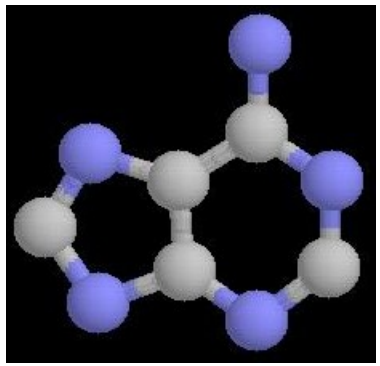
# Molecular Structure Determination
# Distance Geometry – Bound Smoothing

Christodoulos Fragoudakis (cfrag@cs.ntua.gr)



CORELAB seminar

6 December 2004

# Outline

- Introduction – Nuclear Magnetic Resonance,
- Distance Geometry,
- Bound Smoothing,
- Embedding,
- Optimization,

- Triangle Inequality,
- Tetrangle Inequality,
- Discussion – Open Problems.

# Nuclear Magnetic Resonance

- Certain Nuclear Magnetic Resonance (NMR) techniques have made it possible to measure inter-atomic distances for molecular structures as large as 5.000 atoms.

- If two nuclei are very close in space then their spins interact and the frequency required for a spin flip is shifted.

- The peaks in the spectrum become shifted slightly.

- The intensity of this effect depends on the distance between the nuclei.

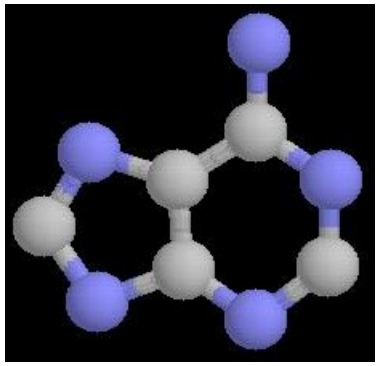
# Distance Geometry

- The distances measured by NMR experiments (usually a small subset of all possible pairs) must be converted into a 3D structure consistent with the measurements.

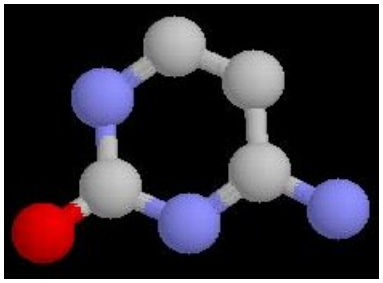
- In general the distances are imprecisely measured: for each distance $d_{ij}$ we have $l_{ij} \leq d_{ij} \leq u_{ij}$.

- The Distance Geometry Method is based on the foundational work of *Cayley* (1841) and *Menger* (1928) who showed how convexity and other basic geometric properties could be defined in terms of distances between pairs of points.

