal{D}(W) := \{D : D \text{is an ultrametric on} V \text{that is bounded from above by} W\}$, W^* must indeed be an ultrametric, called the *subdominant ultrametric* for W.

In this paper, we will study the *ultrametric network* $G(V|W)$ associated with W, i.e. the graph $G(V|W) := (V, E(V|W))$ with vertex set V and edge set $E(V|W) := \left\{\{u,v\} \in \binom{V}{2} : W(u,v) = W^*(u,v)\right\}$.

It is easy to see that $E(V|W)$ is actually the union of the edge sets of all minimum spanning trees with vertex set V relative to W, considered as a weighting of the complete graph $G(V) = (V, E(V))$ with vertex set V and edge set $E(V) := \binom{V}{2}$.

Indeed, continuing with the notations and assumptions introduced above, W^* can be constructed as follows: Put $|e| := W(u,v)$ for every $e = \{u,v\} \in E(V)$ and, given any path $P = v_0 v_1 ... v_k$ in $G(V)$, define the *support* $supp(P)$ of P by

$$supp(P) := \{\{v_{i-1}, v_i\} : i = 1, 2, \ldots, k\},$$

and the *bottleneck* $B(P) = B(P|W)$ of P (with respect to W) by

$$B(P) = B(v_0 v_1 ... v_k) := \max_{i=1...k} \{|e| : e \in supp(P)\}.$$

Then, given any two vertices $u, v \in V$, $W^*(u,v)$ coincides with the *least-resistance* bottleneck between v and u, i.e., we have

$$W^*(u,v) = \min_{\text{all paths } P \text{ in } G(V) \text{ from} u \text{ to } v} B(P).$$

Any path for which this minimum is attained represents a *minimum-bottleneck* path (for u and v). Clearly, $W^*(u,v)$ is the lowest weight possible for the highest weight in any path leading from u to v. It can be computed by a straightforward adaptation of the Floyd-Warshall algorithm in $O(|V|^3)$ time.

Remarkably, $E \subseteq E(V|W)$ holds for every subset E of $E(V)$ for which the graph (V,E) is connected and minimizes the sum $|E| := \sum_{e \in E} |e|$, that is, for the edge set of any *minimum spanning tree* $T = (X,E)$ for W. This follows from a result generally credited to [3], that we formalize as follows:

Theorem 1. *With V and W as above, the edge set of every minimum spanning tree for W is contained in $E(V|W)$ while, conversely, there exists, for any edge $e \in E(V|W)$, a minimum spanning tree for W whose edge set contains e. In particular, the network $G(V|W)$ is always connected.*

Proof. Indeed, given any such subset $E \subseteq E(V)$ and any edge $e = \{u,v\} \in E$, we may denote by $\Pi(e) = \Pi_E(e)$ the bi-partition of V given by the (vertex sets of the) two connected components of the graph $(V, E - \{e\})$, and by $A(w) = A_E(w)$, for any $w \in V$, the unique component $A(w) \in \Pi(e)$ with $w \in A(e)$. Clearly, we have $\Pi(e) = \{A(u), A(v)\}$ for every edge $e = \{u,v\} \in E$.

Now, assume that there exists some $e = \{u,v\} \in E$ with $e \notin E(V|W)$. Then, we could find some $P = v_0 v_1 ... v_{k-1} v_k$ from $v_0 := u$ to $v_k := v$ in $G(V)$ with $B(P) < |e|$. Furthermore, as $A(v_0) = A(u) \neq A(v) = A(v_k)$ must hold, there must be some $i \in \{1, \ldots, k\}$ with $A(v_{i-1}) \neq A(v_i)$, eg the smallest i in $\{1, \ldots, k\}$ with $A(u) \neq A(v_i)$. Consequently, exchanging the edge e with the edge $e_i := \{v_{i-1}, v_i\}$ in E would also give rise to a spanning tree for $G(V)$, and we would have $|E'| = |E| + |e_i| - |e| < |E|$ in view of $|e_i| \leq B(P) < |e|$, thus contradicting our choice of E. So, $E \subseteq E(V|W)$ must hold, as claimed.

To establish the converse, assume that $e = \{u,v\} \in E(V|W)$ is not contained in any minimum spanning tree. Then, given any such tree, let $P = v_0 v_1 ... v_{k-1} v_k$ denote the unique path from $v_0 := u$ to $v_k := v$ in that tree. Then, exchanging any edge e' in the support of P with the edge e will produce a spanning tree for $G(V)$ of larger weight, implying that $|e'| < |e|$ must hold for every such edge e' implying that also $B(P) < |e|$ must hold. This, however, would clearly contradict our assumption $e \in E(V|W)$. □

Optimal computation of the ultrametric network

Clearly, given V and W as above, the ultrametric network can be produced in time $O(|V|^3)$ by a straightforward adaptation of the Floyd-Warshall all-pairs shortest-path algorithm [22]. In view of Theorem 1, this network could be produced in time $O(|V|^3)$ also by first computing one MST by, say, Prim's algorithm, and then computing W^*

using the paths in this tree. We present here an algorithm to compute the entire network in time $O(|V|^2)$. This is optimal since any algorithm must produce $\Theta(|V|^2)$ values at the outset.

