Efficient use of semidefinite programming for the selection of rotamers in protein conformation

Forbes Burkowski, **Yuen-Lam Cheung** &
Henry Wolkowicz

Retrospective Workshop on
Discrete Geometry, Optimization and Symmetry

November 2013

Outline

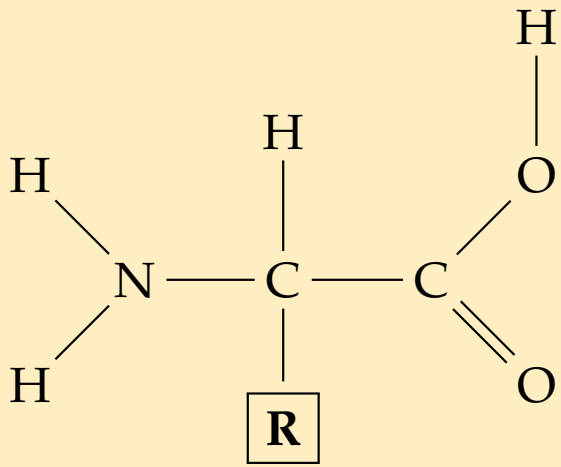
- Primer on protein conformation
- **Side chain positioning**: IP formulation
- SDP relaxation and minimal face
- Implementation: a cutting plane technique
- Quality measurement for integral solutions, and numerics

Protein conformation: a primer

Basics about proteins

An amino acid has five components:

- alpha carbon
- hydrogen atom
- carboxyl group
- amino group
- side chain



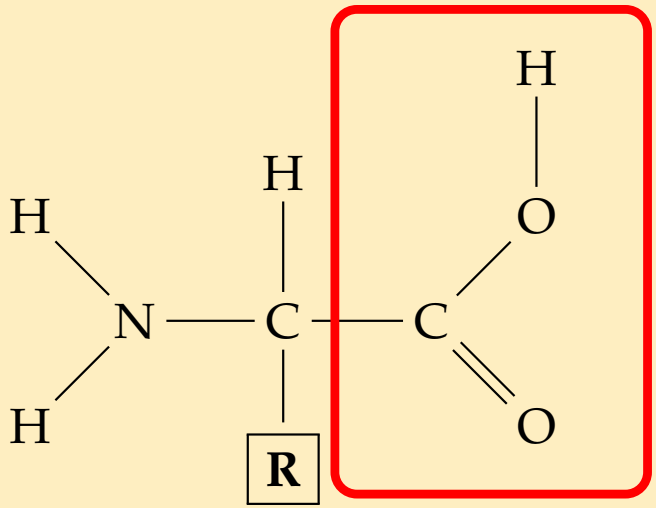
A protein is a polymer formed from a chain of amino acids (with different side chains).

Protein conformation: a primer

Basics about proteins

An amino acid has five components:

- alpha carbon
- hydrogen atom
- carboxyl group
- amino group
- side chain



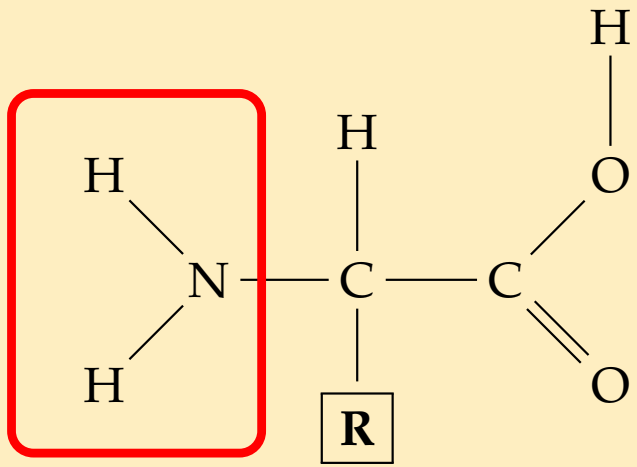
A protein is a polymer formed from a chain of amino acids (with different side chains).

Protein conformation: a primer

Basics about proteins

An amino acid has five components:

- alpha carbon
- hydrogen atom
- carboxyl group
- amino group
- side chain



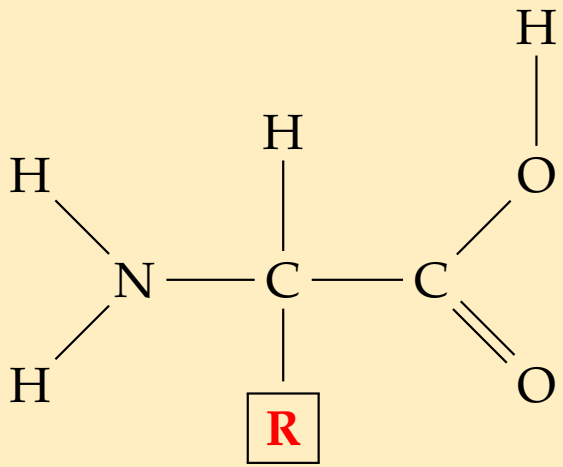
A protein is a polymer formed from a chain of amino acids (with different side chains).

