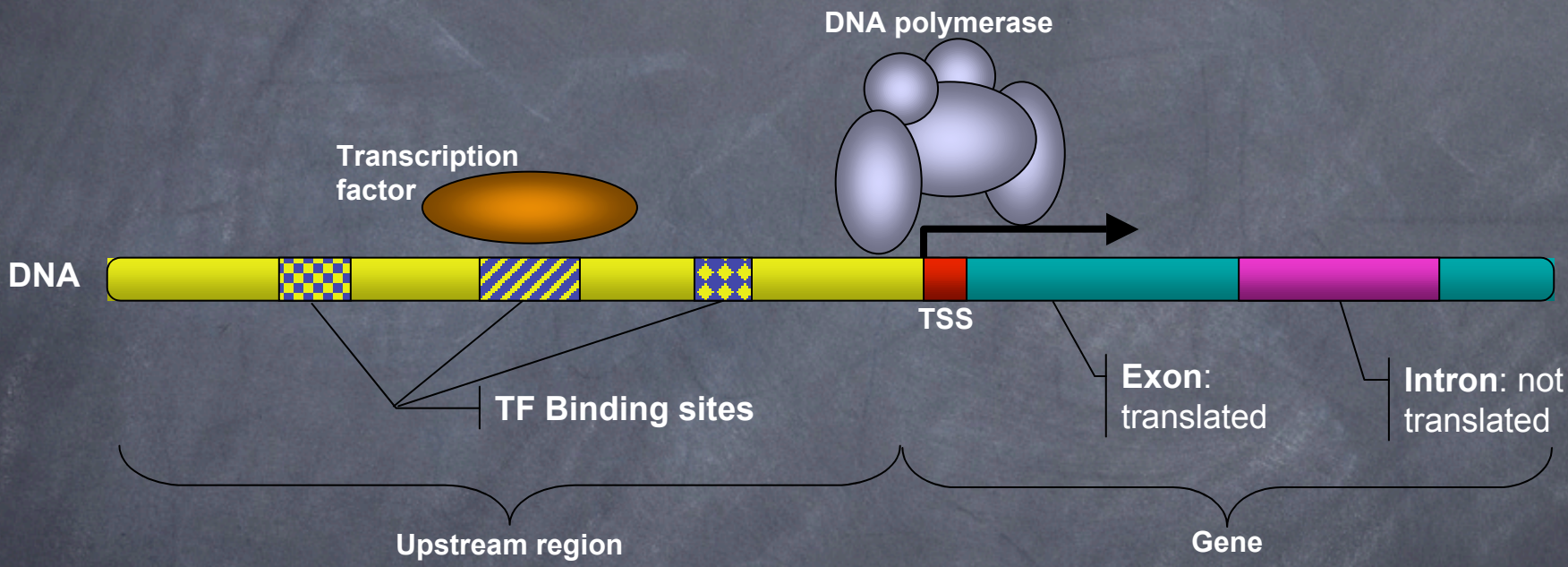
A Compact Mathematical Programming Formulation for DNA Motif Finding

Carl Kingsford
Center for Bioinformatics & Computational Biology
University of Maryland, College Park

Elena Zaslavsky and Mona Singh
Computer Science Dept. and Lewis-Sigler Institute for Integrative Genomics, Princeton University

DNA -> mRNA -> Protein



- Finding transcription factor binding sites can tell us about the cell's regulatory network.

Motif Finding

Transcription factor

1. ttgccacaaaataatccgccttcgcaaattgaccTACCTCAATAGCGGTAgaaaaacgcaccactgcctgacag
2. gtaagtacctgaaagttacggtctgcgaacgctattccacTGCTCCTTTATAGGTAcaacagtatagtctgatgga
3. ccacacggcaaataaggagTAACTCTTTCCGGGTAtgggtatacttcagccaatagccgagaatactgccattccag
4. ccatacccggaaagagttactccttatttgccgtgtggttagtcgcttTACATCGGTAAGGGTAgggattttacagca
5. aaactattaagatttttatgcagatgggtattaaggaGTATTCCCCATGGGTAacatattaatggctctta
6. ttacagtctgttatgtggtggctgttaaTTATCCTAAAGGGGTAtcttaggaatttactt

