Molecular Reconstruction via Douglas–Rachford

Matthew K. Tam

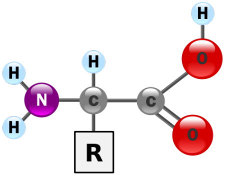
Joint work with Dr Fran Aragón and Laur. Prof Jon Borwein

School of Mathematical and Physical Sciences
University of Newcastle, Australia

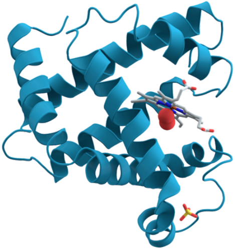
CARMA Retreat, 17th August 2013

Introduction and Preliminaries

Proteins are large biomolecules comprising of multiple amino acid chains.



Generic amino acid



Myoglobin

Proteins perform a vast range of functions and participate in virtually every cellular process!

Introduction and Preliminaries

If the structure of a protein is known, it can be used to predict how it performs its functions. Using NMR spectroscopy, the Nuclear Overhauser effect can be used to determine a subset of the interatomic distances (i.e. less than 6Å).

We say $D = (d_{ij}) \in \mathbb{R}^{n \times n}$ is a Euclidean distance matrix (EDM) if there exists points $p_1, \dots, p_n \in \mathbb{R}^r$ such that

$$d_{ij} = \|p_i - p_j\|^2.$$

If this holds for a set of points in $\mathbb{R}^r$ then D is said to be embeddable in $\mathbb{R}^r$. If D is embeddable in $\mathbb{R}^r$, but not in $\mathbb{R}^{r-1}$, then D is said to be irreducibly embeddable in $\mathbb{R}^r$.

We formulate protein reconstruction as a matrix completion problem:

Find a matrix having certain properties of interest,
knowing only a subset of its entries.

Feasibility formulation

Let D denote the partial EDM, and $\Omega \subset \mathbb{N} \times \mathbb{N}$ the set of indices for known entries. We have the following constraints:

$$C_1 := \{X \in \mathbb{R}^{n\times n} | X_{ii} = 0, X_{ij} \geq 0, X_{ij} = X_{ji} = D_{ij} \text{ for all } (i,j) \in \Omega\},$$
$$C_2 := \{X \in \mathbb{R}^{n\times n} | X \text{ is embeddable in } \mathbb{R}^3\}.$$

The reconstructed EDM is the solution to the feasibility problem

$$\text{Find } X \in C_1 \cap C_2.$$

Now,

- C_1 is a convex set (intersection of cone and affine subspace).
- C_2 is convex iff $n \leq 2$ (in which case $C_2 = \mathbb{R}^{n\times n}$).

For interesting problems, C_2 is **never convex**.

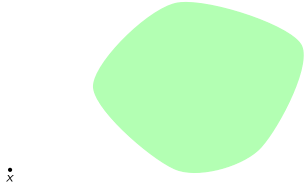
A Variational Toolkit

Let $S \subseteq \mathcal{H}$. The (nearest point) projection onto S is the (set-valued) mapping,

$$P_S x := \operatorname*{argmin}_{s \in S} \|s - x\|.$$

The reflection w.r.t. S is the (set-valued) mapping,

$$R_S := 2P_S - I.$$



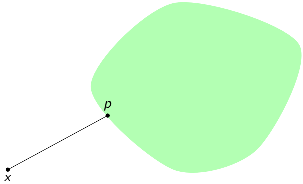
A Variational Toolkit

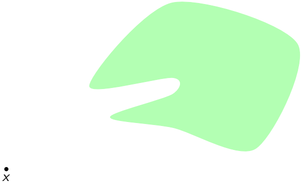
Let $S \subseteq \mathcal{H}$. The (nearest point) projection onto S is the (set-valued) mapping,

$$P_S x := \operatorname*{argmin}_{s \in S} \|s - x\|.$$

The reflection w.r.t. S is the (set-valued) mapping,

$$R_S := 2P_S - I.$$





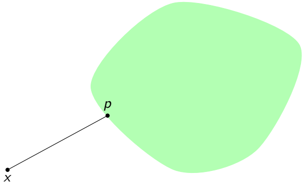
A Variational Toolkit

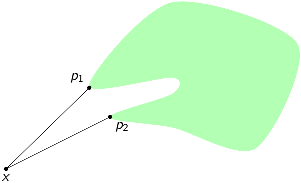
Let $S \subseteq \mathcal{H}$. The (nearest point) projection onto S is the (set-valued) mapping,