The main idea is that the computation can be cast within a control structure that is strongly reminiscent of Prim's MST - or Dijkstra's single-source shortest-path algorithm (refer to, e.g., [22]): starting with an arbitrary vertex r, a subset $\bar{V}$ of V is progressively expanded by annexing, at each step, the one vertex u in $V - \bar{V}$ that is connected to $\bar{V}$ by an edge (v, u) that minimizes cost. As is well known, in Prim's MST the cost to be minimized is the weight of the partial tree over the vertices in $\bar{V}$, whence the edge to be chosen is one of minimum weight. In Dijkstra's algorithm, the cost to be minimized is the sum of weights on the arcs connecting u to r, whence the edge to be chosen is the one minimizing this sum. Note that in both cases there can be more than one vertex that minimizes the cost, however they will all produce the same global minimum. One important point of our algorithm is that (see Theorem 1) choosing (v, u) as in Prim's MST computes the ultrametric distance not only between u and r but between u and any other vertex in $\bar{V}$. Moreover, it can be seen that the pairwise ultrametric distances between any pair of vertices in $\bar{V}$ are not affected by the introduction of u in this set. This last circumstance yields the speedup from $O(|V|^3)$ to $O(|V|^2)$.

The algorithm starts with the original weights $W(u, v)$, computes the ultrametric W^* and identifies the subset of edges that form the ultrametric network $\bar{E}$. In the following pseudo-code, $d(u, v)$ is initialized to ∞ and then used to store consecutively refined estimates of the value of $W^*(u, v)$, for any pair u and v of vertices. It will be seen that at the end $d(u, v) = W^*(u, v)$.

ULTRAMETRIC-NETWORK(G, W, r**)**
INITIALIZATION
for every pair (u, v) of vertices in V **do**
  $d(u, v) = \infty$ **[initial estimate of** W^* **]**
end for
$\bar{V} = \{r\}$ **[Invariant:** $\bar{V}$ **is the subset of** V **with pairwise bottlenecks already computed]**
$Q = \bar{V} - \{r\}$ **[Invariant:** Q **is the subset of** V **not yet processed]**
$\bar{E} = \emptyset$ **[Invariant:** $\bar{E}$ **contains the edges** $\{u, v\} \in \binom{\bar{V}}{2}$ **with** $d(u, v) = W(u, v)$**]**
for every $u \in Q$ **do**
  $prec[u] = r$; $key[u] = W(u, r)$ **[Invariant:** $key[u] = W(prec[u], u)$ **is the minimal weight of an edge** $\{v, u\}$ **connecting** u **to some vertex** v **in** $\bar{V}$**, and** $prec[u] = r$ **is that vertex in** $\bar{V}$ **which is contained in this edge]**
end for
BODY
while $Q \neq \emptyset$ **[there are still vertices disconnected from** $\bar{V}$ **] do**
  ANNEX THE "CLOSEST" VERTEX
  $u = \arg\min(key[u] : u \in Q)$
  $v = prec[u]$; $\bar{E} = \bar{E} \cup \{u, v\}$; $d(u, v) = W(u, v)$
  COMPUTE THE BOTTLENECK BETWEEN u **AND EACH** $v' \in \bar{V}$
  for every $v' \in \bar{V}$ **do**
    $d(v', u) = \max\left(d(v', v), W(v, u)\right)$
    if $d(v', u) = W(v', u)$ **[more edges from** u **to** $\bar{V}$ **of weight** $W^*(v', u)$ **?] then**
      $\bar{E} = \bar{E} \cup \{v', u\}$ [$W^*(v', u) = max\left(W^*(v', v), W(u, v)\right)$, hence $W^*(v', u) = W(v', u)$]
    end if
  end for
  $\bar{V} = \bar{V} \cup \{u\}$ **[Vertex** u **is added to** $\bar{V}$ **]**
  $Q = Q \setminus \{u\}$ **[Vertex** u **is subtracted from** Q**]**
  UPDATING ***key*** **VALUES OF VERTICES IN** Q
  for every $v' \in Q$ **do**
    if $key[v'] > W(u, v')$ **then**
      $key[v'] = W(u, v')$
      $prec[v'] = u$
    end if
  end for
end while

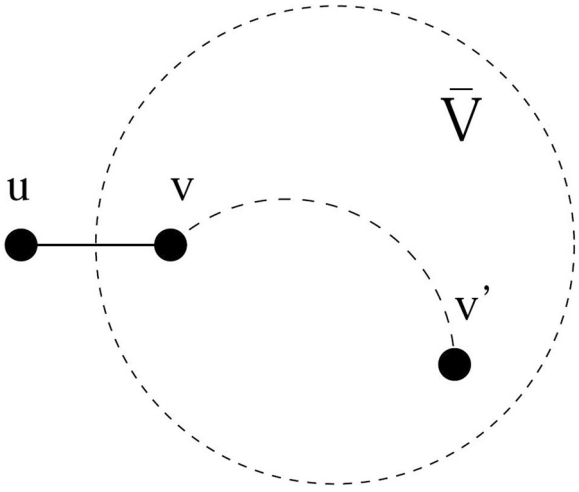

Figure 1 Illustrating the construction of the ultrametric network.

