Molecular Distance Geometry and Atomic Orderings

Antonio Mucherino
IRISA, University of Rennes 1

Workshop "Set Computation for Control"
ENSTA Bretagne, Brest, France
December 5^{th} 2013

The Distance Geometry Problem
(for molecular conformations)

Let $G = (V, E, d)$ be a simple weighted undirected graph, where

- V the set of vertices of G – it is the set of atoms;
- E the set of edges of G – it is the set of known distances;
- $E' \subset E$ the subset of E where distances are exact;
- d the weights associated to the edges of G
  the numerical value of each weight corresponds to the known distance; it can be an interval.

Definition

The **DGP**.
Determine whether there exists a function $x : V \longrightarrow \Re^K$ for which, for all edges $(u, v) \in E$, $||x_u - x_v|| = d(u, v)$.

Sphere intersections and discretization

In the 3-dimensional space, the intersection of

- 2 spheres *gives* one circle
- 3 spheres *gives* two points
- 2 spheres and 1 spherical shell *gives* two disjont curves

Definition of the spheres / spherical shells:

- center $=$ vertex w with known position
- radius $=$ known distance between w and a common vertex (to be placed)

Precision of distance information $\Longrightarrow$ sphere / spherical shell

Importance of orders

The definition of an order on the vertices in V allows us to *ensure that vertex coordinates* are available when needed.

Given

1. a simple weighted undirected graph $G = (V, E, d)$
2. a vertex $v \in V$

how to identify K vertices w_i, with $i = 1, 2, \ldots, K$, for which

- the coordinates of every w_i are available
- every edge $(w_i, v) \in E$

????

We refer to w_i as a reference vertex for v
(w_i, v) as a reference distance for v

Definition of order

Definition

An order for V is a sequence $r : \mathbb{N} \to V \cup \{0\}$ with length $|r| \in \mathbb{N}$ (for which $r_i = 0$ for all $i > |r|$) such that, for each $v \in V$, there is an index $i \in \mathbb{N}$ for which $r_i = v$.

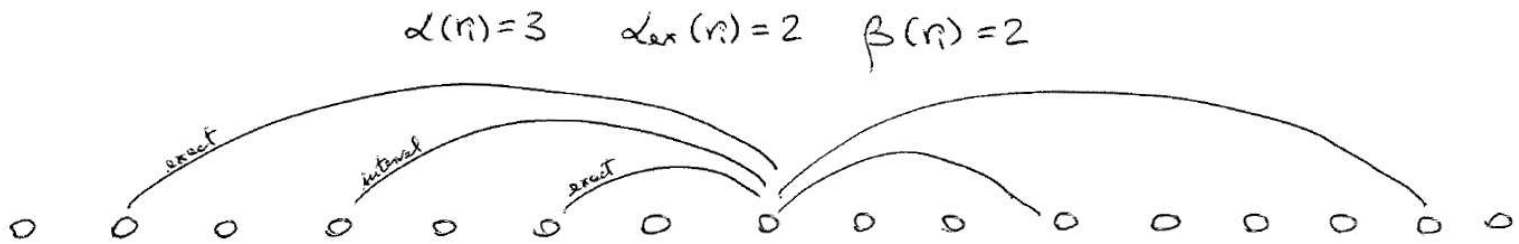
Orders and vertex repetitions:

- they allow for vertex repetitions ($|r| \geq |V|$);
- however, each vertex can be used as a reference only once;
- simplex inequalities (generally satisfied with probability 1) would not be satisfied if the same vertex were used twice as a reference.

Counting the reference vertices

Let r be an order for V.
Let us consider the following counters.

- $\alpha(r_i)$: counter of adjacent predecessors of r_i;
- $\beta(r_i)$: counter of adjacent successors of r_i;
- $\alpha_{ex}(r_i)$: counter of adjacent predecessors of r_i related to an exact distance.



Necessary condition for V to admit a *discretization order* is that, for any order r on V without repetitions,

$$\forall i \in \{1, 2, \ldots, |r|\}, \ \alpha(r_i) + \beta(r_i) \geq K.$$

The ordering problem

Definition

The *interval* Discretization Vertex Order Problem (*i*DVOP).