$$P_S x := \operatorname*{argmin}_{s \in S} \|s - x\|.$$

The reflection w.r.t. S is the (set-valued) mapping,

$$R_S := 2P_S - I.$$





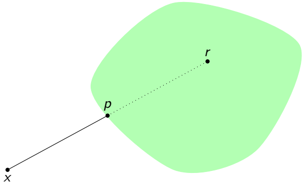
A Variational Toolkit

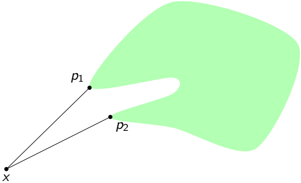
Let $S \subseteq \mathcal{H}$. The (nearest point) projection onto S is the (set-valued) mapping,

$$P_S x := \operatorname*{argmin}_{s \in S} \|s - x\|.$$

The reflection w.r.t. S is the (set-valued) mapping,

$$R_S := 2P_S - I.$$





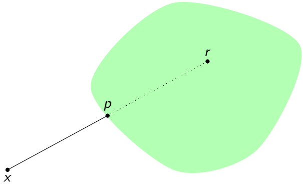
A Variational Toolkit

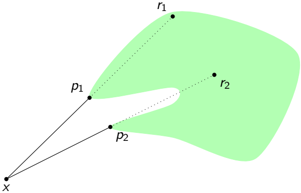
Let $S \subseteq \mathcal{H}$. The (nearest point) projection onto S is the (set-valued) mapping,

$$P_S x := \operatorname*{argmin}_{s \in S} \|s - x\|.$$

The reflection w.r.t. S is the (set-valued) mapping,

$$R_S := 2P_S - I.$$





Computing Projections and Reflections

The projection onto C_1 is given (point-wise) by

$$P_{C_1}(X)_{ij} = \begin{cases} D_{ij} & \text{if } (i,j) \in \Omega, \\ X_{ij} & \text{otherwise.} \end{cases}$$

Theorem (Hayden–Wells)

Let Q be the Householder matrix defined by

$$Q := I - \frac{2vv^T}{v^Tv}, \text{ where } v = \left[1, 1, \dots, 1, 1 + \sqrt{n}\right]^T \in \mathbb{R}^n.$$

Then a distance matrix, X, is a EDM iff the $(n-1) \times (n-1)$ block, $\widehat{X}$, in

$$Q(-X)Q = \begin{bmatrix} \widehat{X} & d \\ d^T & \delta \end{bmatrix}$$

is positive semidefinite. In this case, X is irreducibly embeddable in $\mathbb{R}^r$ where $r = \operatorname{rank}(\widehat{X}) \leq n - 1$.

Computing Projections and Reflections

The projection onto C_1 is given (point-wise) by

$$P_{C_1}(X)_{ij} = \begin{cases} D_{ij} & \text{if } (i,j) \in \Omega, \\ X_{ij} & \text{otherwise.} \end{cases}$$

A projection onto C_2 is given by

$$P_{C_2}(X) = -Q \left[\begin{array}{cc} U\Lambda_+ U^T & d \\ d^T & \delta \end{array} \right] Q,$$

where $X = U\Lambda U^T$ is a spectral decomposition with

$$\begin{aligned} \Lambda \; &:= \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_{n-1}) \quad \text{for } \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{n-1}, \\ \Lambda_+ &:= \operatorname{diag}(0, \dots, 0, \max\{0, \lambda_{n-3}\}, \max\{0, \lambda_{n-2}\}, \max\{0, \lambda_{n-1}\}). \end{aligned}$$

Recall that a spectral decomposition of real symmetric matrix, A, is given by

$$A = U\Lambda U^T$$

where U is an orthogonal matrix, and Λ a diagonal matrix whose entries are eigenvalues of A.

