

Influence of Pruning Devices on the Solution of Molecular Distance Geometry Problems

Antonio Mucherino

www.antoniomucherino.it

CERFACS, Toulouse, France

joint work with:

C. LAVOR, L. LIBERTI, N. MACULAN, T. MALLIAVIN, M. NILGES

SEA11, Crete, Greece,
May 6th 2011



Outline

- 1 The DGP
 - Introduction
 - Our discrete DGP
- 2 Algorithms for the DGP
 - The BP algorithm
 - The *i*BP algorithm
- 3 Pruning devices
 - NMR information
 - Some experiments
- 4 Future works



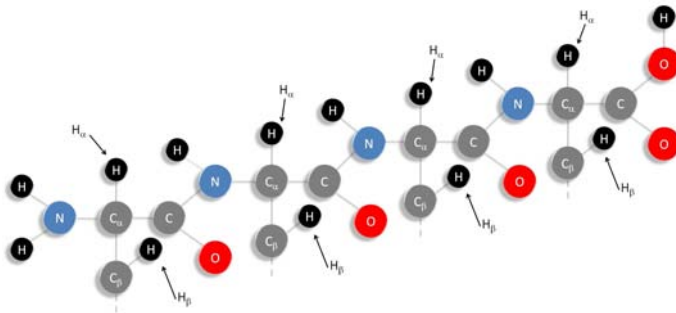
Outline

- 1 The DGP
 - Introduction
 - Our discrete DGP
- 2 Algorithms for the DGP
 - The BP algorithm
 - The *i*BP algorithm
- 3 Pruning devices
 - NMR information
 - Some experiments
- 4 Future works



Protein conformations

Proteins play many vital functions in the bodies of living beings.



They are chains of **amino acids**. They fold into unique three-dimensional conformations, which define their functions.



The Protein Data Bank (PDB)

It's a web database of protein conformations (~72000 on May 2011).

Two techniques are mainly employed:

1 X-ray diffraction

- provides (in general) conformations with a good resolution;
- requires the crystallization of the molecule;

2 Nuclear Magnetic Resonance (NMR)

- analyzes the molecule in solution (crystallization not required);
- only provides information from which the conformation can be obtained.

What does NMR provide?

Distances between some pairs of atoms of the molecule.



The Distance Geometry Problem (DGP)

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

- V the set of vertices of G – corresponds to the subset of atoms;
- E the set of edges of G – corresponds to the set of known distances;
- d the weights associated to the edges of G
the numerical value of each weight corresponds to the known distance.

Definition

Find a conformation $x = \{x_1, x_2, \dots, x_n\}$ such that all the following constraints are satisfied:

$$\|x_i - x_j\| = d_{ij} \quad \forall i, j : i \neq j,$$

where $\|x_i - x_j\|$ is the computed distance between x_i and x_j , and d_{ij} is the generic weight of the graph G .



Outline

- 1 The DGP
 - Introduction
 - **Our discrete DGP**
- 2 Algorithms for the DGP
 - The BP algorithm
 - The *i*BP algorithm
- 3 Pruning devices
 - NMR information
 - Some experiments
- 4 Future works



The Discretizable MDGP (DMDGP)

Instances satisfying the following two assumptions can be discretized.

Given a graph $G = (V, E, d)$,

Ass.1 $(1, 2, 3) \subset V$ must be a clique and,
for each $i \in V$ such that $i > 3$, the distances

$$d_{i-3,i} \quad d_{i-2,i} \quad d_{i-1,i}$$

must be known;

Ass.2 for each $i \in V$ such that $i > 2$, the strict triangular inequality

$$d_{i-2,i} < d_{i-2,i-1} + d_{i-1,i}$$

must hold.



Outline

- 1 The DGP
 - Introduction
 - Our discrete DGP
- 2 Algorithms for the DGP
 - **The BP algorithm**
 - The *i*BP algorithm
- 3 Pruning devices
 - NMR information
 - Some experiments
- 4 Future works



The Branch & Prune algorithm

This reformulation of the problem allows for using a **very efficient** algorithm for its solution.

```

0: BP( $i, n, d$ )
  for ( $k = 1, 2$ ) do
    compute the  $k^{\text{th}}$  atomic position for the  $i^{\text{th}}$  atom:  $x_i^k$ ;
    check the feasibility of the atomic position  $x_i$ :
    if ( $\|x_i - x_j\| = d_{ij}, \forall j < i$ ) then
      the atomic position  $x_i$  is feasible;
      if ( $i = n$ ) then
        a solution is found;
      else
        BP( $i + 1, n, d$ );
      end if
    else
      the current branch is pruned;
    end if
  end for
  
```

Because of the **pruning phase**, branches of the binary tree are pruned quickly, so that an exhaustive search on the remaining branches is not expensive.



