Combinatorial Optimization Models for Protein Structure Analysis

Harvey J. Greenberg
Director, Center for Computational Biology
http://www.cudenver.edu/ccb/
University of Colorado at Denver and UC Health Sciences Center
hgreenbe@www.cudenver.edu
http://www.cudenver.edu/~hgreenbe/

- Protein folding
- Protein alignment

Protein Folding

HP Lattice Model (Dill)

- Data: $\mathcal{L}$ = lattice; $\mathcal{N}(p) = \{q \in \mathcal{L} \colon |x_p - x_q| + |y_p - y_q| + |z_p - z_q| = 1\}$;

  $\{H_i\}_1^n$ = 0-1 sequence, where $H_i = 1$ means the i-th amino acid is hydrophobic.

- Variables: $$v_{ip} = \begin{cases} 1 & \text{if acid } i \text{ assigned to point } p; \\ 0 & \text{otherwise.} \end{cases}$$

  $$h_{pq} = \begin{cases} 1 & \text{if } \exists i, j \ni v_{ip} H_i = v_{jq} H_j = 1; \\ 0 & \text{otherwise.} \end{cases}$$

- IP:

$$\max \sum_{p \in \mathcal{L}} \sum_{q \in \mathcal{N}(p)} h_{pq} \colon h_{pq}, v_{ip} \in \{0,1\} \, \forall i, p, q,$$

$$\text{assignments:} \quad \sum_{p=1}^{n^2} v_{ip} = 1, \ \sum_{i=1}^{n} v_{ip} \leq 1$$

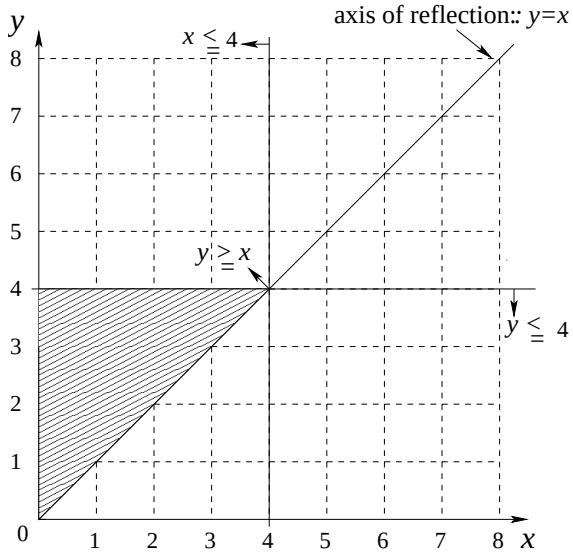
$$\text{sequence order:} \quad \sum_{q \in \mathcal{N}(p)} v_{i+1,q} \geq v_{ip} \ (i < n), \ \sum_{q \in \mathcal{N}(p)} v_{i+1,q} \geq v_{ip} \ (i > 1)$$

$$\text{scoring:} \quad h_{pq} \leq \sum_{i=1}^{n} v_{ip} H_i, \ h_{pq} \leq \sum_{i=1}^{n} v_{iq} H_i,$$

$$\text{symmetry exclusion:} \quad v_{m,p_m} = 1, \ \sum_{p \in \mathcal{P}} v_{1p} = 1.$$

$$\mathcal{P} = \{p \in \mathcal{L} \colon x_p \leq 4, y_p \leq 4, y_p \geq x_p\} = \tfrac{1}{2}\text{Quadrant III}.$$

Region Restricting a_1 to Eliminate Some Symmetric Folds



Linear Programming Relaxation (LPR)

(Used in Branch & Bound Algorithms)

$$\max \sum_{p\in\mathcal{L}} \sum_{q\in\mathcal{N}(p)} h_{pq} : \boxed{h_{pq}, v_{ip} \in [0,1]} \forall i,p,q,$$

$$\text{assignments:} \quad \sum_{p=1}^{n^2} v_{ip} = 1, \ \sum_{i=1}^{n} v_{ip} \le 1$$

$$\text{sequence order:} \quad \sum_{q\in\mathcal{N}(p)} v_{i+1,q} \ge v_{ip} \ (i<n), \ \sum_{q\in\mathcal{N}(p)} v_{i+1,q} \ge v_{ip} \ (i>1)$$

$$\text{scoring:} \quad h_{pq} \le \sum_{i=1}^{n} v_{ip} H_i, \ h_{pq} \le \sum_{i=1}^{n} v_{iq} H_i,$$