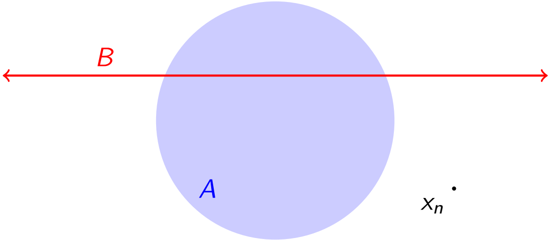
The Douglas–Rachford Algorithm

Theorem (Douglas–Rachford, Lions–Mercier)

Suppose $C_1, C_2 \subseteq \mathcal{H}$ are closed and convex with $C_1 \cap C_2 \neq \emptyset$. For any $x_0 \in \mathcal{H}$ define

$$x_{n+1} := Tx_n \text{ where } T := \frac{I + R_{C_2}R_{C_1}}{2}.$$

Then (x_n) converges (weakly) to a point x such that $P_{C_1}x \in C_1 \cap C_2$.



$$A = \{x \in \mathcal{H} : \|x\| \leq 1\}, \quad B = \{x \in \mathcal{H} : \langle a, x\rangle = b\}.$$

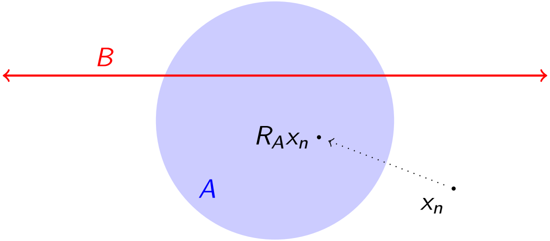
The Douglas–Rachford Algorithm

Theorem (Douglas–Rachford, Lions–Mercier)

Suppose $C_1, C_2 \subseteq \mathcal{H}$ are closed and convex with $C_1 \cap C_2 \neq \emptyset$. For any $x_0 \in \mathcal{H}$ define

$$x_{n+1} := Tx_n \text{ where } T := \frac{I + R_{C_2} R_{C_1}}{2}.$$

Then (x_n) converges (weakly) to a point x such that $P_{C_1} x \in C_1 \cap C_2$.



$$A = \{x \in \mathcal{H} : \|x\| \leq 1\}, \quad B = \{x \in \mathcal{H} : \langle a, x \rangle = b\}.$$

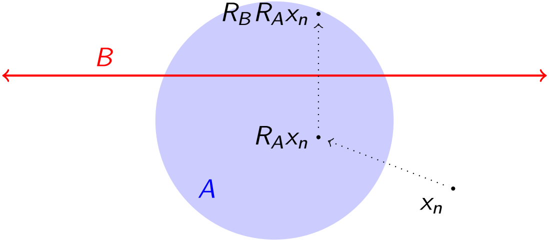
The Douglas–Rachford Algorithm

Theorem (Douglas–Rachford, Lions–Mercier)

Suppose $C_1, C_2 \subseteq \mathcal{H}$ are closed and convex with $C_1 \cap C_2 \neq \emptyset$. For any $x_0 \in \mathcal{H}$ define

$$x_{n+1} := Tx_n \text{ where } T := \frac{I + R_{C_2}R_{C_1}}{2}.$$

Then (x_n) converges (weakly) to a point x such that $P_{C_1}x \in C_1 \cap C_2$.



$$A = \{x \in \mathcal{H} : \|x\| \leq 1\}, \quad B = \{x \in \mathcal{H} : \langle a, x\rangle = b\}.$$

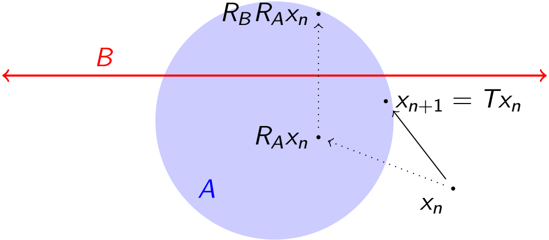
The Douglas–Rachford Algorithm

Theorem (Douglas–Rachford, Lions–Mercier)

Suppose $C_1, C_2 \subseteq \mathcal{H}$ are closed and convex with $C_1 \cap C_2 \neq \emptyset$. For any $x_0 \in \mathcal{H}$ define

$$x_{n+1} := Tx_n \text{ where } T := \frac{I + R_{C_2} R_{C_1}}{2}.$$

Then (x_n) converges (weakly) to a point x such that $P_{C_1} x \in C_1 \cap C_2$.



$$A = \{x \in \mathcal{H} : \|x\| \leq 1\}, \quad B = \{x \in \mathcal{H} : \langle a, x \rangle = b\}.$$