Outline

- 1 The DGP
 - Introduction
 - Our discrete DGP
- 2 Algorithms for the DGP
 - The BP algorithm
 - The *i*BP algorithm
- 3 Pruning devices
 - NMR information
 - Some experiments
- 4 Future works



interval Branch & Prune

This is an extension of BP for interval data.

```

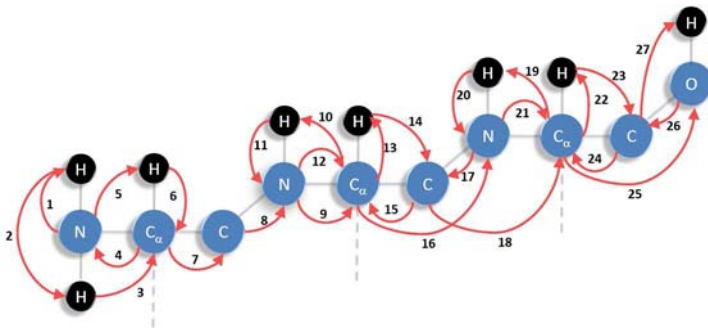
0: iBP(i, n, d, nbranches)
  if (xi is a duplicated atom) then
    assign to xi the same coordinates of its previous copy;
    iBP(i + 1, n, d, nbranches);
  else
    if (d(i - 3, i) is exact) then
      b = 2;
    else
      b = nbranches;
    end if
    for (k = 1, b) do
      compute the kth atomic position for the ith atom: xik;
      check the feasibility of the atomic position xik:
      if (xik is feasible) then
        if (i = n) then
          a solution is found;
        else
          iBP(i + 1, n, d, nbranches);
        end if
      else
        the current branch is pruned;
      end if
    end for
  end if
end if

```



A special ordering for proteins

The *i*BP algorithm can be applied if, for each atom i , no more than one reference distance is represented by an interval.



If this ordering is considered, the distances $d(i, i + 1)$ and $d(i, i + 2)$ are always exact.

Outline

- 1 The DGP
 - Introduction
 - Our discrete DGP
- 2 Algorithms for the DGP
 - The BP algorithm
 - The *i*BP algorithm
- 3 Pruning devices
 - **NMR information**
 - Some experiments
- 4 Future works



What's NMR able to provide?

NMR experiments are *not only able* to provide a list of lower and upper bounds on the distances between pairs of hydrogen atoms, but also:

- a list of lower and upper bounds for the **torsion angles** that can be defined on the protein backbone
- the **secondary structure** in which each amino acid is contained

Our pruning devices:

- 1 **DDF** – Direct Distance Feasibility
- 2 **TAF** – Torsion Angle Feasibility (**new**)
- 3 **SSF** – Secondary Structure Feasibility (**new**)



Outline

- 1 The DGP
 - Introduction
 - Our discrete DGP
- 2 Algorithms for the DGP
 - The BP algorithm
 - The *i*BP algorithm
- 3 Pruning devices
 - NMR information
 - **Some experiments**
- 4 Future works



Computational experiments

<i>instance name</i>	<i>n_{aa}</i>	<i>n</i>	<i>D</i>	<i>iBP</i> calls	#DDF	#TAF	#SSF	CPU time
2jmy	15	134	15	4724652	2356670	-	-	39
2jmy	15	134	15	10482	5244	2695	-	1
2jmy	15	134	15	31986247	15206046	-	6189223	248
2jmy	15	134	15	33709275	16017742	1069321	5156934	298
2ppz	36	323	20	98807	48586	-	-	1
2ppz	36	323	20	91466	43568	41600	-	2
2ppz	36	323	20	414926692	142727215	-	70158539	10263
2ppz	36	323	20	58296108	18941155	10111249	615926	1471
2jwu	56	503	22	6528633	6715391	-	-	117
2jwu	56	503	22	11159985	28183553	1029437	-	396
2jwu	56	503	22	20119294	14742376	-	1601915	432
2jwu	56	503	22	44795676	19494850	9450743	5313997	1363

D is the number of sample distances taken from known intervals.



There is still a lot of interesting work to do!

Future works

Theory

- prove that the number of solutions for the reformulated problems is always a **power of 2**
- prove that BP and *i*BP are **polynomial** on protein instances
- study the **symmetry properties** of the reformulated problems
- ...

Related subproblems

- study the problem of finding a **vertex ordering** for a graph such that the assumptions for applying BP or *i*BP are satisfied
- solve the problem of identifying the conformations of amino acid **side chains** for a given protein backbone
- ...

Implementation

- implement BP and *i*BP algorithms for **parallel** and **distributed** environments
- ...



Thanks!

`www.antoniomucherino.it`