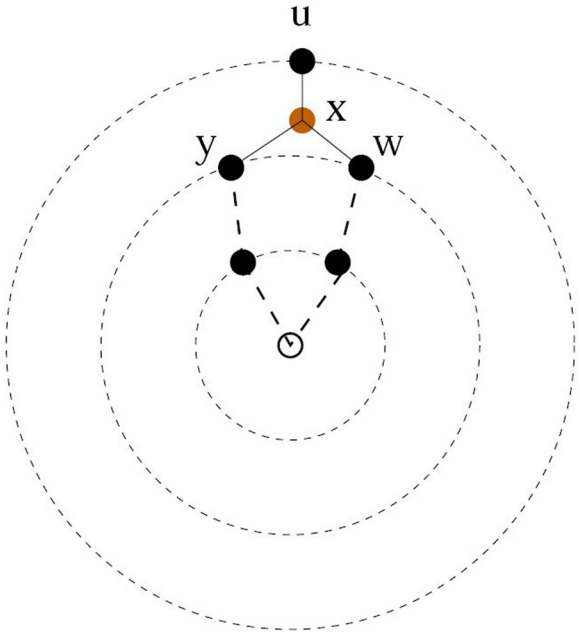

Figure 2 The creation of a new artificial vertex.

Lemma 1. *At each iteration, d coincides with* $(W|_{\bar{V}})^*$*, the subdominant ultrametric of the restriction* $W|_{\bar{V}}$ *of W to* $\bar{V}$*, where the edge* $\{v', v''\} \in \bar{E}$ *if and only if:*

- $d(v', v'') = W(v', v'')$ *and* $d(v', v'') \leq max(W(v', u), W(v'', u))$ *holds for any two vertices* v', v'' *in* $\bar{V}$ *and all* $u \in Q$*,*
- $key[u] = W(prec[u], u) = \min(W(u, v) : v \in \bar{V})$ *holds for every* $u \in Q = V - \bar{V}$*.*

The proof follows easily from observing that $\bar{V}$ contains, at each recursive step, a connected subgraph that is part of a MST for V. Figure 1 shows the generic step of the algorithm. It is easy to check that the set $\bar{E}$ contains all edges of the ultrametric network. We now prove that the algorithm is also optimal.

Lemma 2. *The algorithm constructs the subdominant ultrametric and its associated ultrametric network in optimal time* $O(|V|^2)$*.*

Proof. All the initialization steps, inclusive of the insertion of $(|V| - 1)$ initial values in the queue Q take time $O(|V|^2)$. Following this, each of the $(|V| - 1)$ iterations of the *while* loop contains two cascaded *for* cycles of $O(|V|)$ elementary steps each. The first *for* computes the ultrametric network for the vertices in $\bar{V}$, whereas the second one updates the queue Q, which stores all vertices $v \notin \bar{V}$ according to the index $key[v]$. All the operations in each *for* take trivially constant time, except for the queue updates. If the queue is implemented as a Fibonacci heap, we can extract the minimum element in amortized $O(\log |V|)$ and update the queue in amortized $O(1)$, when $key[v]$ is decreased. There are $(|V| - 1)$ *extractmin* (at the beginning of every iteration of the *while*), which thus charge $O(|V| \log |V|)$ overall. There are $O(|V|^2)$ constant-time updates throughout all the executions of the second *for* loop. Hence the total cost of the algorithm is $O(|V|^2)$. The subdominant ultrametric requires $\Theta(|V|^2)$ entries, and an ultrametric network contains at most $|V|^2$ edges, so that $O(|V|^2)$ time is optimal. □

Ultrametric network relaxations

The algorithm of the preceding section lends itself naturally to variants that accommodate some tolerance in the ultrametric distance and relax the notion of ultrametric network. We outline here these two variants, respectively leading to Δ-ultrametric networks and to the introduction of new artificial vertices.

Δ-Ultrametric extension

We define the Δ *-ultrametric network* in which edges are inserted if their weights do not deviate more that a given threshold Δ from the corresponding ultrametric distance.

Formally the Δ *-ultrametric network* consists of the graph $G_\Delta(V|W) = \left(V, E_\Delta(V|W)\right)$ with vertex set V and edge set $E_\Delta(V|W) := \left\{\{u, v\} \in \binom{V}{2} : W(u, v) \leq W^*(u, v) + \Delta\right\}$, where $W^*(u, v)$ is the standard ultrametric distance. Intuitively, the Δ-ultrametric network is thus a relaxation of the ultrametric network, resulting in increased connectivity. More precisely the graph $G_\Delta(V|W)$ coincides with the map $\min(W, W^* + \Delta)$.