Given a simple weighted undirected graph $G = (V, E, d)$ and a positive integer K, establish whether there exists an order r such that:

(a) $G_C = (V_C, E_C) \equiv G[\{r_1, r_2, \ldots, r_K\}]$ is a clique and $E_C \subset E'$;

(b) $\forall i \in \{K+1, \ldots, |r|\}$, $\alpha(r_i) \geq K$ and $\alpha_{ex}(r_i) \geq K - 1$.

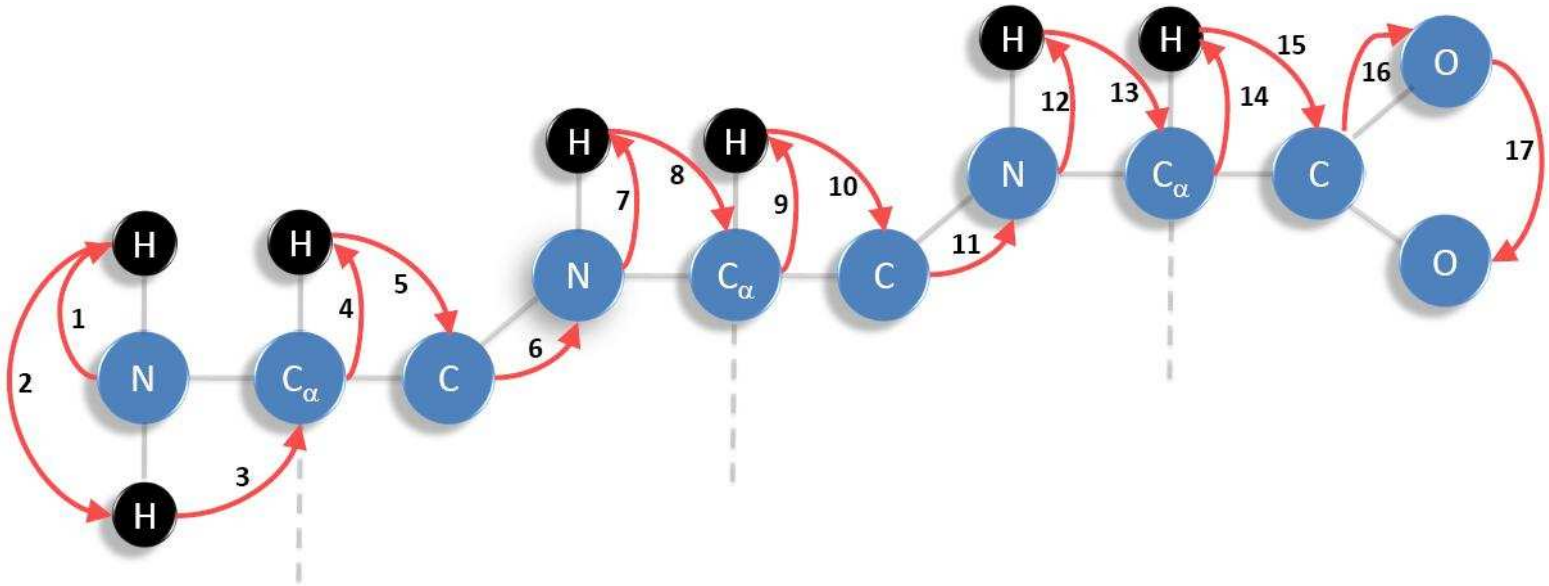
Remarks:

- this problem is NP-complete when K is not fixed
- no consecutivity assumption: solvable in polynomial time when K is known
- when dealing with proteins, $K = 3$

A greedy algorithm

0: reorder(G)
while (a valid order r is not found yet) **do**
  let $i = 0$;
  find a K-clique C in G with exact distances;
  // position C at the beginning of new order
  for (all vertices v in C) **do**
    let $i = i + 1$;
    let $r_i = v$;
  end for
  // greedy search
  while (V is not covered) **do**
    $v = \arg\max\{\alpha(u) \mid \nexists j \leq i : r_j = u \text{ and } \alpha_{ex}(u) \geq K - 1\}$;
    if ($\alpha(v) < K$) **then**
      break the inner loop: there are no possible orderings for C;
    end if
    // adding the vertex to the order
    let $i = i + 1$;
    let $r_i = v$;
  end while
end while
return r;