Protein conformation: a primer

Basics about proteins

An amino acid has five components:

- alpha carbon
- hydrogen atom
- carboxyl group
- amino group
- side chain

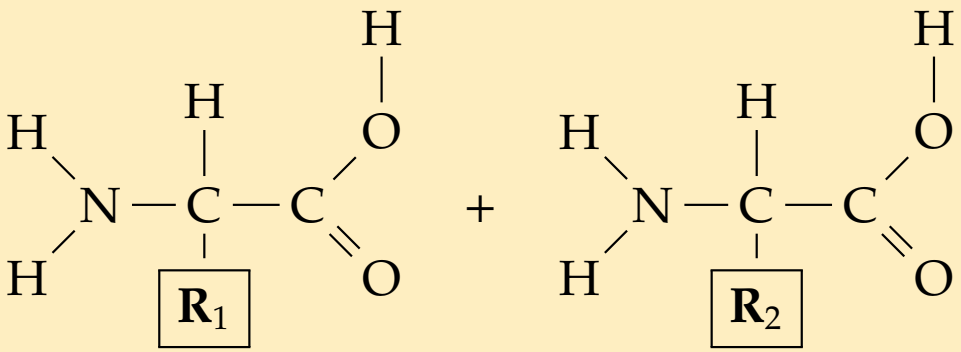


A protein is a polymer formed from a chain of amino acids (with different side chains).

Protein conformation: a primer

Forming a protein through condensation

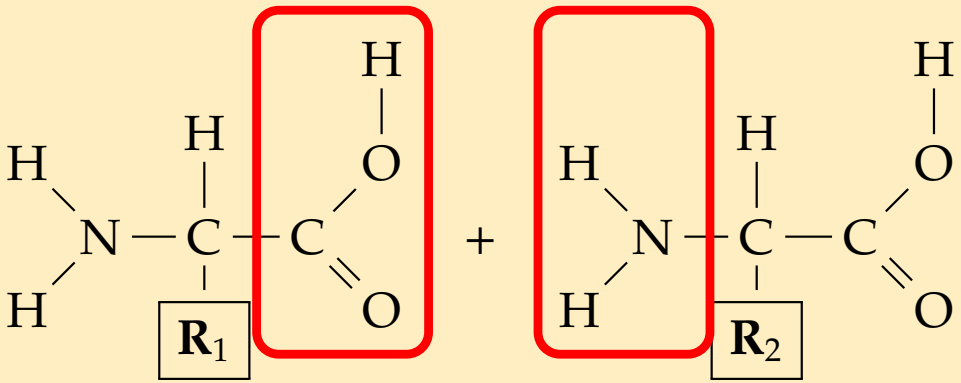
A protein is a polymer formed from a chain of amino acids, bonded via a condensation process:



Protein conformation: a primer

Forming a protein through condensation

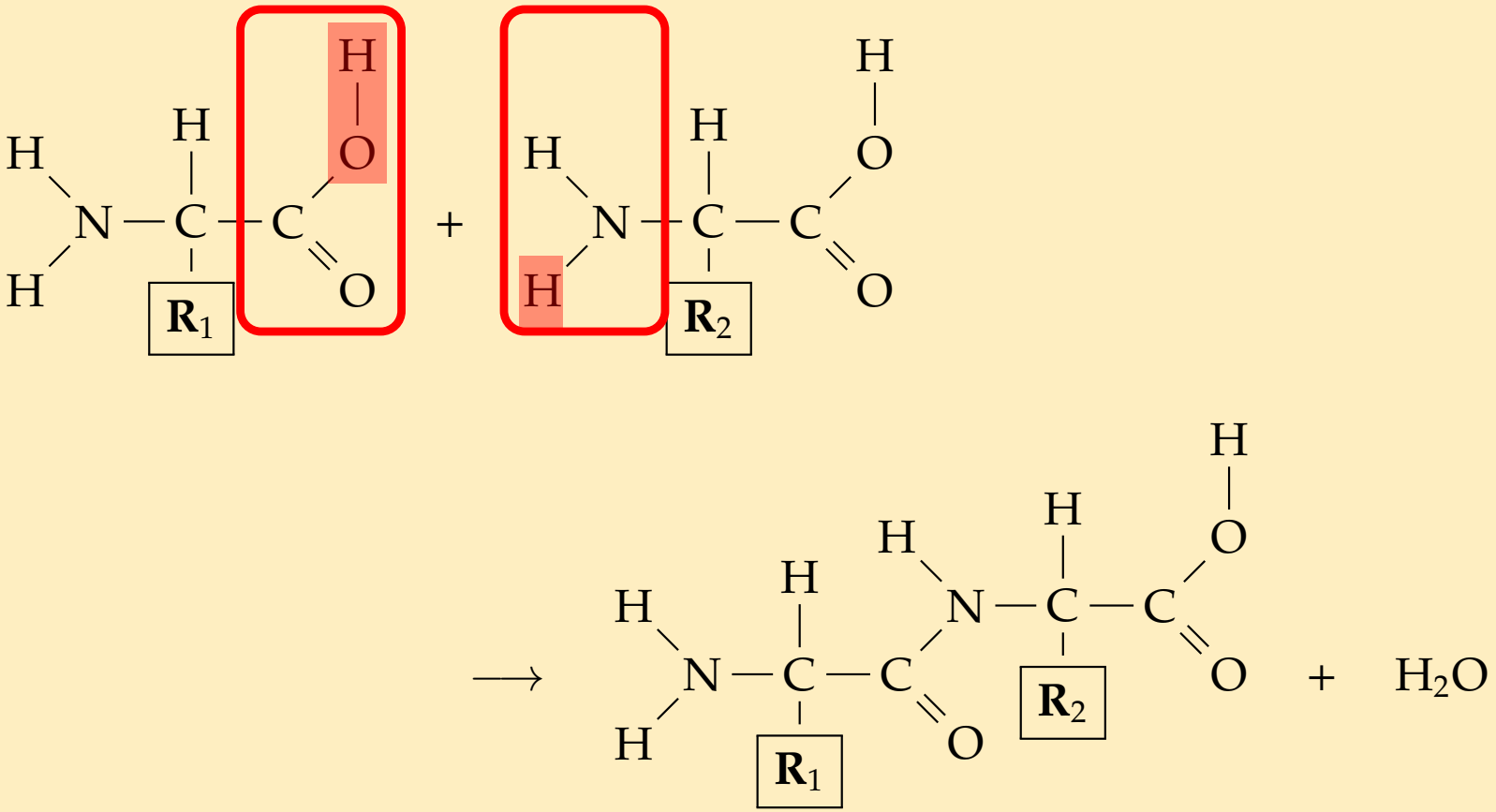
A protein is a polymer formed from a chain of amino acids, bonded via a condensation process:



Protein conformation: a primer

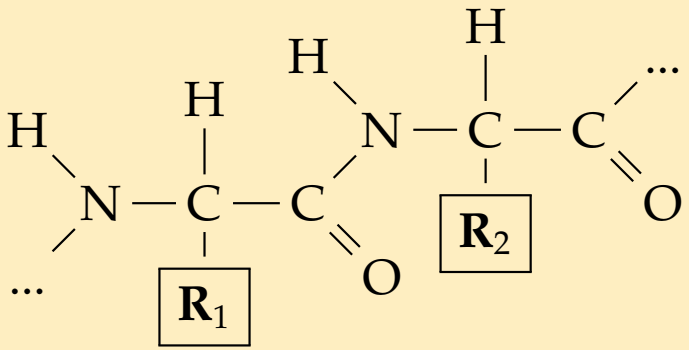
Forming a protein through condensation

A protein is a polymer formed from a chain of amino acids, bonded via a condensation process:



Protein conformation: a primer

Backbone and the side chain positioning



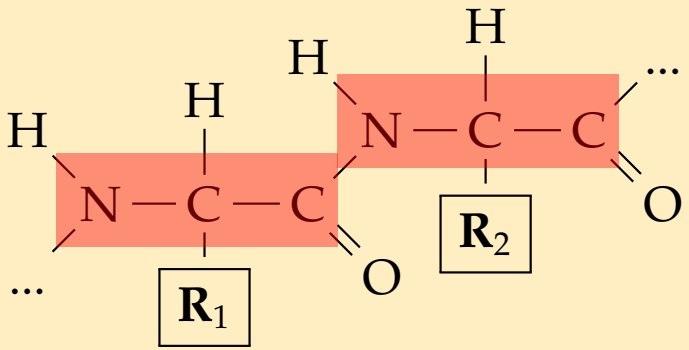
Protein conformation problem:

Given a 2D chain of residues of a protein,
find the 3D positions of all the atoms so that

- the bond lengths and bond angles are respected, and
- the total energy of the resultant protein conformation is at global minimum.

Protein conformation: a primer

Backbone and the side chain positioning



Side chain positioning problem, a subproblem for protein conformation:

- Suppose we know the positions of the *backbone* atoms.
  Find the 3D positions of the atoms in the side chains so that

  the total energy of the resultant conformation is at global min.

- Further assumption:
  each of the side chains can take one of finitely many possible positions, a.k.a. *rotamers*.

Side chain positioning problem

Setup

Given a weighted complete p-partite graph with vertex set

$$\mathcal{V} = \bigcup_{k=1}^{p} \mathcal{V}_k, \quad \text{where } \mathcal{V}_1 = 1 : m_1, \quad \mathcal{V}_k = \left(1 + \sum_{l=1}^{k-1} m_l\right) : \sum_{l=1}^{k} m_l, \ \forall\, k = 2, \ldots, p,$$

(and $m \in \mathbb{Z}^p$ is a positive vector), with

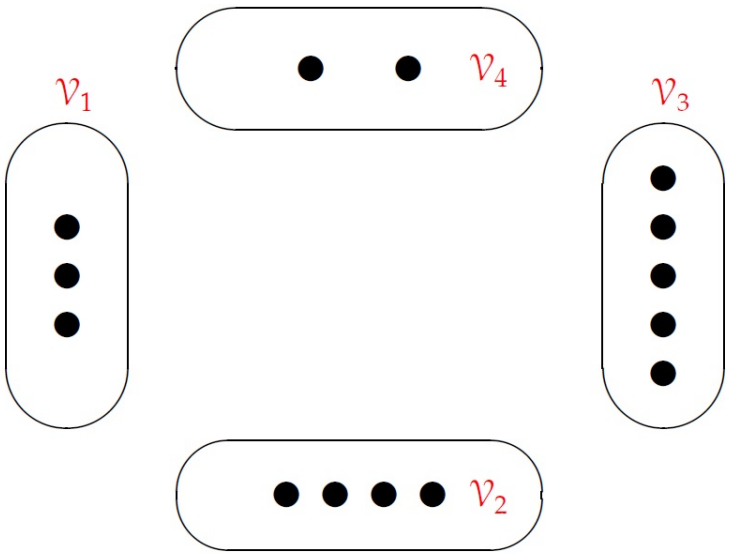
$$\text{edge weight } E_{ij} = E_{ji}, \quad \forall\, \{i, j\} \in (1 : n_0) \times (1 : n_0), \text{ where } n_0 = \sum_{k=1}^{p} m_j.$$

Side chain positioning

Statement of the sidechain positioning problem

Pick exactly one vertex from each partition $\mathcal{V}_k$ $(\forall\, k = 1, 2, \ldots, p)$ s.t.

the total edge weight of the induced subgraph is minimized.