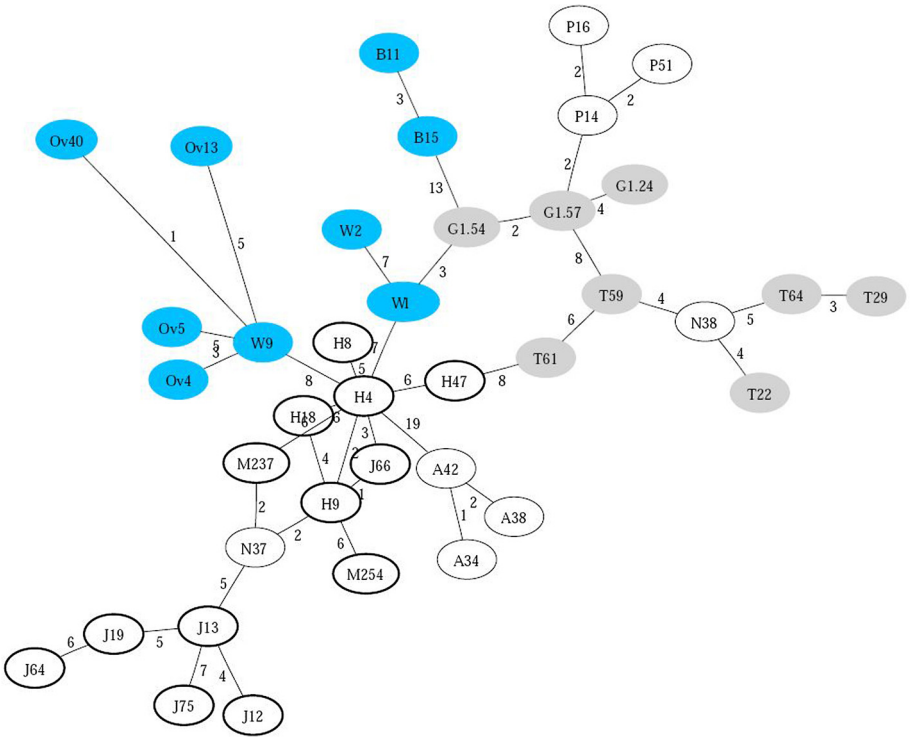

Figure 3 Ultrametric network of 54 individuals from different populations all over the world. $\Delta = 0$ and $\delta = 0$. The labels represent: H = Han, G = German, T = Turk, J = Japanese, Ov = Ovambo, W = Western Pygmy, A = Australian Aboriginal, P = Papuan, B = San (Bushman), and M= Mongolian. A Rough classification is: J+H+M=East Asian, G+T= Caucasoid, A+P=Australian, B+W+Ov= Africans. Nodes filled with gray represent Caucasoid, the ones filled with blue represent Africans, bold nodes are Asians, and the remaining ones are from Australia.

The Δ-ultrametric network can be computed as a post-processing by adding all such edges $E_\Delta(V|W)$ to the *exact* ultrametric network. In summary at first we run the algorithm on the original weights $W(u,v)$ to compute W^*. Then, we apply the postprocess that includes all edges (u,v) with weight $W(u,v)$ that deviates at most Δ from the corresponding ultrametric value $W^*(u,v)$. Similarly to the main algorithm, this postprocess takes optimal $O(|V|^2)$ time. Other alternatives relaxations can be explored, like the subdominant Δ-ultrametrics for which analogous results can be established. The subdominant Δ-ultrametrics relaxation will be addressed in a future paper.

Artificial vertices

In applications such as phylogeny on biological data of extant species/individuals, the topology must account for missing data points. In other words, there is a need to reduce the distance between a pair of vertices by introducing a new artificial vertex in the network.

In the phylogeny construction problem, the given data points are the terminals and the artificial vertices correspond to missing (or ancestral) data points. The traditional Steiner tree problem [23] involves the minimization of the sum of the lengths of all edges used after introducing artificial vertices, as opposed to the sum of the pairwise distances of all the terminals. For different metrics the Steiner tree problem is known to be NP-Hard [24]. Thus in our context, given the graph induced by the ultrametric W^*, the problem of introducing new artificial vertices that minimize the sum of the weights of all edges is still NP-Hard.

In our case the input graph G is not just any graph, but it can be characterized as follow. Suppose we are given a (big) metric space $\mathcal{R} = (V, D)$ (could be the metric space consisting of the vertex set R of a connected weighted graph G with the "induced" metric, i.e., the largest metric D on V with $D(u,v) \leq w(u,v)$ for all edges u, v in G), and a finite subset R of V.