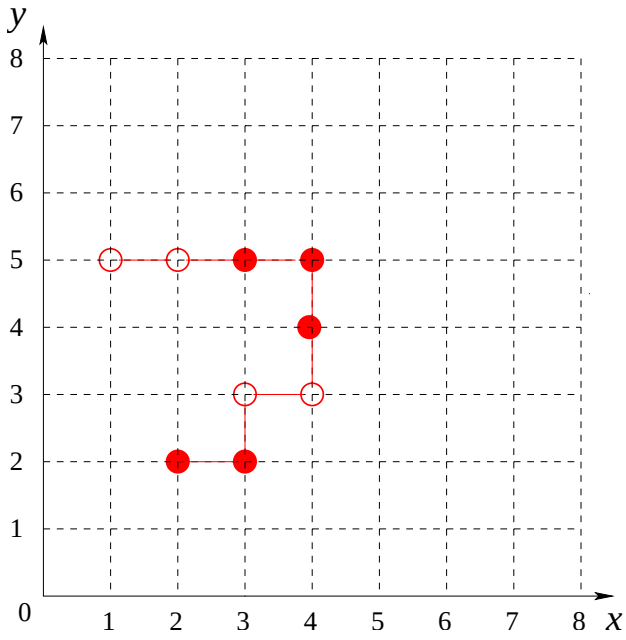
$$\text{symmetry exclusion:} \quad v_{m,p_m} = 1, \ \sum_{p\in\mathcal{P}} v_{1p} = 1.$$

LPR Assignments

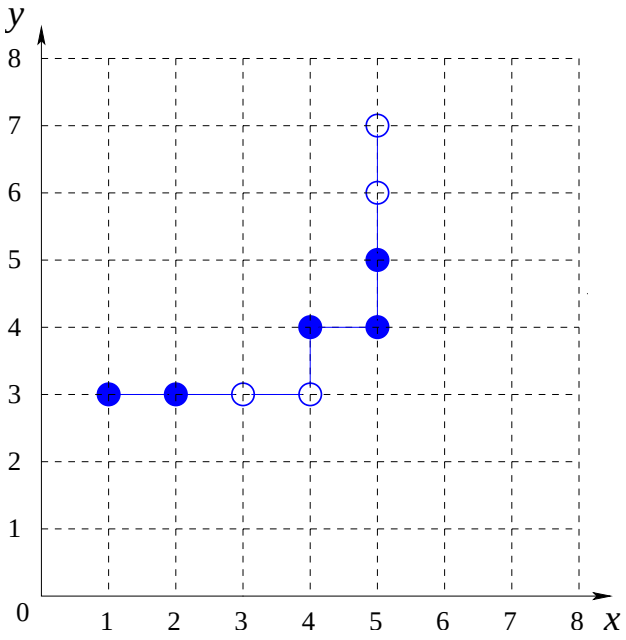
y=8								
y=7					9:25			
y=6					8:25			
y=5	9:25	8:25	7:25	6:25	9:25			
		2:25			7:25			
y=4	2:25	3:50	8:25	5:1	6:25			
		1:25	6:25					
			4:50					
y=3	1:25	2:25	9:25	8:25	7:25			
			3:50	6:25				
			1:25	4:50				
y=2		1:25	2:25	7:25				
y=1								
y=0								
x =	0	1	2	3	4	5	6	7	8