An order for the protein backbone

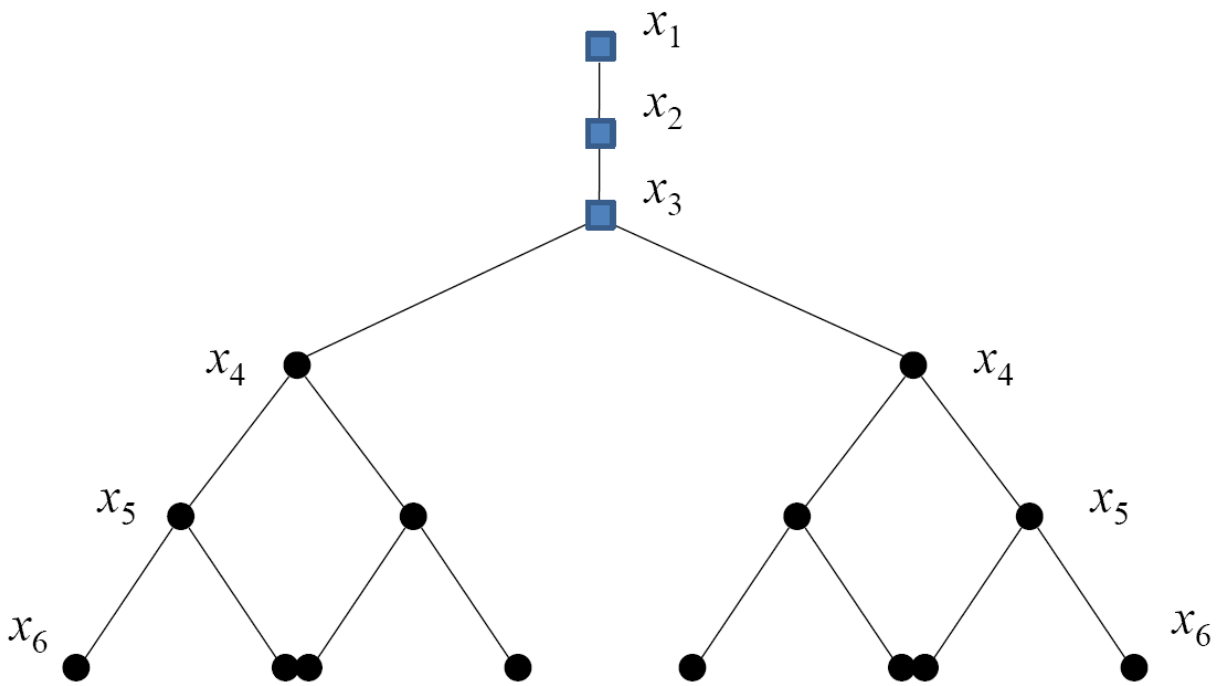


This order was automatically obtained by the greedy algorithm; no NMR distances were supposed to be known.

In presence of NMR data, the algorithm can be applied again for finding an order that is perfectly tailored to the instance at hand.

The Branch & Prune algorithm

The Branch & Prune (BP) algorithm is based on the idea of branching over all possible positions for each vertex, and of pruning by using additional distances that are not used in the discretization process (*pruning distances*).



In this tree, it is supposed that all available distances are exact.

If not, D sample (exact) distances can be taken from interval distances.

Generation of NMR-like instances

We consider artificially generated instances that simulate real **NMR** instances. We consider

- the molecular graph, describing the chemical structure of the molecule
- the set of atomic coordinates from the PDB.