Given p sequences, find the most mutually similar length-k subsequences, one from each sequence:

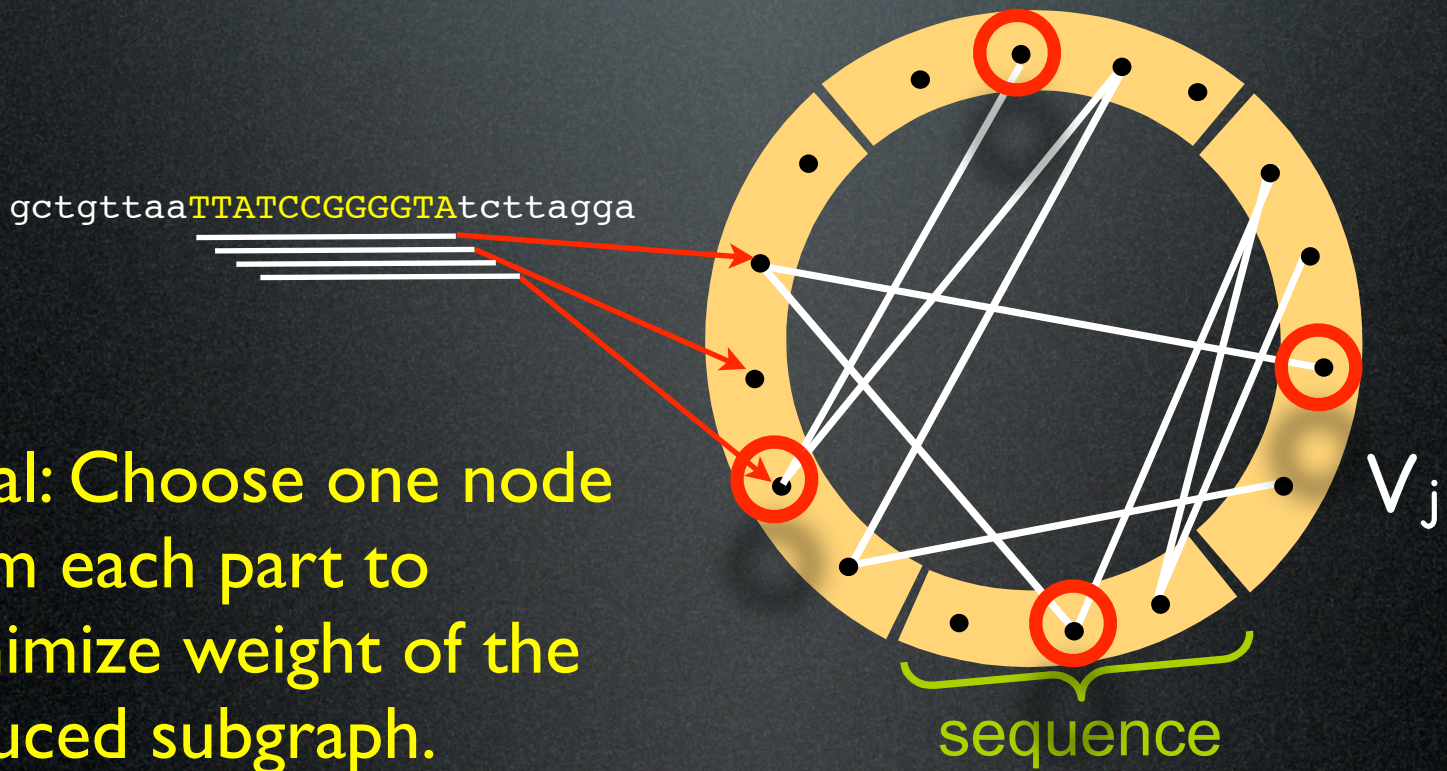
$$\underset{s_1,\ldots,s_p}{\operatorname{argmin}} \sum_{i<j} \operatorname{dist}(s_i, s_j)$$

dist(s_i,s_j) = Hamming distance between s_i and s_j.

Hundreds of papers, many formulations (Tompa05)

Graph Formulation

- For p sequences, complete p-partite graph.
- Node for each sliding window of length k.
- Weight on edge (u,v) = dist(u,v) = # of differences between subsequences u and v.