The solution to the Steiner tree problem is to find a connected graph $G(\mathcal{R}|V)$ with a vertex set containing R and contained in V and edge set $E(\mathcal{R}|V)$ such that $\sum_{\{u,v\}\in E(\mathcal{R}|V)} D(u,v)$ is minimized. In case you just have a metric D on V, in our case W^*, the most natural choice for (V, D) is the tight span

$$T(D) := \{f \in R^V : f(v) = \sup(D(u,v) - f(u)) : u \in V\}$$

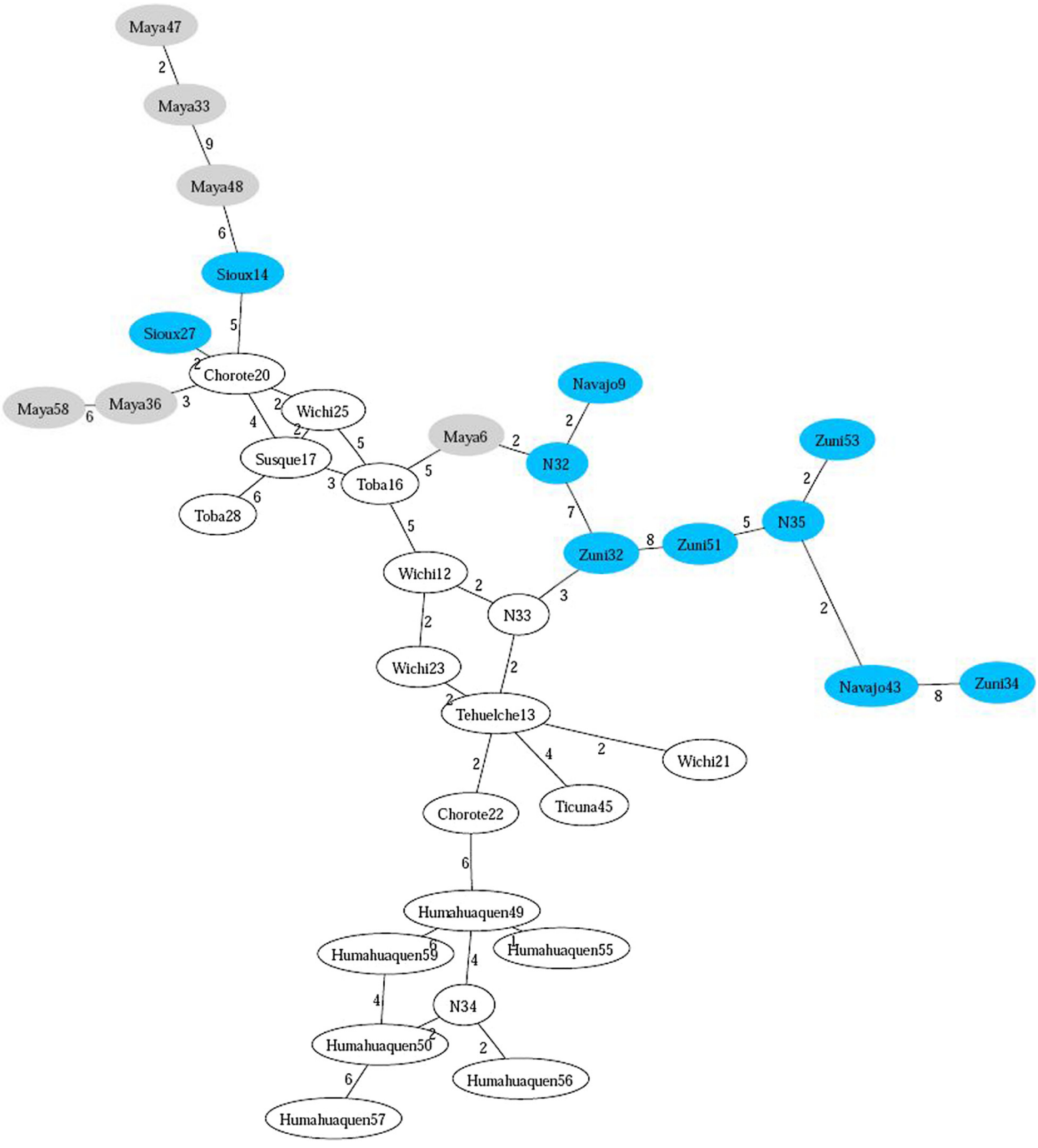

Figure 4 Ultrametric network of 41 new world natives. A rough classification is: North America (Navajo, Zuni, Sioux) in blue, Central (Maya) in grey, South (Ticuna, Wichi, Toba, Chorote, Tehuelche), (Susque, Humahuaquen) in white.

of (V, D). It is natural in this case as any Steiner tree for V can be mapped into $T(D)$ by a non-expanding map that preserves the distances between the points in V[25]. We don't know how to efficiently search for the best Steiner tree within $T(D)$, also known as the optimal realization, without an exhaustive enumeration [26,27]. Instead starting from a given graph G_{W^*} over R we look at the "neighborhood" of this seed in $T(D)$. The input graph G_{W^*} over R is the ultrametric network computed in the previous section and we are interested in the "neighborhood" of this network such that the sum of all edges is smaller and that the pairwise distances in R are preserved.