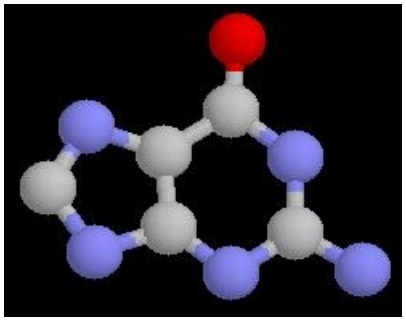
# Small Molecules



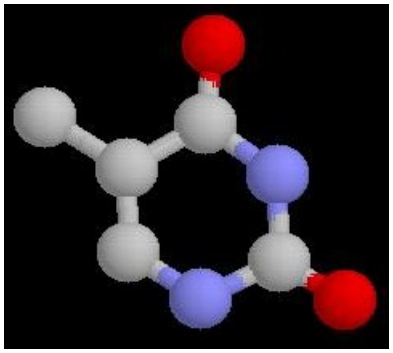
Adenine



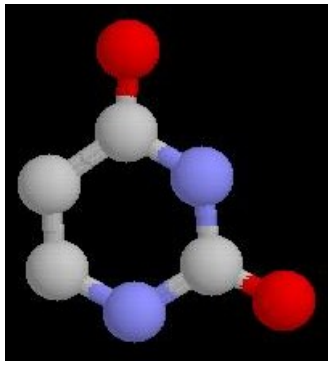
Cytosine



Guanine



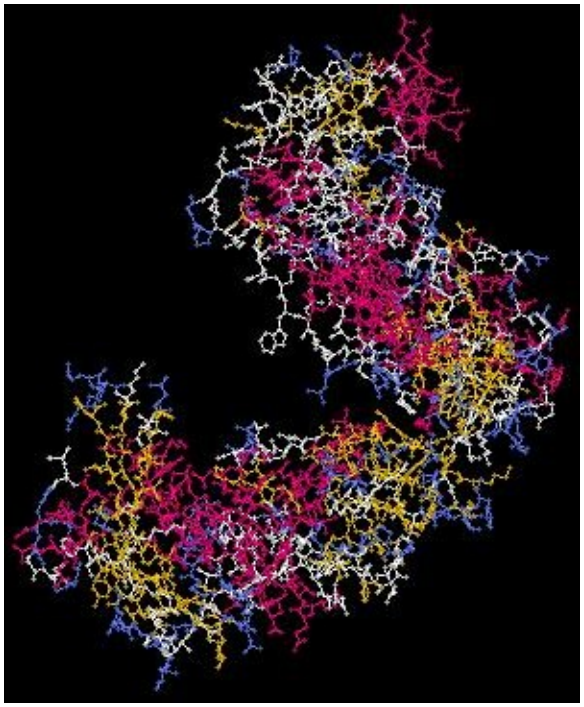
Thymine



Uracile

# A Big Molecule (HIV Retrotranscriptase)

4200 atoms, 554 amino-acids.



# Distance Geometry (cont.)

- Cayley and Menger gave necessary and sufficient conditions for a set of distances to express the inter-point distances of a set of points in the 3-dimensional Euclidean space.

- Blumenthal's monograph:

  L.M. Blumenthal, *Theory and Applications of Distance Geometry*, Oxford, Clarendon Press, 1953.

  is explicitly cited in all relevant papers (afaik).

# Distance Geometry (cont.)

- The Distance Geometry Method has three steps:
  - Bound Smoothing
  - Embedding
  - Optimization

# Bound Smoothing

- Due to imprecision in measurements, inter-atomic distances are specified as pairs of upper and lower bounds.

- It is necessary to tighten the intervals defined by the distance bounds.

- Moreover, since only a small subset of all the pairwise distances can be measured experimentally, bounds on the remaining distances must be computed.

# Embedding

- For each atom pair $i$, $j$ a distance is selected randomly from the tightened interval $\bar{l}_{ij}, \overline{u}_{ij}$.

- The selected distances are inserted into a special $n \times n$ matrix and the calculation of the coordinates of the points, is reduced to the calculation of the eigenvalues of the special matrix. This is done using the Singular Value Decomposition method which finishes after $O(n^3)$ floating point operations.

- Dong and Wu proposed an $O(n^3)$ "Geometric Build-Up" Algorithm (2002). Recently, Hilaris and Piliouras improved to $O(e)$.

# Optimization

In order to further optimize the obtained structure, the chemist uses his a priori knowledge of the molecule structure (chilarity, torsion angles, energy constraints, planarity, atomic repulsion) employing iterative methods, such as:

- simulated annealing
- conformational space sampling
- torsion angle space minimization

# General Practice

- The steps of embedding and optimization are repeated for several selections of the special matrix and the best obtained structure is chosen.

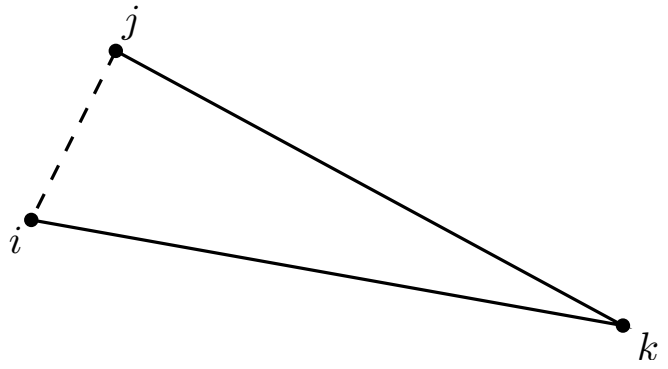
- The quality of the computed structure depends crucially on how tight the distance bounds are.

- Bound Smoothing is an extremely important step.