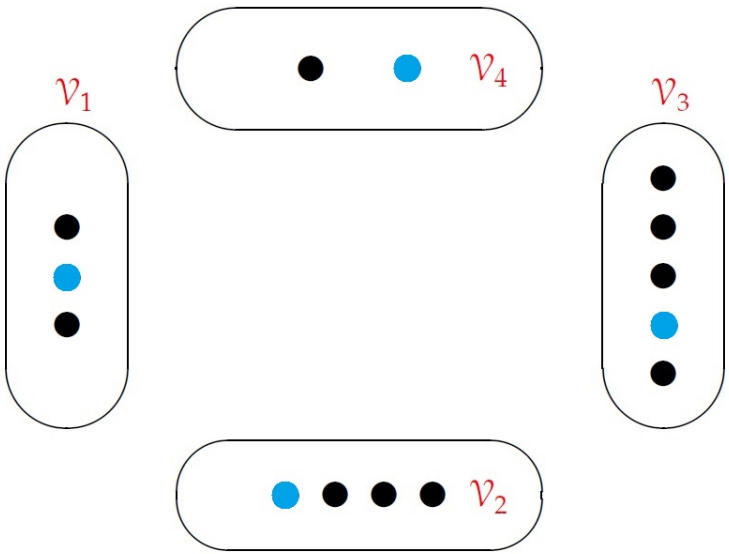


Side chain positioning

Statement of the sidechain positioning problem

Pick exactly one vertex from each partition $\mathcal{V}_k$ $(\forall k = 1, 2, \ldots, p)$ s.t.

the total edge weight of the induced subgraph is minimized.

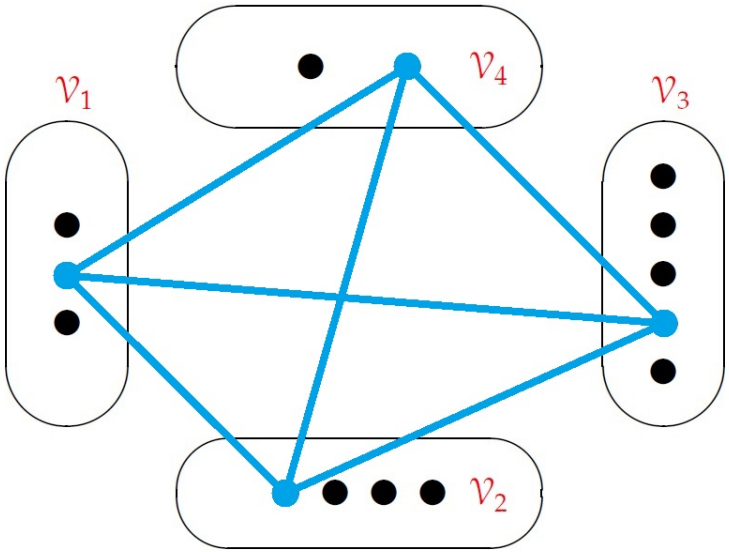


Side chain positioning

Statement of the sidechain positioning problem

Pick exactly one vertex from each partition $\mathcal{V}_k$ $(\forall k = 1, 2, \ldots, p)$ s.t.

the total edge weight of the induced subgraph is minimized.



Side chain positioning problem

Statement of the sidechain positioning problem

Pick exactly one vertex from each partition $\mathcal{V}_k$ ($\forall\, k = 1, 2, \ldots, p$) s.t.

the total edge weight of the induced subgraph is minimized.

Complexity of sidechain positioning problem

- NP-hard [Akutsu, 1997; Pierce and Winfree, 2002]
- Special cases of the sidechain positioning problem:
  - MAX 3-SAT [Chazelle *et al.*, 2004]
    $\Longrightarrow$ side chain positioning problem is "inapproximable"
  - maximum k-cut problem

Side chain positioning problem: IP formulation

Statement of the sidechain positioning problem

Pick exactly one vertex from each partition $\mathcal{V}_k$ ($\forall\, k = 1, 2, \ldots, p$) s.t.

the total edge weight of the induced subgraph is minimized.

Integer quadratic programming formulation

$$
\begin{aligned}
v_{\mathrm{SCP}} = \quad & \min_{x} \quad x^\top E x \\
& \text{s.t.} \qquad x \quad = \left[v^{(1)} ;\, v^{(2)} ;\, \cdots ;\, v^{(p)} \right] \in \{0,1\}^{n_0}, \\
& \qquad\qquad \bar{e}^\top v^{(k)} = 1, \quad \forall\, k = 1, \ldots, p.
\end{aligned}
$$

$x \in \mathbb{R}^{n_0}$ is an incident vector for the choices of vertices in each partition.