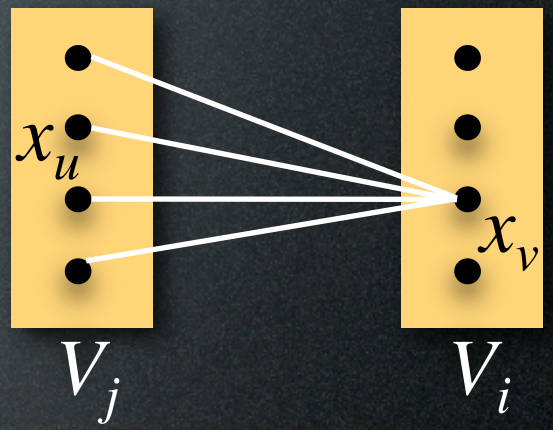
Goal: Choose one node from each part to minimize weight of the induced subgraph.

Hardness

- NP-hard (Wang+94, Akutsu+00).
- General distance measure ⇒ inapproximatible within O(|V|) = # nodes in the graph (Chazelle+04).
- Triangle inequality ⇒ constant-factor approximation (Bafna+97).
- Interested in provably optimal solutions.

Integer Programming Formulation

- Binary variables x_u for each node
- Binary variables x_{uv} for each edge



Minimize $\sum_{\{u,v\} \in E} w_{uv} x_{uv}$

1. $\sum_{u \in V_j} x_u = 1$ for every part j

2. $\sum_{u \in V_j} x_{uv} = x_v$ for every part j, node v

(IP1)

(Zaslavsky & Singh, 05)

Problem: Instances are Huge

205,146 to 16,637,889 variables

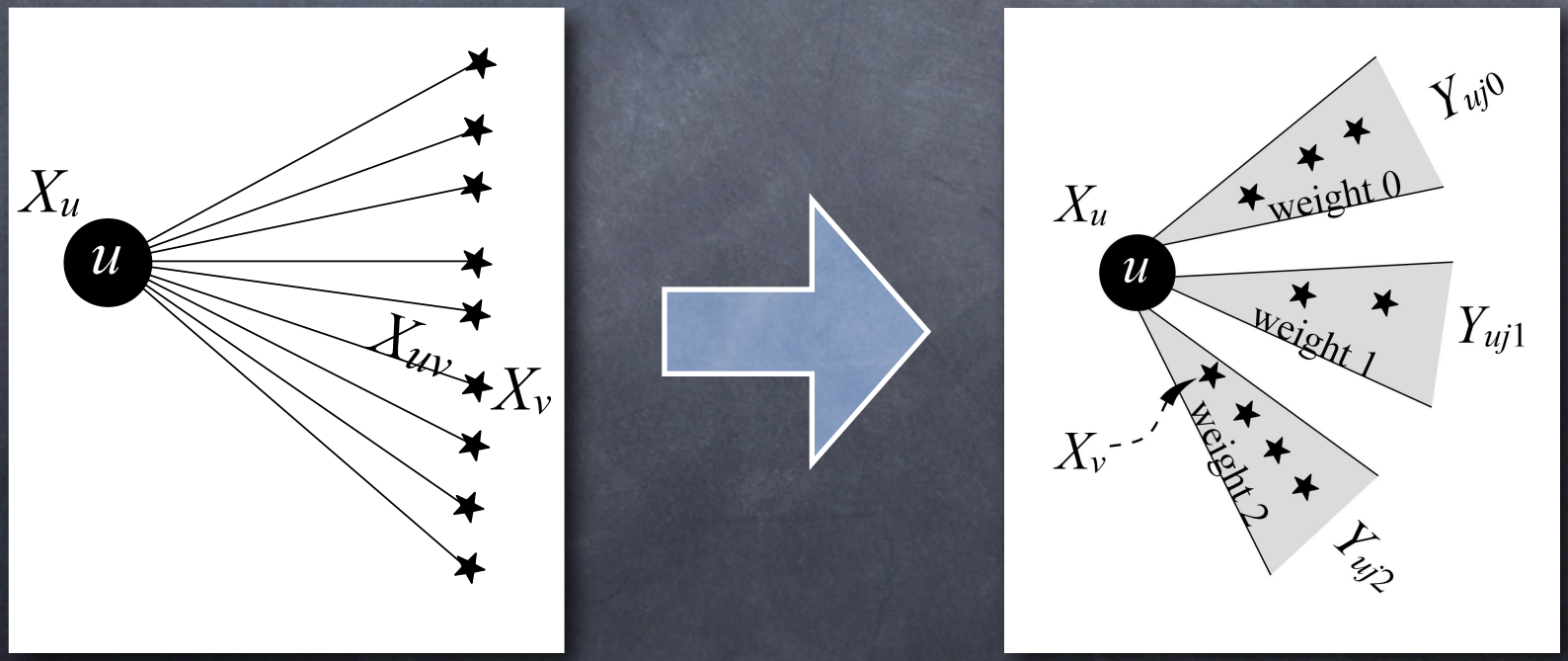
(For the reasonably sized instances in our test set)