# A triple of atoms

For any three points in $\mathbb{R}^3$ the triangle inequality holds:

$$\left|d_{ik} - d_{jk}\right| \leq d_{ij} \leq d_{ik} + d_{jk}$$




$$\begin{array}{l|c} l_{ij} \leq d_{ij} \leq u_{ij} \\ l_{ik} \leq d_{ik} \leq u_{ik} \\ l_{jk} \leq d_{jk} \leq u_{jk} \end{array} \quad \begin{array}{c} \overline{u}_{ij} = \mathsf{min}\{u_{ij}, u_{ik} + u_{jk}\} \\ \bar{l}_{ij} = \mathsf{max}\{l_{ij}, l_{ik} - u_{jk}, l_{jk} - u_{ik}\} \end{array}$$

# Remarks

- The tightened upper bounds can be computed independently of the lower.

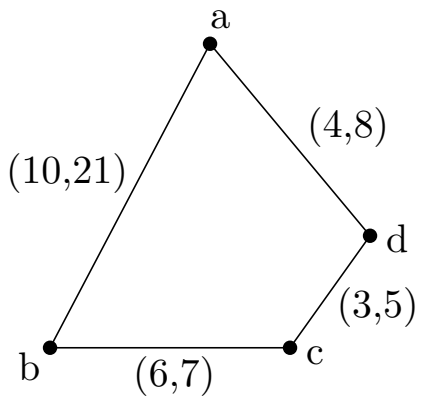
- The tightened upper bounds further improve the tightened lower bounds:

  - $\overline{u}_{ik} \leq u_{ik}$, $\overline{u}_{jk} \leq u_{jk}$, $\bar{l}_{ij} = \max\{l_{ij}, l_{ik} - u_{jk}, l_{jk} - u_{ik}\}$, so

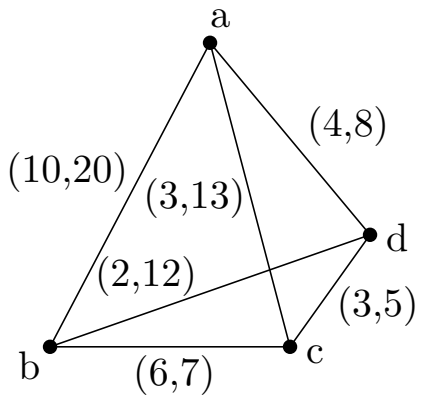
  - $\bar{l}_{ij} = \max\{l_{ij}, l_{ik} - \overline{u}_{jk}, l_{jk} - \overline{u}_{ik}\}$

- If $\bar{l}_{ij} > \overline{u}_{ij}$ we have a triangle inequality violation so we can safely discard all the erroneous constraints that involve atoms $i$, $j$ and $k$.

# Example




$$
\begin{array}{lcll}
l_{bd} & = & l_{ab} - u_{ad} & = & 2 \\
l_{ac} & = & l_{ab} - u_{bc} & = & 3 \\
u_{bd} & = & u_{bc} + u_{cd} & = & 12 \\
u_{ac} & = & u_{ad} + u_{cd} & = & 13 \\
u_{ab} & = & u_{ad} + u_{bd} & = & 20
\end{array}
$$



# The Algorithm

TriangleBoundSmoothing $(L, U)$

**for** each $\binom{n}{3}$ triples of atoms $(i, j, k)$ **do**

$\overline{u}_{ij} \leftarrow \min\{u_{ij}, u_{ik} + u_{jk}\}$

**end for**

**for** each $\binom{n}{3}$ triples of atoms $(i, j, k)$ **do**

$\overline{l}_{ij} \leftarrow \max\{l_{ij}, l_{ik} - u_{jk}, l_{jk} - u_{ik}\}$

**if** $\overline{l}_{ij} > \overline{u}_{ij}$ **then**

**error** ▷ Triangle Inequality Violation!

**end if**

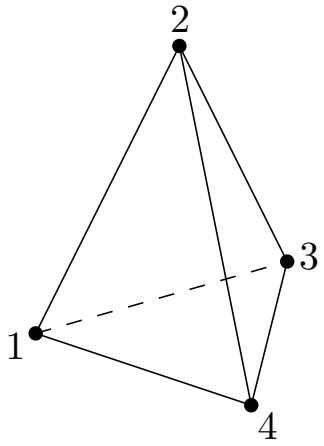
**end for**

**return** $\overline{L}$ and $\overline{U}$

After $O(n^3)$ steps we have a complete set of bounds without incosistences that cannot be further improved.

