A Branch and Prune Algorithm for the Molecular Distance Geometry Problem

Carlile Lavor\textsuperscript{1}, Leo Liberti\textsuperscript{2}, Nelson Maculan\textsuperscript{3}

\textsuperscript{1} IMECC-UNICAMP, Campinas, Brazil
\textsuperscript{2} LIX, École Polytechnique, France
\textsuperscript{3} COPPE, UFRJ, Rio de Janeiro, Brazil
Summary of talk

- Problem definition
- Literature
- Discrete formulation for protein backbone
- Branch and Prune
- Computational results
Problem definition

- Protein functionality determined by 3D structure
- Known set of atoms $V$
- Some inter-atomic distances $d_{ij}$ known (NMR)
- Find atomic positions which preserve distances
- Molecular Distance Geometry Problem (MDGP): find an immersion in $\mathbb{R}^3$ of a given undirected, weighted graph $G = (V, E, d)$
Problem formulation

- Solve set of $|E|$ nonlinear equations

$$||x_i - x_j|| = d_{ij} \quad \forall \{i, j\} \in E,$$  \hspace{1cm} (1)

where $x_i = (x^1_i, x^2_i, x^3_i) \in \mathbb{R}^3$ is the position of the $i$-th atom

- Usual formulation: nonconvex least-squares problem

$$\min_x \left( f(x) = \sum_{\{i, j\} \in E} (||x_i - x_j||^2 - d_{ij}^2)^2 \right)$$  \hspace{1cm} (2)

- Leads to nonlinear programming problem (NLP) with many local minima (practically hard to solve)

- Known fact: $x^*$ solves the MDGP $\Leftrightarrow f(x^*) = 0$
Solution accuracy vs. time

• Many different molecular structures with near-zero objective accuracy of solution is of **utmost importance** (CPU time less important)

• Accuracy usually measured as Largest Distance Error:

\[
LDE(x) = \frac{1}{|E|} \sum_{\{i,j\} \in E} \left| \frac{||x_i - x_j|| - d_{ij}}{d_{ij}} \right|
\]
Literature

- Applications in wireless network localization and graph drawing
- Problem is NP-hard even in 1 dimension (reduction from \textsc{Subset-Sum})
- If $G$ is complete, there is a linear-time algorithm
- All methods based on continuous optimization:
  1. Global Continuation Algorithm [Moré & Wu 1999] (solves smoothed problems progressively more similar to original problem)
  2. ABBIE [Hendrickson 1995] (global optimization and recombination of sub-clusters)
  3. DCA [An & Tao 2003] (D.C. formulation)
  4. Double VNS with Smoothing [L., Lavor & Maculan 2005] (VNS applied to smoothed then to original p.)
Competition

- GCA code (dgsol): very fast, not very accurate
- Double VNS with Smoothing: slow, quite accurate
- ABBIE & DCA: could not test them
- On average, satisfactory LDE’s (less than $10^{-5}$) for up to 100 atoms
- Solutions with LDE’s in the order of $10^{-2}$ (unsatisfactory) may be found for up to 700-800 atoms in hours of CPU time on multi-processors / clusters; order of $10^{-1}$ (wrong) for up to 1000 atoms in 15 mins of CPU time on PCs
- **Branch and Prune: LDE of about $10^{-6}$ for up to 1000 atoms found in 1s of CPU time on a PIV 2.66GHz with 1GB RAM**
- Branch and Prune limitation: special class of instances (protein backbones)
Discrete formulation

- **Assumption 1**: presence of linear atomic ordering (proteins)

- Bond lengths/angles known $\Rightarrow$ can compute distances $d_{i,i+2}$

- **Assumption 2**: all distances $d_{i,i+3}$ must be known (fair assumption for NMR)

- **Assumption 3**: bond angles $\neq k\pi$, $k \in \mathbb{Z}$

- Assumptions 1, 2, 3 applicable to almost all proteins
Discretization

- Under assumptions 1, 2, 3 problem can be discretized

\[ i \]

\[ i \]

\[ i + 1 \]

\[ i + 2 \]

\[ i + 3 \]

\[ (i + 3)' \]

- Suppose atoms \( i, i + 1, i + 2 \) fixed; how do we fix \( i + 3 \)?

- **Geometrical intuition:** \( x_{i+3} \) must be on the intersection of the three spheres \( S(x_i, d_{i,i+3}), S(x_{i+1}, d_{i+1,i+3}), S(x_{i+2}, d_{i+2,i+3}) \), which generally consists of at most two points

- Inductive process starts because atoms 1, 2, 3 can be fixed
Failure of assumption 3 (angle multiple of $\pi$) leads to situation below:

Every position for atom $i + 3$ on dashed circle preserves both bond angle $\theta_{i+2}$ and distance $d_{i,i+3}$

Apparently, no real protein has bond angles of $k\pi$, $k \in \mathbb{Z}$
DMDGP

- **Discretizable Molecular Distance Geometry Problem (DMDGP):** MDGP instances for which a discrete formulation can be supplied

**Discretizable Molecular Distance Geometry Problem (DMDGP):** given a weighted undirected graph $G = (V, E, d)$ such that there exists an ordering $v_1, \ldots, v_n \in V$ satisfying the following requirements:

1. $E$ contains all cliques on quadruplets of consecutive vertices:
   \[ \forall i \in \{4, \ldots, n\} \forall j, k \in \{i - 3, \ldots, i\} \ (\{j, k\} \in E); \]
2. for $i \in \{2, \ldots, n - 1\}$, the following strict triangular inequality holds:
   \[ d(v_{i-1}, v_{i+1}) < d(v_{i-1}, v_i) + d(v_i, v_{i+1}) \]

is there a function $x : G \rightarrow \mathbb{R}^3$ such that $||x(u) - x(v)|| = d(u, v)$ for each $\{u, v\} \in E$?