Decompose

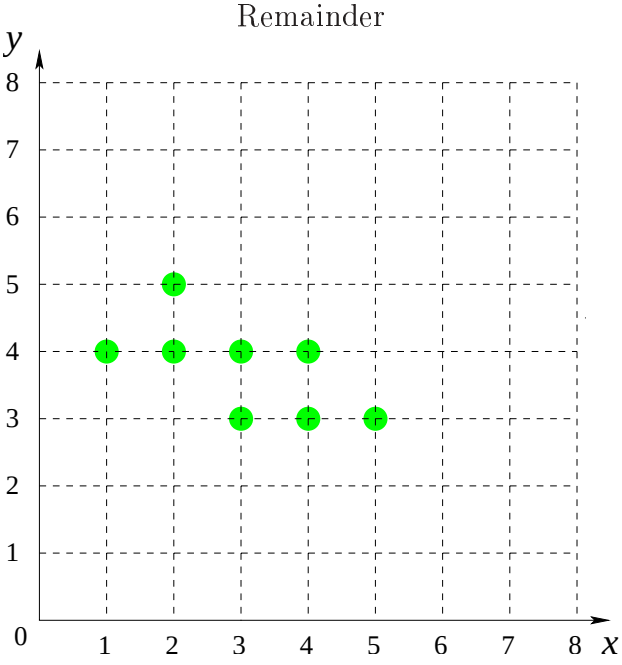
Fold 1



Fold 2

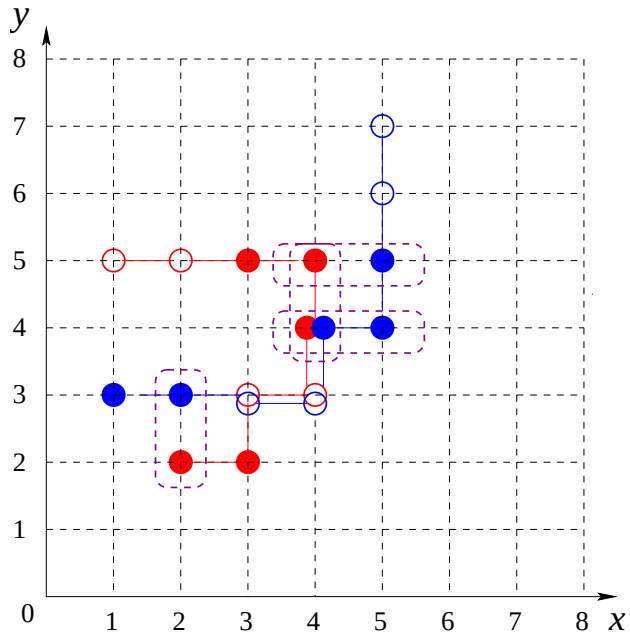


Remainder

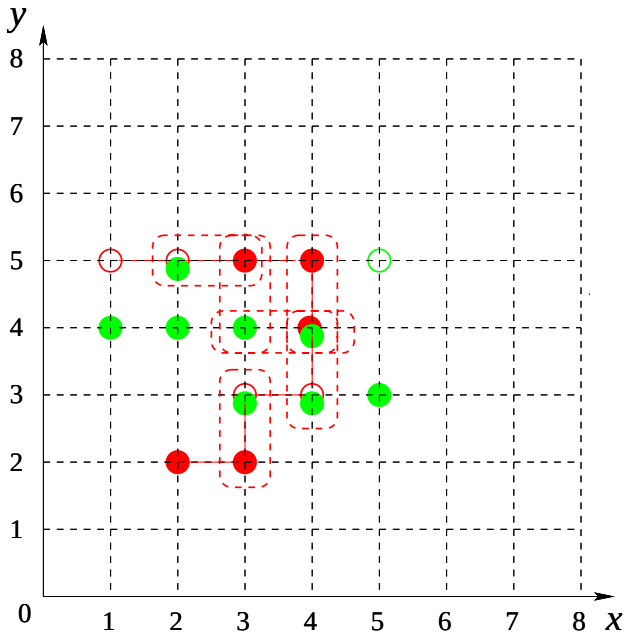


Interactions from Superposition

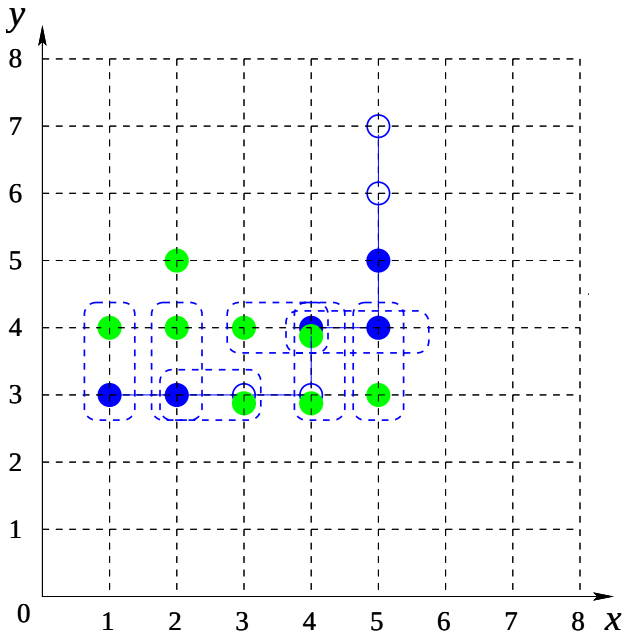
Fold 1–Fold 2 Interactions



Fold 1–Remainder Interactions



Fold 2–Remainder Interactions



Fold Cuts

Given fold v^f, its *support set* is $\sigma(v^f) = \{(i,p) : v^f_{ip} = 1\}$.