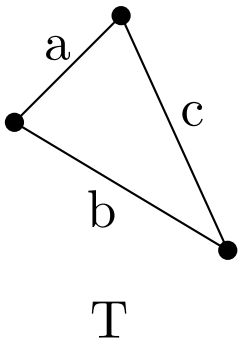
# A quadruple of atoms

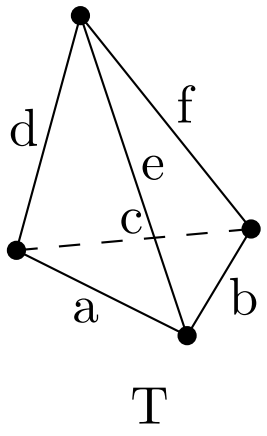
Often tighter bounds can be obtained by using the tetrangle inequality, which in the case of four atoms $\{1, 2, 3, 4\}$, is expressed by means of Cayley–Menger determinants:




$$CM = \begin{vmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 & d_{14}^2 \\ 1 & d_{21}^2 & 0 & d_{23}^2 & d_{24}^2 \\ 1 & d_{31}^2 & d_{32}^2 & 0 & d_{34}^2 \\ 1 & d_{41}^2 & d_{42}^2 & d_{43}^2 & 0 \end{vmatrix} > 0$$

# Heron's formula




- $CM(a,b,c) = 16(\text{Area of T})^2 = (a+b+c)(-a+b+c)(a-b+c)(a+b-c)$

- $CM(a,b,c,d) = 288(\text{Volume of T})^2$

# Necessary and sufficient conditions

- It is well known that $CM(d_{12}, d_{13} \ldots, d_{34}) > 0$ along with the triangle inequality, are the necessary and sufficient conditions for the numbers $d_{12}, \ldots, d_{34}$ to be the distances among a quadruple of points in the 3–D Euclidean space.

- There are exactly $2^6$ inequalities of the form $CM(\ldots) > 0$ each with either $u_{ij}$ or $l_{ij}$ for a $d_{ij}$.

- From now on, w.l.o.g., we are going to consider the case of the $(3, 4)$ distance.

# Non-redundant Inequalities

It turns out that only seven inequalities are non-redundant:

- For the upper limit $u_{34}$:

$$\begin{aligned} CM(l_{12}, u_{13}, u_{14}, u_{23}, u_{24}, u_{34}) &> 0 \\ CM(u_{12}, l_{13}, l_{14}, u_{23}, u_{24}, u_{34}) &> 0 \\ CM(u_{12}, u_{13}, u_{14}, l_{23}, l_{24}, u_{34}) &> 0 \end{aligned}$$

- For the lower limit $l_{34}$:

$$\begin{aligned} CM(u_{12}, u_{13}, l_{14}, l_{23}, u_{24}, l_{34}) &> 0 \\ CM(u_{12}, l_{13}, u_{14}, u_{23}, l_{24}, l_{34}) &> 0 \\ CM(l_{12}, l_{13}, u_{14}, l_{23}, u_{24}, l_{34}) &> 0 \\ CM(l_{12}, u_{13}, l_{14}, u_{23}, l_{24}, l_{34}) &> 0 \end{aligned}$$

# The tightened bounds

- The smallest value $u'_{34}$ obtained from the first three inequalities is compared to the given $u_{34}$ value. We set $\overline{u}_{34} = \min\{u_{34}, u'_{34}\}$.

- The largest value $l'_{34}$ obtained from the last four inequalities is compared to the given $l_{34}$ value. We set $\bar{l}_{34} = \max\{l_{34}, l'_{34}\}$.

# Solving a CM inequality

- Consider the inequality $CM(\sqrt{a}, \sqrt{b}, \sqrt{c}, \sqrt{d}, \sqrt{e}, \sqrt{x}) > 0$ where $x$ is the only unknown.