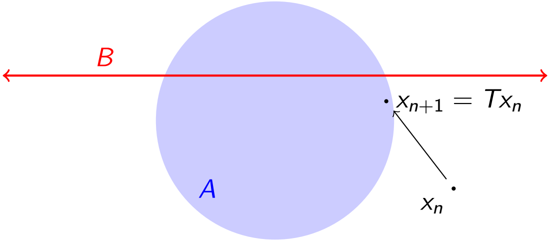
The Douglas–Rachford Algorithm

Theorem (Douglas–Rachford, Lions–Mercier)

Suppose $C_1, C_2 \subseteq \mathcal{H}$ are closed and convex with $C_1 \cap C_2 \neq \emptyset$. For any $x_0 \in \mathcal{H}$ define

$$x_{n+1} := Tx_n \text{ where } T := \frac{I + R_{C_2} R_{C_1}}{2}.$$

Then (x_n) converges (weakly) to a point x such that $P_{C_1} x \in C_1 \cap C_2$.



$$A = \{x \in \mathcal{H} : \|x\| \leq 1\}, \quad B = \{x \in \mathcal{H} : \langle a, x \rangle = b\}.$$

Results: Six Proteins

Interatomic distances below 6Å typically constitute less than 8% of the total nonzero entries of the distance matrix.

Table 1. Six Proteins: average (maximum) errors from five replications.

Protein	# Atoms	Rel. Error (dB)	RMSE	Max Error
1PTQ	404	-83.6 (-83.7)	0.0200 (0.0219)	0.0802 (0.0923)
1HOE	581	-72.7 (-69.3)	0.191 (0.257)	2.88 (5.49)
1LFB	641	-47.6 (-45.3)	3.24 (3.53)	21.7 (24.0)
1PHT	988	-60.5 (-58.1)	1.03 (1.18)	12.7 (13.8)
1POA	1067	-49.3 (-48.1)	34.1 (34.3)	81.9 (87.6)
1AX8	1074	-46.7 (-43.5)	9.69 (10.36)	58.6 (62.6)

$$\text{Rel. error} := 10\log_{10}\left(\frac{\|P_{C_2}P_{C_1}X_N - P_{C_1}X_N\|^2}{\|P_{C_1}X_N\|^2}\right),$$

$$\text{RMSE} := \sqrt{\frac{\sum_{i=1}^{m}\|\hat{p}_i - p_i^{true}\|_2^2}{\#\text{ of atoms}}}, \qquad \text{Max} := \max_{1\le i\le m}\|\hat{p}_i - p_i^{true}\|_2.$$

The points $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$ denote the best fitting of $p_1, p_2, \ldots, p_n$ if rotation, translation and reflections are allowed.

Results: Six Proteins

Interatomic distances below 6Å typically constitute less than 8% of the total nonzero entries of the distance matrix.

Table 1. Six Proteins: average (maximum) errors from five replications.

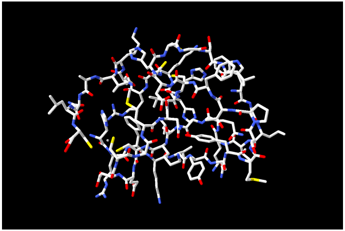
Protein	# Atoms	Rel. Error (dB)	RMSE	Max Error
1PTQ	404	-83.6 (-83.7)	0.0200 (0.0219)	0.0802 (0.0923)
1HOE	581	-72.7 (-69.3)	0.191 (0.257)	2.88 (5.49)
1LFB	641	-47.6 (-45.3)	3.24 (3.53)	21.7 (24.0)
1PHT	988	-60.5 (-58.1)	1.03 (1.18)	12.7 (13.8)
1POA	1067	-49.3 (-48.1)	34.1 (34.3)	81.9 (87.6)
1AX8	1074	-46.7 (-43.5)	9.69 (10.36)	58.6 (62.6)

$$\text{Rel. error} := 10\log_{10}\left(\frac{\|P_{C_2}P_{C_1}X_N - P_{C_1}X_N\|^2}{\|P_{C_1}X_N\|^2}\right),$$

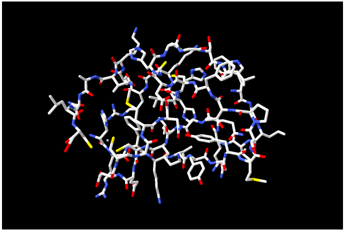
$$\text{RMSE} := \sqrt{\frac{\sum_{i=1}^{m}\|\hat{p}_i - p_i^{true}\|_2^2}{\#\text{ of atoms}}}, \qquad \text{Max} := \max_{1\le i\le m}\|\hat{p}_i - p_i^{true}\|_2.$$

The points $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$ denote the best fitting of $p_1, p_2, \ldots, p_n$ if rotation, translation and reflections are allowed.