Side chain positioning problem: IP formulation

Statement of the sidechain positioning problem

Pick exactly one vertex from each partition $\mathcal{V}_k$ $(\forall\, k = 1, 2, \ldots, p)$ s.t.

the total edge weight of the induced subgraph is minimized.

Integer quadratic programming formulation

$$\begin{aligned} v_{\text{SCP}} = \quad & \min_{x} \quad x^\top E x \\ & \text{s.t.} \quad Ax = \bar{e} \ \in \mathbb{R}^p, \\ & \qquad\quad x \in \{0,1\}^{n_0}, \end{aligned}$$

where

$$A = \begin{array}{c} \\ 1 \\ 1 \\ \\ 1 \end{array} \begin{array}{c} \begin{array}{cccc} m_1 & m_2 & & m_p \end{array} \\ \begin{bmatrix} \bar{e}^\top & 0 & \cdots & 0 \\ 0 & \bar{e}^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{e}^\top \end{bmatrix} \end{array} \in \mathbb{R}^{p \times n_0}.$$

Side chain positioning problem: IP formulation

$$
\begin{aligned}
v_{\mathrm{SCP}} = \quad & \min_{x} \quad && x^\top E x \\
& \text{s.t.} \quad && Ax = \bar{e} \ \in \mathbb{R}^p, \\
& && x \in \{0,1\}^{n_0}\,.
\end{aligned}
\tag{SCP}
$$

Valid constraints on x and $X := xx^\top$

- nonegativity, i.e., $X \geqslant 0$;
- all the diagonal blocks of X are diagonal, i.e., $(A^\top A - I) \circ X = 0$;
- the "arrow" constraint, i.e., $\mathrm{diag}(X) = x$;
- $\|Ax - \bar{e}\|_2^2 = 0$, i.e., $\langle A^\top A, X\rangle - 2\bar{e}^\top x + p = 0$.

Indeed any $x \in \mathbb{R}^{n_0}$ together with $X = xx^\top$ satisfying the 3rd-4th constraints is feasible for (SCP).

Side chain positioning problem: IP formulation

Valid constraints on x and $X := xx^\top$

- nonegativity, i.e., $X \geqslant 0$;
- all the diagonal blocks of X are diagonal, i.e., $(A^\top A - I) \circ X = 0$;
- the "arrow" constraint, i.e., $\mathrm{diag}(X) = x$;
- $\|Ax - \bar{e}\|_2^2 = 0$, i.e., $\langle A^\top A, X\rangle - 2\bar{e}^\top x + p = 0$.

Indeed any $x \in \mathbb{R}^{n_0}$ together with $X = xx^\top$ satisfying the 3rd-4th constraints is feasible for (SCP).

Equivalent formulation of (SCP)

$$
\begin{array}{rll}
v_{\mathrm{SCP}} = & \min\limits_{x,X} & \langle E, X\rangle \\
& \text{s.t.} & (A^\top A - I) \circ X = 0, \\
& & \langle A^\top A, X\rangle - 2\bar{e}^\top x + p = 0, \\
& & \mathrm{diag}(X) = x, \\
& & X = xx^\top.
\end{array}
$$

Side chain positioning problem: SDP relaxation

SDP relaxation of (SCP)

$$
\begin{aligned}
v_{\mathrm{SCP}} \geqslant v_{\mathrm{SCP(SDP)}} := \quad & \min_{x,X} && \langle E, X\rangle \\
& \text{s.t.} && \operatorname{diag}(X) = x, \\
& && \langle A^\top A, X\rangle - 2\bar{e}^\top x + p = 0, \\
& && (A^\top A - I) \circ X = 0, \\
& && X = xx^\top.
\end{aligned}
$$

Side chain positioning problem: SDP relaxation

SDP relaxation of (SCP)

$$
\begin{aligned}
v_{\mathrm{SCP}} \geqslant v_{\mathrm{SCP(SDP)}} := \quad & \min_{x,X} && \langle E, X\rangle \\
& \text{s.t.} && \operatorname{diag}(X) = x, \\
& && \langle A^\top A, X\rangle - 2\bar{e}^\top x + p = 0, \\
& && (A^\top A - I) \circ X = 0, \\
& && X \succeq xx^\top \ (\text{i.e., } X - xx^\top \in \mathbb{S}^n_+).
\end{aligned}
$$

Side chain positioning problem: SDP relaxation

SDP relaxation of (SCP)

$$
\begin{aligned}
v_{\mathrm{SCP}} \geqslant v_{\mathrm{SCP(SDP)}} := \quad & \min_{x,X} && \langle E, X\rangle \\
& \text{s.t.} && \operatorname{diag}(X) = x, \\
& && \langle A^\top A, X\rangle - 2\bar{e}^\top x + p = 0, \\
& && (A^\top A - I) \circ X = 0, \\
& && Y = \begin{bmatrix} 1 & x^\top \\ x & X \end{bmatrix} \succeq 0.
\end{aligned}
\tag{SCP-SDP}
$$

Side chain positioning problem: SDP relaxation

SDP relaxation of (SCP)

$$
\begin{aligned}
v_{\mathrm{SCP}} \geqslant v_{\mathrm{SCP(SDP)}} := \quad & \min_{x,X} && \langle E, X \rangle \\
& \text{s.t.} && \operatorname{diag}(X) = x, \\
& && \left\langle \begin{bmatrix} p & -\bar{e}^\top \\ -\bar{e} & A^\top A \end{bmatrix}, Y \right\rangle = 0, \\
& && (A^\top A - I) \circ X = 0, \\
& && Y = \begin{bmatrix} 1 & x^\top \\ x & X \end{bmatrix} \succeq 0.
\end{aligned}
\tag{SCP-SDP}
$$