- Solving for $x$ involves finding the roots of the quadratic equation $Ax^2 + Bx + C = 0$ where $A$, $B$, $C$ can be computed in 19 additions and 9 multiplications as:

$$
\begin{aligned}
A &= (-a) \\
B &= (d-b)(c-e) + a(b+c+d+e-a) \\
C &= be(a-b+c+d-e) + cd(a+b-c-d+e) - a(ce+bd)
\end{aligned}
$$

# The Algorithm

TetrangleBoundSmoothing $(L, U, TOL)$

Initialize using TriangleBoundSmoothing

**repeat**

**for** each of $\binom{n}{4}$ quadruples of atoms $(i, j, k, m)$ **do**

**for** each of $d_{ij}$, $d_{ik}$, $d_{im}$, $d_{jk}$, $d_{jm}$, $d_{km}$ **do**

Calculate the $\overline{l}$, $\overline{u}$ values

**end for**

**end for**

**until** The largest change in any bound is $< TOL$

**return** $\overline{L}$ and $\overline{U}$

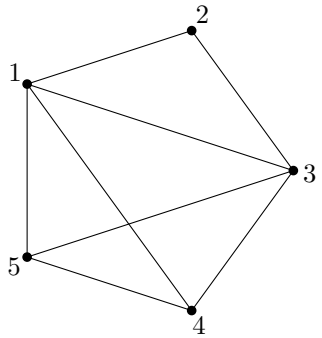
# Experimental results

- The order in which quadruples are considered does not affect the quality of the produced bounds.

- The same holds for the order that the distances are tightened

- The number of passes, until the algorithm converges, occasionally varies with respect to the considered example. Unfortunately the procedure may get caught in a cycle and may progress very slowly towards the final result.

# An example of cycling




- Five sides: $u = l = 1$
- Three diagonals: $l = 1.617$
- True diagonal length: $1.618 = 2\cos 36^\circ$

- After TriangleBoundSmoothing an upper bound of 2 is obtained for each of the 5 diagonals A new value of $u_{24}$ is obtained by tightening the bounds in the $(2, 3, 4, 5)$ quadruple. $u_{24}$ is then used in the $(2, 4, 5, 1)$ quadruple in order to obtain a new value on $u_{25}$.

- After 30 passes, with tolerance $10^{-14}$ the values of $u_{24}$ and $u_{25}$ are 1.6207323507579925 and 1.6207323507579441

# Graph Embedding

- $G(V, E, w)$ is an incomplete undirected graph, with vertex set $V$, edge set $E$ and for every $e = (v_i, v_j) \in E$ there is a non negative weight $w(e)$.

- The Graph Embedding problem asks for a mapping $\phi : V \rightarrow \mathbb{R}^k$ such that the Euclidean Distance $\|\phi(v_i) - \phi(v_j)\|$ is equal to $w(e)$.

- Graph Embedding is NP-hard even when $k = 1$ and all the edge weights are restricted to either 1 or 2 (SAXE, 1979).

- For practical purposes, $k$ is either 2 or 3.

# Software Packages

- The `EMBED` package (Crippen and Havel).
- The `ABBIE` package (Hendrickson).
- The `DGSOL` package (Moré and Wu).

# Conclusion – Open Problems

- The only bound smoothing method efficient enough for large molecules is based on the triangle inequality but empirical study shows that tetrangle inequalities produce significantly tighter bounds.

- Given is a possibly incomplete graph with a real positive interval assigned to each edge and a tolerance $e > 0$. Tighten the intervals in a way that any interval is smaller than $e$ and the following property is respected: If we can select values inside the given intervals that make the graph $d$-embeddable then the same holds for the tightened intervals.

# Conclusion – Open Problems (cont.)

- Is the above an NP-hard problem? Is it approximable within a function of $e$?

- What about 3-dimensional embeddings? Consider using tetrangle, pentagle or higher order inequalities and find any approximation algorithm.

- Any other algorithm for any $d$?

# Bibliography


- L.M. Blumenthal, *Theory and Applications of Distance Geometry*, Oxford, Clarendon Press, 1953.

- G.M. Cripen and T.F. Havel, *Distance Geometry and Molecular Conformations*, John Wiley & Sons, 1988.

- B.A. Hendrickson, *The Molecular Problem: Determining Conformation from Pairwise Distances*, Ph.D. thesis, Cornell University, 1991.

- A.I. Barvinok, *Problems of distance geometry and convex properties of quadratic maps*. Discrete and Computational Geometry, 13:189–202, 1995

- J.B. Saxe, *Embeddability of Weighted Graphs in k-space is Strongly NP-hard*, in proc. 17th Allerton Conference in Communications, Control and Computing, pp. 480–489, 1979.