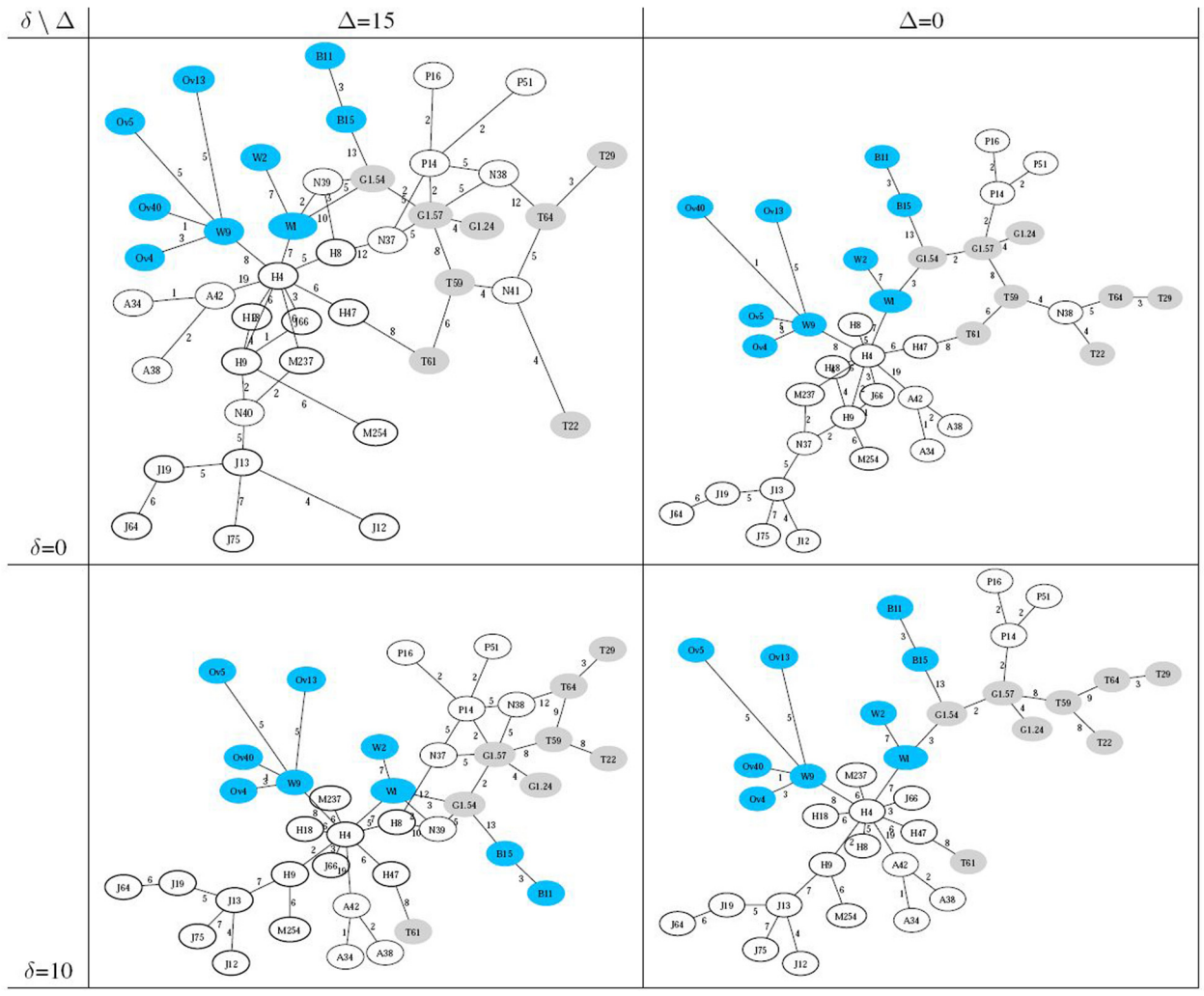

Figure 5 Ultrametric network of 54 individuals from different populations all over the world (same as in Figure 3) for a few values of parameters Δ and δ.

For every edge $\{y, w\}$ in G_{W^*} we search in its neighborhood by looking at the vertices directly connected to y and w. If there is some vertex u that is connected to y and w we explore the possibility to insert a new node x as the median of y, w and u. Clearly if we add the artificial vertex x and replace the edges that create a cycle (u, w), (u, y) and (w, y), with the edges (u, x), (w, x) and (y, x) (see Figure 2) this new configuration does not increase the contributions of the distances involving the three nodes.