Goal: Can we exploit features of the instances
to reduce sizes (or get better formulation)

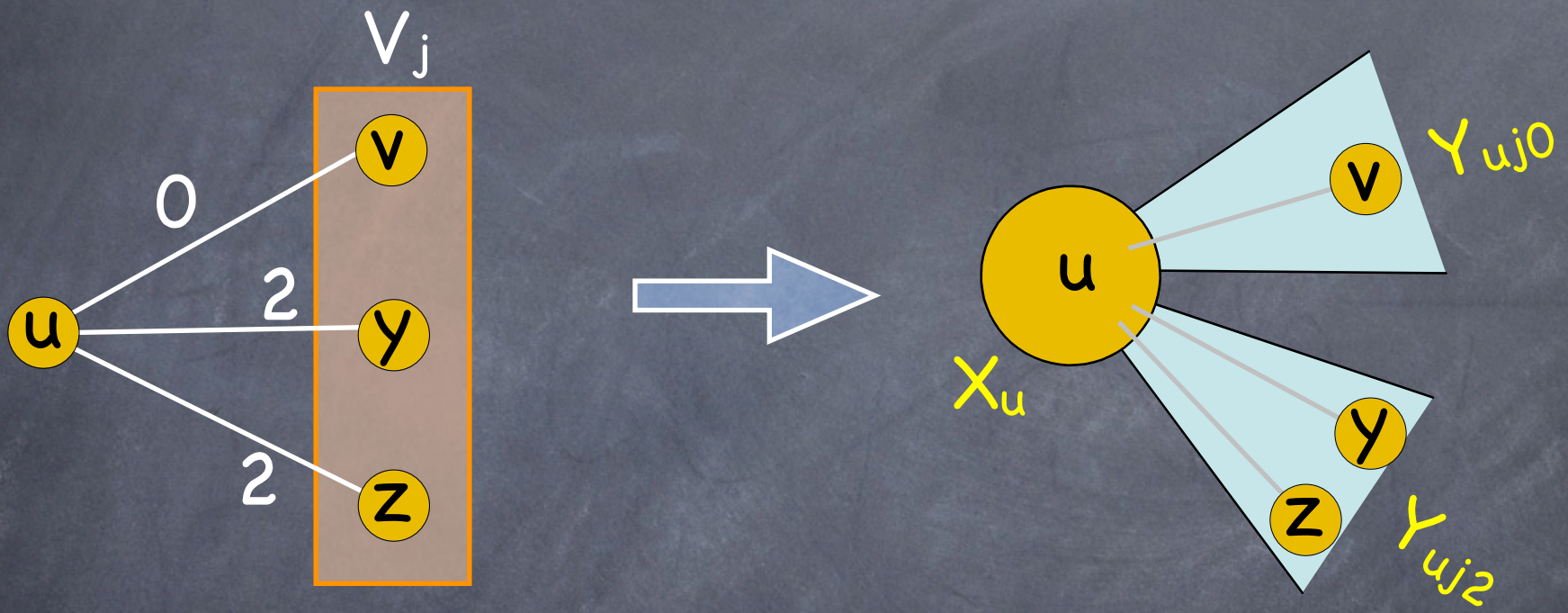
New Formulation

- Only small number of possible edge weights, ≤ window length.
- Don't care which edge is chosen, only that correct cost is paid.