- **Theorem:** the DMDGP is **NP-complete** (reduction from **SUBSET-SUM**)

- **Note:** requiring quintuplet cliques puts this problem in **P**
The Branch-and-Prune algorithm

- Branch the search at each possible atomic position, verify feasibility with distances, prune as necessary
  1. Position atoms 1, 2, 3
  2. Atom \( i \) can be placed in two possible positions \( x_i, x'_i \)
  3. For \( x_i, x'_i \), check feasibility with all distances \( \{ j, i \} \in E \) with \( j < i \) (i.e. check that \( (||x_j - x_i||^2 - d^2_{j,i})^2 < \varepsilon \), where \( \varepsilon > 0 \) is a given tolerance)
  4. Recurse from step 2 on each feasible position with \( i \) replaced by \( i + 1 \); if no position is feasible, prune this search branch and return to previous recursion layer

- Stopping criteria: for one solution, terminate as soon as the last atom is placed in a feasible position. For all solutions, explore the whole search tree.
Example
Tested instances

- Moré-Wu instances (cubic lattices of various sizes): these are “undiscretizable”, so we perturbed the bond angles slightly

- Sizes up to 216 atoms ($= 6^3$)
Tested instances

- Lavor instances (protein backbone shaped; randomly generated from a realistic potential energy model)

- Sizes of up to 1000 atoms
## Results on Moré-Wu instances

| Name    | $n$ | $|E|$ | CPU  | LDE     | CPU  | LDE     |
|---------|-----|------|------|---------|------|---------|
| mmorewu-2 | 8   | 28   | 0.02 | 2.63E+5 | 0.00 | 4.37E-10 |
| mmorewu-3 | 27  | 331  | 0.23 | 6.99    | 0.00 | 2.97E-09 |
| mmorewu-4 | 64  | 1882 | 0.67 | 7.79E-6 | 0.00 | 5.56E-09 |
| mmorewu-5 | 125 | 7105 | 2.94 | 3.54E-6 | 0.00 | 1.67E-08 |
| mmorewu-6 | 216 | 21461| 18.65| 0.032   | 0.02 | 4.91E-08 |

Calculated on PIV 2.66GHz 1GB RAM
## Results on Lavor instances

<table>
<thead>
<tr>
<th>Size</th>
<th>dgsol CPU</th>
<th>dgsol LDE</th>
<th>BP-One CPU</th>
<th>BP-One LDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10.57</td>
<td>3.53E-01</td>
<td>0.57</td>
<td>2.46E-09</td>
</tr>
<tr>
<td>200</td>
<td>57.34</td>
<td>3.61E-01</td>
<td>0.02</td>
<td>2.00E-08</td>
</tr>
<tr>
<td>300</td>
<td>109.91</td>
<td>4.03E-01</td>
<td>0.03</td>
<td>1.90E-08</td>
</tr>
<tr>
<td>400</td>
<td>173.54</td>
<td>6.69E-01</td>
<td>0.10</td>
<td>1.05E-08</td>
</tr>
<tr>
<td>500</td>
<td>273.66</td>
<td>6.19E-01</td>
<td>0.16</td>
<td>4.92E-07</td>
</tr>
<tr>
<td>600</td>
<td>351.15</td>
<td>5.75E-01</td>
<td>0.01</td>
<td>5.47E-08</td>
</tr>
<tr>
<td>700</td>
<td>365.37</td>
<td>7.03E-01</td>
<td>0.82</td>
<td>2.65E-06</td>
</tr>
<tr>
<td>800</td>
<td>583.65</td>
<td>6.54E-01</td>
<td>2.72</td>
<td>1.90E-06</td>
</tr>
<tr>
<td>900</td>
<td>714.39</td>
<td>6.88E-01</td>
<td>1.68</td>
<td>2.85E-06</td>
</tr>
<tr>
<td>1000</td>
<td>787.30</td>
<td>6.88E-01</td>
<td>0.41</td>
<td>1.45E-06</td>
</tr>
</tbody>
</table>

Calculated on PIV 2.66GHz 1GB RAM

Averages taken on 10 instances for each size
Conclusion and future work

- New discrete formulation for the MDGP restricted to proteins
- Branch and Prune algorithm
- 6-9 orders of magnitude improvement on solution accuracy
- Up to 3 orders of magnitude improvement on CPU time
- Can potentially find all solutions (no other method can do this)

**Future work**: test on real proteins, consider experimental NMR errors