To control the number of artificial vertices, the new vertex x is created only if the sum of pairwise distances of the triangle among u, w and y exceeds the threshold δ. Note that if two triangles share an edge we need to select where to insert the new artificial vertex. A canonical order can be established by ranking all candidate triangles by the sum of pairwise distances. This ensures that, at least generically, the introduction of new artificial vertices is unique and does not depend on the input order.

Experimental results

To conclude our presentation, we report two examples of inference of Human Y-chromosome phylogeny from Short Tandem Repeats. This can be based on the study of Human migration and the associated relationships among different populations. In typical experiments, we are interested in constructing a network from the STRs information of various individuals and in comparing the results with known paths of migrations. An interesting example of such a phylogeny reconstruction can be found in [28], which discusses the significance of STRs data as markers for human evolution, but also highlights the difficulties that the analysis of this data derives from the lack of an appropriate methodology.

Using the same data of [28,29], we study migration histories within two different scenarios. In the first experiment, we analyze a very broad spectrum of populations: Africans, Europeans, Asians and Australians. In the second, we concentrate on Native Americans spanning North America (Navajo, Zuni, Sioux), Central America (Maya), and South America (Ticuna, Wichi, Toba, Chorote, Tehuelche, Susque, Humahuaqueño).

For both experiments the data available include a number of different STRs, specifically, 12 in the first experiment and 7 in second. The first step is to establish for all STRs a weighting scheme reflecting the different mutation rates. To this end, we use the three weights 1,2 and 4, and assign to each STR a weight proportional to its mutation rate.

Figure 3 shows the ultrametric network computed from the first dataset. The labels associated with each node are reported in the figure's caption, and new nodes are tagged with the letter "N". The pairwise distances between nodes are reported as attributes of the edges. In Figure 3, the populations are grouped by continent; the nodes filled with gray represent Caucasoid, the ones filled with blue represent Africans, bold nodes are Asians and the remaining ones are from Australia.

We can observe that, in general, all different continents are well separated, and that most of the individuals belonging to the same population are clustered together: Japanese, Han, Turkish, Australian, and so on. Moreover, the known paths of migration support the view that Han are close relatives of Africans and that Japanese evolved from Central Asian populations. Also, Germans appear to be related to Northern Africans and Turks, the latter are also connected with Han, thus supporting the idea that Turks are partially Asians. The only population slightly misplaced are Papuans probably because the STRs examined do not resolve for this population, a problem already observed in [28]. Figure 4 shows the ultrametric network of Native Americans, using the same data as in [29]. This experiment is a particularly difficult test, due to the high level of homoplasy and the small number of STRs available. Nevertheless the network still exposes the structural diversity between North American Natives (blue), Central (gray) and South Americans (white).

As discussed, the connectivity of the inferred network can be fine-tuned by setting the two control parameters δ and Δ. The first one is used to filter out the feeblest edges: with $\delta = 0$, all links are selected. The value assigned to the second parameter sets the tolerance within which edges are included in the Δ-ultrametric network. Thus, large values of δ reduce the number of artificial points introduced, large values of Δ increase the connectivity of the network. Space limitation prevents a thorough analysis of these variants. As an illustration, Figure 5 displays the networks obtained in correspondence with a few different settings.

Conclusions

In conclusion, this paper expands the *subdominant ultrametric* perspective by studying ultrametric *networks*. We shown that, for a graph with n vertices, the construction of such a network can be carried out by a simple algorithm in optimal time $O(n^2)$. This algorithm can be easily adapted to compute relaxed networks, such as Δ-ultrametric networks and to support the introduction of artificial points to reduce the maximum distance between vertices in a pair. Finally, we discussed a few experiments to demonstrate the applicability of subdominant ultrametric networks.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed equally to this work. All authors read and approved the final manuscript.

Acknowledgements

Enrico Guariento implemented and tested the software Ultranet. M. Comin was partially supported by the Ateneo Project *CPDA110239*.

Author details

[1]College of Computing, Georgia Institute of Technology, 801 Atlantic Drive, Atlanta, GA 30332, USA. [2]Department of Information Engineering, University of Padova, Via Gradenigo 6/A, 35131 Padova, Italy. [3]CAS-MPG Partner Institute and Key Lab for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, and infinity, Bielefeld, Germany. [4]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA.