We compute all distances between pairs of atoms and we add a distance in our instance if this is a distance between

- two bonded atoms (*exact*)
- two atoms that are bonded to a common atom (*exact*)
- two atoms belonging to a quadruplet of bonded atoms forming a torsion angle (*interval*)
- two hydrogen atoms whose distance is in the interval $[2.5, 5]$ (*interval*)

Computational experiments

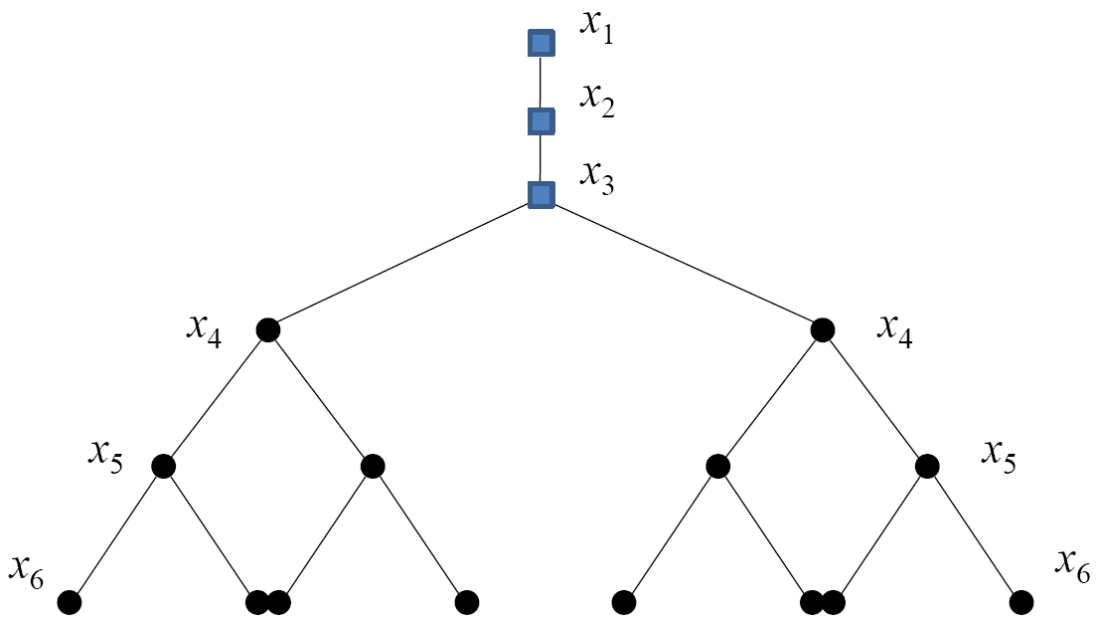
NMR-like instances are considered in these experiments, which include protein backbones and side chains.

Instance	n	$\|E\|$	D	BP calls	Time
1niz	219	1470	6	16425	0.11
1u6u	258	1757	9	1931	0.01
1b03	280	1913	8	4723	0.02
2jnr	293	2004	7	2442	0.01
2pv6	374	2428	8	1100	0.01
1zec	370	2496	9	696836	5.40
2m1a	433	2857	6	10974	0.07
2me1	446	2893	10	801719	13.71
2me4	458	3002	7	96863	1.07
1dsk	465	3181	8	33984	0.16

All instances were automatically reordered by the greedy algorithm, and the BP algorithm was invoked for finding one solution.

The *generalized* BP algorithm ???

The discretization of the curves can lead to the propagation of errors. The number D of sample distances that are taken from each curve plays an inportant role, but it is rather difficult to predict.



What if the nodes of the tree do not represent vertex positions, but rather intervals???

Work in progress . . .

Can we find orders that help BP in finding solutions?

- minimize in length the subsequences in the order having no pruning distances
- minimize the number of vertices that are crossed by the same pruning distance
- *(for proteins)* maximize the interval distances that are related to pairs of hydrogen atoms
- . . .

Other work in progress . . .

- make the branching phase of BP adaptive
- identify clusters of solutions in BP solution sets
- improve and tailor the parallel versions of BP to interval data
- *(for proteins)* exploit energy-based information for pruning purposes
- *(for proteins)* use real NMR data and compare our results to what is currently available on the PDB
- . . . . . .

Thanks!