Eliminated by $\sum_{(i,p)\in\sigma(v^f)} v_{ip} \leq n-1$.

$$\boldsymbol{v_{1,21} + v_{2,22} + v_{3,31} + v_{4,32} + v_{5,41} + v_{6,50} + v_{7,49} + v_{8,48} + v_{9,47} \leq 8.}$$

$$\boldsymbol{v_{1,29} + v_{2,30} + v_{3,31} + v_{4,32} + v_{5,41} + v_{6,42} + v_{7,51} + v_{8,60} + v_{9,69} \leq 8.}$$

In both cases, LHS $= 3\frac{1}{2}$, so no constraint!

Tucking in the Folds

Require $v_{ip} = 0$ if $|x_p - x_{p_m}| + |y_p - y_{p_m}| > R$ = radius (parameter)

$$R = 2$$

```
y=8
y=7
y=6                          59
y=5                    49    50    51
y=4              39    40   [41]   42    43
y=3                    31    32    33
y=2                          23
y=1
y=0
-------------------------------------------------------
  x =  0     1     2     3     4     5     6     7     8
```

$R = 2$ in example $\Rightarrow$ OPT $= 11\frac{1}{3}$ $(< 12\frac{1}{2})$

```
y=8
y=7
y=6
y=5                    3:33  4:33  7:33
                       7:33  6:33
                       8:33  8:33
                       9:33
y=4              1:33  2:33  5:1   6:33  9:67
                       4:33
                       6:33
y=3                    1:66  2:66  3:33
                             4:33  7:33
                                   8:33
y=2                          3:33
y=1
y=0
-------------------------------------------------------
  x = 0    1      2      3      4      5      6      7      8
```

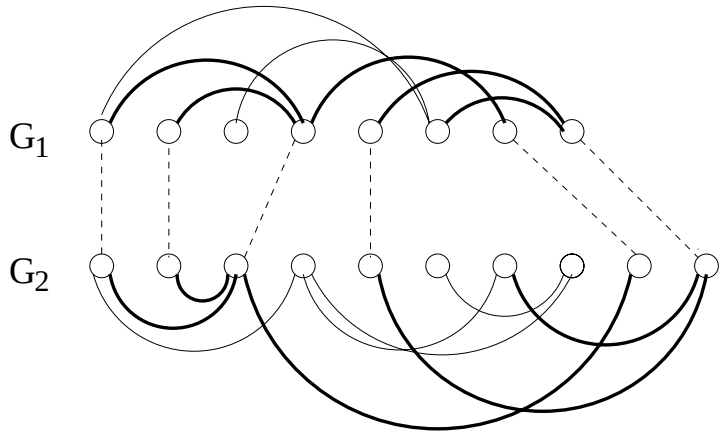
Cannot decompose!

Protein Alignment by Contact Maps

Contact map = graph: $V = \{1, \ldots, n\} \Leftrightarrow$ residues
$E = \{(i, j) : i < j,\ d_{ij} \leq \tau\}$

Given 2 contact maps, their *similarity* is measured by relative size of maximal subgraphs that are isomorphic while preserving order of backbone sequences.

Example: Non-crossing Association with Value = 5



(From Carr-Istrail-Lancia-Walenz [2001].)

0-1 IP Formulation

Given: $G_1 = [V_1, E_1]$ and $G_2 = [V_2, E_2]$.

Variables:

$$x_{ij} = \begin{cases} 1 & \text{if nodes } i \in V_1 \text{ and } j \in V_2 \text{ are associated;} \\ 0 & \text{otherwise.} \end{cases}$$

$$y_{(i,k)(j,l)} = \begin{cases} 1 & \text{if edges } (i,k) \in E_1 \text{ and } (j,l) \in E_2 \text{ are selected;} \\ 0 & \text{otherwise.} \end{cases}$$

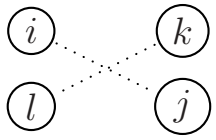
Objective: $\max \sum_{\substack{(i,k) \in E_1 \\ (j,l) \in E_2}} y_{(i,k)(j,l)}$.