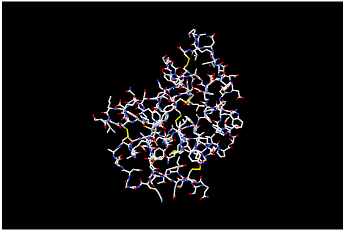
What do the reconstructions look like?



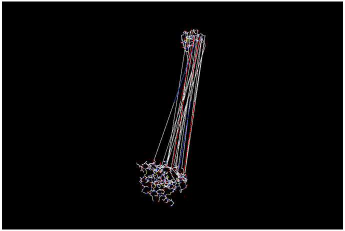
1PTQ (actual)



5,000 steps, -83.6dB

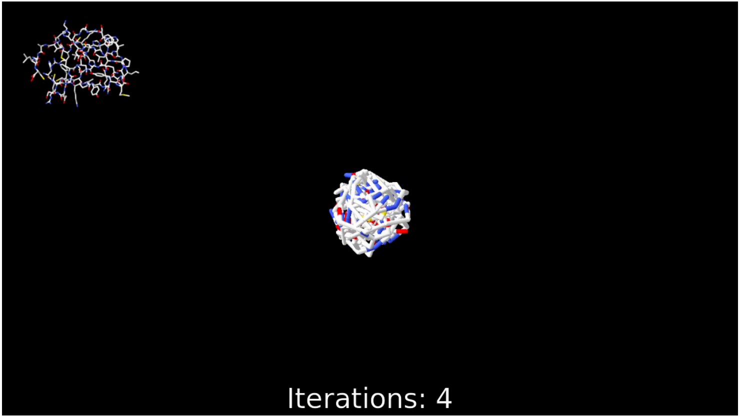


1POA (actual)



5,000 steps, -49.3dB

What do reconstructions look like?

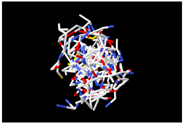


First 3,000 steps of the 1PTQ reconstruction

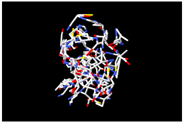
What do reconstructions look like?

There are many projection methods, so why Douglas-Rachford?

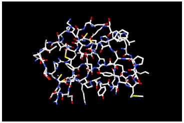
Douglas–Rachford reconstruction:



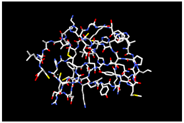
500 steps, -25 dB.



1,000 steps, -30 dB.

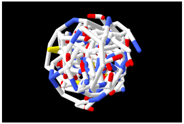


2,000 steps, -51 dB.

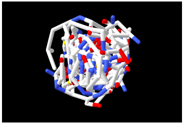


5,000 steps, -84 dB.

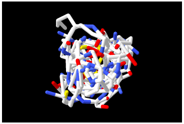
Alternating projections reconstruction:



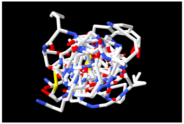
500 steps, -22 dB.



1,000 steps, -24 dB.



2,000 steps, -25 dB.



5,000 steps, -28 dB.

Concluding Remarks and Future Work

- We presented with a feasibility problem, it is well worth see if Douglas–Rachford can deal with it – it is conceptually simple and easy to implement.
- More efficient implementation (including computation of P_{C_2}).
- Refine the method applied to large molecules.
  - Reasonable upper bounds from bond lengths.
  - Splitting approach.
- Other non-convex applications
  - Hadamard matrices, Sudoku, Nonograms, ILs.
- Extensions to non-convex convergence theory *á la* Aragón–Borwein–Sims, Hesse–Luke?
- Can these unjustifiably good results be explained in CAT(0) spaces?

Douglas–Rachford feasibility methods for matrix completion problems
with F.J. Aragón Artacho & J.M. Borwein. *Soon to be submitted*, 2013.

Many resources can be found at the companion website:

http://carma.newcastle.edu.au/DRmethods/