Received: 15 October 2012 Accepted: 18 February 2013
Published: 5 March 2013

References

1. Agarwala R, Bafna V, Farach M, Narayanan B, Paterson M, Thorup M: **On the approximability of numerical taxonomy: Fitting distances by tree metrics.** *Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms* 1996, **28**(3):1073–1085 .
2. Farach M, Kannan S, Warnow T: **A Robust model for finding optimal evolutionary trees.** *Algorithmica, special issue on Computational Biology* 1996, **13:**155–179.
3. Gower J, Ross G: **Minimum spanning trees and single linkage cluster analysis.** *Appl Stat* 1969, **18:**54–64.
4. Florek K, Lukaszewickz J, Perkal H, Steinhaus H, Zubrzycki S: **Sur la Liaison et la Division des Points d'un Ensemble Fini.** *Colloq Matematicum* 1951, **2:**282–285.
5. Edwards A, Sforza LC: **Reconstruction of evolutionary trees.** *Phenetic Phylogenet Classif* 1964, **6:**67–76.
6. Gromov M: **Hyperbolic groups, essays in group theory, MSRI series vol. 8, S.Gersten, ed., Springer-Verlag (1988). trees and single linkage cluster analysis.** *Appl Stat* 1969, **18:**54–64.
7. Bowditch B: **Notes on Gromov's hyperbolicity criterion for path metric spaces.** In *E. Ghys et al.: E. Ghys et al., Proceedings of Group Theory from a Geometric Viewpoint, World Scientific*. Singapore; 1991:64–167.
8. Dress A, Huber K, Moulton V: **Some uses of the Farris Transform in Mathematics and Phylogenetics – A Review.** *Ann Combinatorics, Special Volume Biomathematics* 2007, **11:**1–37.
9. Bayod JM, Martinez-Maurica J: **Subdominant ultrametrics.** *Proc Am Math Soc* 1990, **109**(3):829–834.
10. Parisi G: **Spin glasses and fragile glasses: Static, dynamics and complexity.** *PNAS* 2006, **103**(21):7948–7955.

11. Rammal R, Toulouse G, Virasoso MA: **Ultrametricity for Physicists.** *Rev Mod Phys* 1986, **58:**765–788.
12. Soete GD: **Ultrametric tree representations of incomplete dissimilarity data.** *J Classif* 1984, **1:**235–242.
13. Huson D, Nettles S, Warnow T: **Obtaining accurate topology estimates of evolutionary trees from very short sequences.** *Proc RECOMB* 1999 :198–207.
14. Bandelt HJ, Forster P, Rühl A: **Median-joining networks for inferring intraspecific phylogenies.** *Mol Biol Evol* 1999, **16:**37–48.
15. Bryant D, Moulton V: **Neighbor-Net: An agglomerative method for the construction of phylogenetic networks.** *Mol Biol Evol* 2004, **21:**255–265.
16. Grünewald S, Forslund K, Dress A, Moulton V: **QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets.** *Mol Biol Evol* 2007, **24:**532–538.
17. Saitou N, Imanishi T: **Relative effciencies of the fitch-margoliash, maximum parsimony, maximum likelihood, minimum evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree.** *Mol Biol Evol* 1989, **6:**514–525.
18. Excoffier L, Smouse P: **Using allele frequencies and geographic subdivision to reconstruct gene genealogies within a species: Molecular variance parsimony.** *Genetics* 1994, **136:**343–359.
19. Templeton A, Crandall K, Sing C: **A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation.** *Genetics* 1992, **132:**619–633.
20. Hart M, Sunday J: **Things fall apart: biological species form unconnected parsimony networks.** *Biol Lett* 2007, **3:**509–512.
21. Clement M, Posada D, Crandall K: **TCS: a computer program to estimate gene genealogies.** *Mol Ecol* 2000, **9:**1657–1659.
22. Cormen TH, Leiserson CE, Rivest RL, Stein C: *Introduction to Algorithms*. Cambridge: MIT Press; 2001.
23. Hwang FK, Richards DS, Winter P: *The Steiner Tree Problem, Volume Annals of Discrete Mathematics*. Amderdam: Elsevier Science Publishes; 1992.
24. Bern M, Eppstein D: **Hardness of approximations algorithms for geometric problems.** In *Approximation algorithms for NP-Hardness Problems edited by D. S. Hochbaum*. Boston, MA: PWS Publishing Company; 1997:296–345.
25. Dress A, Huber KT, Moulton V: **Metric spaces in pure and applied mathematics.** *Documenta Mathematica (Proceedings Quadratic Forms LSU)* 2001:121–139.
26. Dress A, Huber KT, Koolen JH, Moultonm V: **Block realizations of finite metrics and the tight-span construction I: The embedding theorem.** *Discrete Appl Math* 2008, **21**(12):1306–1309.
27. Dress A, Wu T, Xu X: **A note on single-linkage equivalence.** *Appl Math Lett* 2010, **23:**432–435.
28. Forster P, Rohl A, Lunnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, Brinkmann B: **A short tandem repeat-based phylogeny for the human Y chromosome.** *Am J Hum Genet* 2000, **67:**182–196.
29. Bianchi N, Catanesi C, Bailliet G, Martinez-Marignac V, Bravi C, Vidal-Rioja L, Herrera R, Lopez-Camelo J: **Characterization of ancestral and derived Y-chromosome haplotypes of new world native populations.** *Am J Hum Genet* 1998, **63**(6):1862–1871.

doi:10.1186/1748-7188-8-7
Cite this article as: Apostolico *et al.*: Ultrametric networks: a new tool for phylogenetic analysis. *Algorithms for Molecular Biology* 2013 **8**:7.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central