Constraints:

Endpoints associated: $y_{(i,k)(j,l)} \leq x_{ij}$ and $y_{(i,k)(j,l)} \leq x_{kl}$;

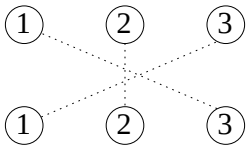
Associations unique: $\sum_{i \in V_1} x_{ij} \leq 1, \ \forall j \in V_2$ and $\sum_{j \in V_2} x_{ij} \leq 1, \ \forall i \in V_1$;

Non-crossing: $x_{ij} + x_{kl} \leq 1$ for $1 \leq i < k \leq |V_1|$ and $1 \leq l < j \leq |V_2|$.



Clique Inequalities

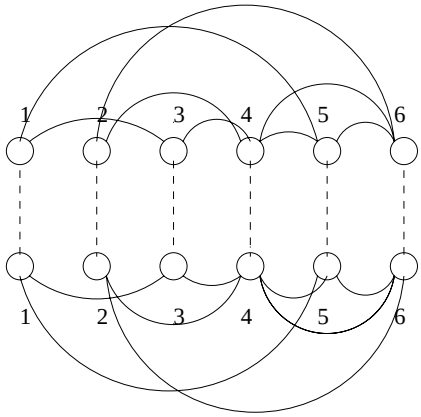
$$x_{13} + x_{22} \leq 1, \ x_{13} + x_{31} \leq 1, \ x_{22} + x_{31} \leq 1.$$



$$\Rightarrow \underbrace{x_{13} + x_{22} + x_{31} \leq 1}_{\text{stronger}}.$$

General: $\sum_{(i,j) \in C} x_{ij} \leq 1$ for any (maximal) clique, C.

Another example: isomorphic contact maps



LPR Solution

	$j \in V_2$					
$i \in V_1$	1	2	3	4	5	6
1	0.167	0.167	0.167	0.167	0.167	0.167
2	0.167	0.333	0.167	0.167	0.167	
3	0.167	0.167	0.167	0.167	0.167	0.167
4	0.167	0.167	0.167	0.167	0.167	0.167
5	0.167	0.167	0.167	0.167	0.167	0.167
6	0.167		0.167	0.167	0.167	0.333

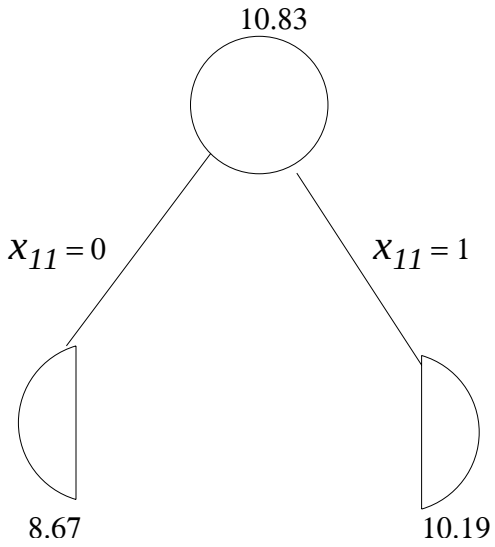
3-cliques $\Rightarrow$ no effect;
4-cliques $\Rightarrow$ no effect;
5-cliques $\Rightarrow$ no effect.

Only 6-clique is:

$$x_{16} + x_{25} + x_{34} + x_{43} + x_{52} + x_{61} \leq 1$$

$\Rightarrow$ no effect!

Branching on one variable ineffective!



	$j \in V_2$					
$i \in V_1$	1	2	3	4	5	6
1	1					
2		0.333	0.333	0.333		
3		0.333	0.333		0.333	
4		0.333		0.333		0.333
5			0.333		0.333	0.333
6				0.333	0.333	0.333

	$j \in V_2$					
$i \in V_1$	1	2	3	4	5	6
1		0.308	0.192	0.038	0.308	
2	0.308	0.192	0.192	0.192	0.115	
3	0.192	0.192	0.077	0.192	0.038	0.308
4	0.038	0.192	0.192	0.192	0.192	0.192
5	0.308	0.115	0.038	0.192	0.038	0.308
6			0.308	0.192	0.308	0.192

Conclusions

- LPR is highly fractionated.
- Symmetry exclusion crucial, even during search.
- Usual cuts and branches ineffective.
- Usual decomposition ineffective.
- Tight folds show promise.