Side chain positioning problem: SDP relaxation

SDP relaxation of (SCP)

$$
\begin{aligned}
v_{\mathrm{SCP}} \geqslant v_{\mathrm{SCP(SDP)}} := \quad \min_{x,X} \quad & \langle E, X \rangle \\
\text{s.t.} \quad & \operatorname{diag}(X) = x, \\
& \left\langle \begin{bmatrix} p & -\bar{e}^\top \\ -\bar{e} & A^\top A \end{bmatrix}, Y \right\rangle = 0, \qquad \text{(SCP-SDP)} \\
& (A^\top A - I) \circ X = 0, \\
& Y = \begin{bmatrix} 1 & x^\top \\ x & X \end{bmatrix} \succeq 0.
\end{aligned}
$$

Failure of the Slater condition

- But $\begin{bmatrix} p & -\bar{e}^\top \\ -\bar{e} & A^\top A \end{bmatrix} \succeq 0 \implies$ Slater condition fails for (SCP-SDP),

  i.e., (SCP-SDP) does not have a feas. solution Y that is positive definite

Side chain positioning problem: SDP relaxation

Failure of the Slater condition

- $\begin{bmatrix} p & -\bar{e}^\top \\ -\bar{e} & A^\top A \end{bmatrix} \succeq 0 \implies$ Slater condition fails for (SCP-SDP).
- Y feasible $\implies$ $Y = WXW^\top$ for some $X \in \mathbb{S}_+^{n_0-p+1}$ (W: full col. rank).
- $W\mathbb{S}_+^{n_0-p+1}W^\top$ is a proper face of $\mathbb{S}_+^n$.
- In fact, $W\mathbb{S}_+^{n_0-p+1}W^\top$ is the minimal face of (SCP-SDP).

Side chain positioning problem: SDP relaxation

Facial reduction

Using the minimal face $W\mathbb{S}_+^{n_0-p+1}W^\top$,

i.e., the substitution $Y = W\hat{Y}W^\top$,

$$
\begin{aligned}
v_{\mathrm{SCP(SDP)}} = \quad & \min_{x,X} && \langle E, X\rangle \\
& \text{s.t.} && \operatorname{diag}(X) = x, \\
& && (A^\top A - I) \circ X = 0, \\
& && \qquad \text{(i.e., the diagonal blocks are diagonal matrices)} \\
& && \left\langle \begin{bmatrix} p & -\bar{e}^\top \\ -\bar{e} & A^\top A \end{bmatrix}, Y \right\rangle = 0, \\
& && Y = \begin{bmatrix} 1 & x^\top \\ x & X \end{bmatrix} \succeq 0.
\end{aligned}
$$

Side chain positioning problem: SDP relaxation

Facial reduction

Using the minimal face $W\mathbb{S}_+^{n_0-p+1}W^\top$,

i.e., the substitution $Y = W\hat{Y}W^\top$,

$$
\begin{aligned}
v_{\mathrm{SCP(SDP)}} = \quad & \min_{x,X} && \langle E, X\rangle \\
& \text{s.t.} && \operatorname{diag}(X) = x, \\
& && (A^\top A - I) \circ X = 0, \\
& && \qquad \text{(i.e., the diagonal blocks are diagonal matrices)} \\
& && Y = W\hat{Y}W^\top \\
& && Y = \begin{bmatrix} 1 & x^\top \\ x & X \end{bmatrix} \succeq 0.
\end{aligned}
$$

Side chain positioning problem: SDP relaxation

Facial reduction

Using the minimal face $W\mathbb{S}_+^{n_0-p+1}W^\top$,
i.e., the substitution $Y = W\hat{Y}W^\top$,

$$
\begin{aligned}
v_{\mathrm{SCP(SDP)}} = \quad & \min_{\hat{x},\hat{X}} \quad \left\langle W^\top \begin{bmatrix} 0 & 0 \\ 0 & E \end{bmatrix} W, \begin{bmatrix} 1 & \hat{x}^\top \\ \hat{x} & \hat{X} \end{bmatrix} \right\rangle \\
& \text{s.t.} \quad \operatorname{diag}(\hat{X}) = \hat{x}, \\
& \qquad (A^\top A - I) \circ \hat{X} = 0, \\
& \qquad\qquad \text{(i.e., the diagonal blocks are diagonal matrices)} \\
& \qquad \begin{bmatrix} 1 & \hat{x}^\top \\ \hat{x} & \hat{X} \end{bmatrix} \in \mathbb{S}_+^{n_0-p+1}.
\end{aligned}
$$

$\Longrightarrow$ a "smaller" and equiv. SDP relaxation of (SCP)

Cutting plane technique

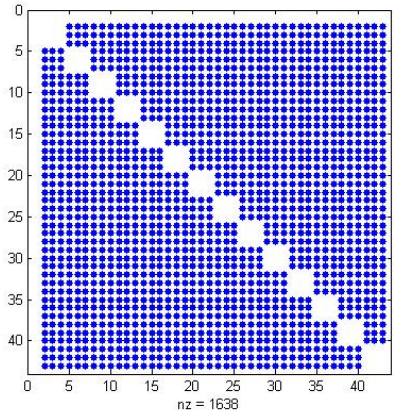
Nonnegativity constraint $Y \geqslant 0$

- Y being doubly nonnegative is a valid constraint.
- But the constraint

$$\begin{bmatrix} 0 & 0 \\ 0 & A^\top A - I \end{bmatrix} \circ Y = 0$$

and $Y \in \mathbb{S}^n_+$ implies that enforcing $Y \geqslant 0$ in the SDP relaxation necessarily lead to the failure of the Slater condition.

- Only need $\mathcal{B} :=$



- It is still too expensive to enforce $Y_{ij} \geqslant 0$ for all $(i,j) \in \mathcal{B}$.

Cutting plane technique

Cutting plane technique

Initial cuts: $\mathcal{I} \subseteq \mathcal{B}$; **repeat** the following:

- Solve

$$
\begin{aligned}
v_{\mathrm{SCP(SDP)}}(\mathcal{I}) = \quad & \min_{\hat{x},\hat{X}} \quad \left\langle W^\top \begin{bmatrix} 0 & 0 \\ 0 & E \end{bmatrix} W, \begin{bmatrix} 1 & \hat{x}^\top \\ \hat{x} & \hat{X} \end{bmatrix} \right\rangle \\
& \text{s.t.} \quad \operatorname{diag}(\hat{X}) = \hat{x}, \ (A^\top A - I) \circ \hat{X} = 0, \\
& \qquad \hat{Y} = \begin{bmatrix} 1 & \hat{x}^\top \\ \hat{x} & \hat{X} \end{bmatrix} \in \mathbb{S}_+^{n_0-p+1}, \\
& \qquad (W\hat{Y}W^\top)_{ij} \geqslant 0, \ \forall\, (i,j) \in \mathcal{I}.
\end{aligned}
\tag{SDP($\mathcal{I}$)}
$$

  for solution $\hat{Y}^*$.

- If $W\hat{Y}^*W^\top \succcurlyeq 0$ or if $\hat{Y}^*$ is "good enough", then stop; else, find a group $\mathcal{I}'$ of indices $(i,j) \in \mathcal{B}$ s.t.

$$(W\hat{Y}^*W^\top)_{ij} << 0, \quad E_{i-1,j-1} >> 0.$$

- Update: $\mathcal{I} \leftarrow \mathcal{I} \cup \mathcal{I}'$.

Existing rounding technique

The typical rounding techniques; also in Chazelle *et al.*, 2004

- Projection rounding: use the diagonal of $W\hat{Y}^*W^\top$;
- Perron Frobenius rounding: use the principal eigenvector of $W\hat{Y}^*W^\top$, which empirically is nonnegative.

The fractional vector u from either of the rounding method satisfies

$$\bar{e}^\top u^{(k)} = 1 \;\; \forall k \in 1:p, \quad u = [u^{(1)}\,;\, u^{(2)}\,;\, \ldots\,;\, u^{(p)}].$$

If $u \geqslant 0$, then we can use $u^{(1)}$, $u^{(2)}$, $\ldots$, $u^{(p)}$ as vectors of probability distributions:

$$v^{(k)} = e_j \in \mathbb{R}^{m_k} \;\; \text{with prob. } u_j^{(k)}, \; \forall j \in 1:m_k, \; k \in 1:p.$$

Quality measurement for integral solutions

Measuring the quality of integral solutions

- Let x be a feasible integral solution of (SCP).
- Bound: $t^* \leqslant v_{\mathrm{SCP(SDP)}} \leqslant v_{\mathrm{SCP}} \leqslant x^\top Ex$,
  where t^* is the opt. value of the dual of the SDP relaxation.
- Relative difference:

$$\frac{x^\top Ex - t^*}{\frac{1}{2}|x^\top Ex + t^*|}$$

Numerics

Results on small proteins

Protein	n_0	p	run time (sec)		relative diff	
			SCPCP	orig	SCPCP	orig
1AAC	117	85	6.58	296.06	5.75E-11	1.72E-05
1AHO	108	54	7.97	364.73	8.44E-11	4.95E-05
1BRF	130	45	14.96	977.08	3.92E-11	2.27E-05
1CC7	160	66	28.60	1059.06	1.13E-11	2.01
1CKU	115	60	5.46	815.18	7.17E-11	4.79E-05
1CRN	65	37	12.76	46.42	1.64E-12	3.05E-05
1CTJ	153	61	16.15	777.31	2.98E-11	2.00
1D4T	188	89	41.32	2775.34	3.88E-11	2.00
1IGD	82	50	5.51	189.04	4.79E-10	2.74E-06
1PLC	129	82	14.32	1766.03	1.28E-11	7.28E-04
1VFY	134	63	23.49	1765.36	1.67E-11	-1.11E-05
4RXN	98	48	18.44	366.48	1.48E-11	2.62E-05

Numerics

Results on medium-sized proteins

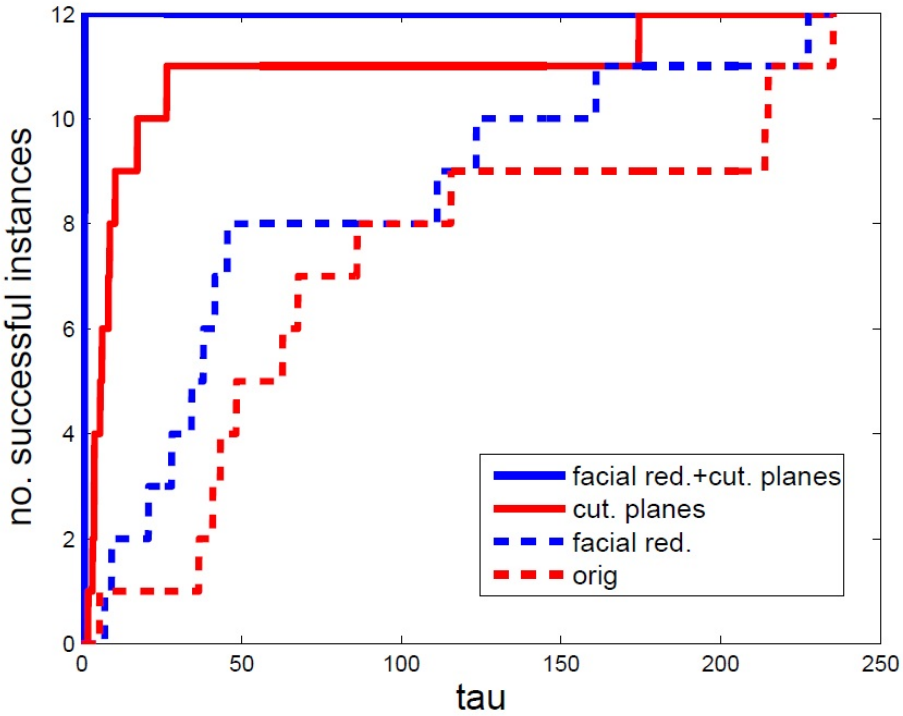
Protein	n_0	p	run time (min)		relative diff	
			SCPCP	orig	SCPCP	orig
1B9O	265	112	0.64	254.85	1.19E-11	2.14
1C5E	200	71	2.59	70.63	4.93E-11	2.01
1C9O	207	53	2.15	66.50	3.35E-12	2.00
1CZP	237	83	1.90	143.95	8.30E-11	2.24
1MFM	216	118	0.19	102.11	2.01E-11	2.00
1QQ4	365	143	5.70	-	6.49E-11	-
1QTN	302	134	5.04	-	2.24E-11	-
1QU9	287	101	7.55	-	1.80E-11	-

Numerics

Results on large proteins

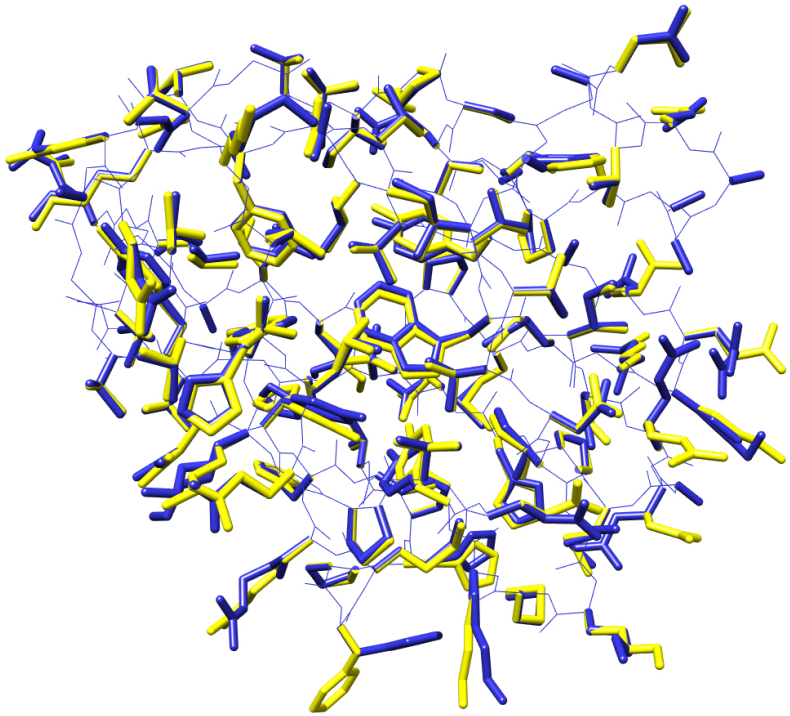
Protein	n_0	p	run time (hr)	rel. diff	numcut	# iter	Final # cuts
1CEX	435	146	0.08	1.26E-11	40	9	485
1CZ9	615	111	3.96	2.98E-13	60	25	1997
1QJ4	545	221	0.15	5.31E-12	60	14	1027
1RCF	581	142	0.85	3.71E-12	60	17	1305
2PTH	930	151	29.65	8.69E-09	120	34	7247
5P21	464	144	0.31	1.39E-12	40	16	822

Individual speedup contribution



$$
\begin{aligned}
t_{i,j} &:= \text{run time for getting the final solution of IQP for instance } i \text{ by method } j, \\
r_{i,j} &:= \frac{t_{i,j}}{\min\{t_{i,j} : j = 1,2,3,4\}}, \\
\rho_j(\tau) &:= \text{number of instance } i \text{ such that } r_{i,j} \leqslant \tau
\end{aligned}
$$

An illustration in protein conformation



Yellow : reconstruction of the protein `1AAC`

Blue : crystallized form of `1AAC` from the Protein Data Bank

Thank you!