➡ Merge edges that have same cost; vastly reduce # of variables.



Merging Edges



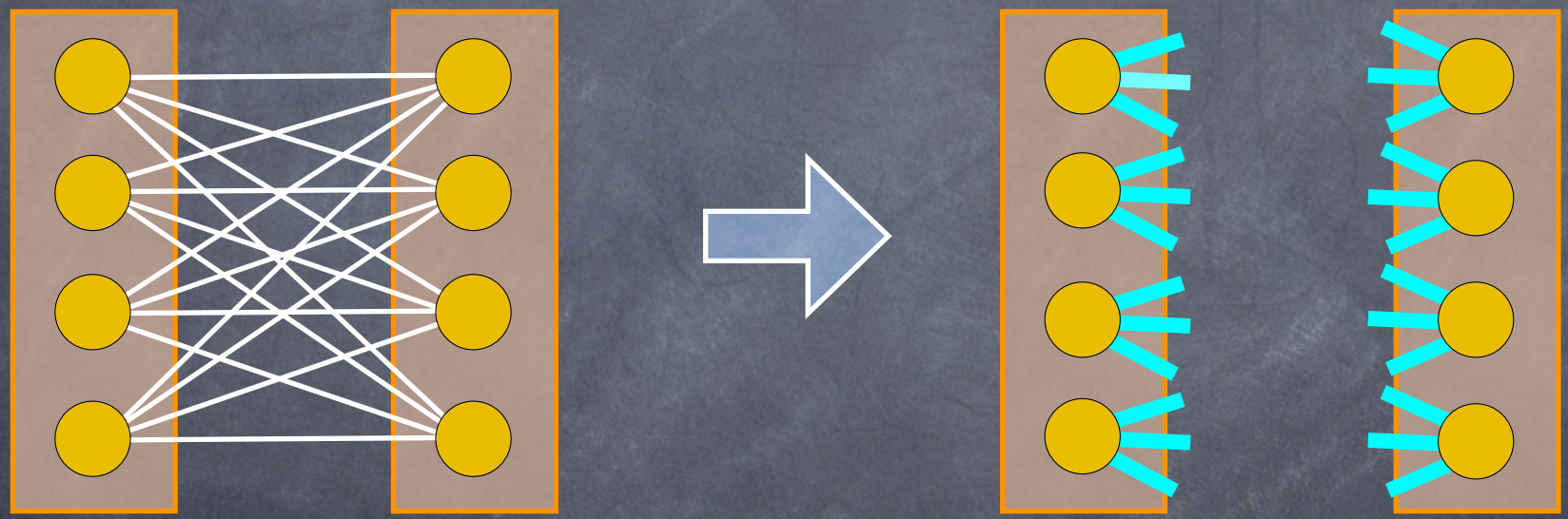
Binary variables:

- X_u on the nodes.
- Y_{ujc} for each node u, position j not containing u, and weight c.

Idea: $Y_{ujc} = 1$ if we choose an edge of weight c between node u and position j.

Reduced # Variables

- N = # nodes per position
- k = length of window = # of possible edge weights - 1



N^2 edge variables

2N(k+1) edge-group variables

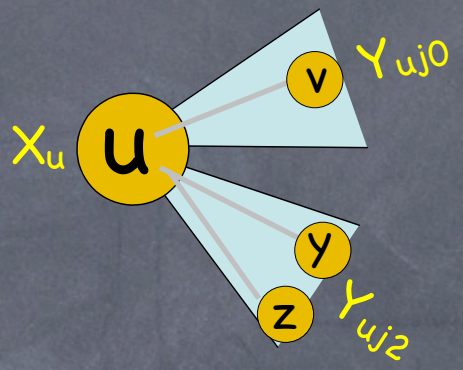
k << N

➡ Need to ensure compatible edge-groups are chosen.

New Integer Program

(IP2)

$$\min \frac{1}{2} \sum_{(u,j,c)} c Y_{ujc}$$



$\sum_{u \in V_j} X_u = 1$	Choose one node from each position.
$\sum_{c \in D} Y_{ujc} = X_u$	If we choose a node, choose a edge group next to it.
$\sum_{\substack{v \in V_j: \\ \mathrm{dist}(u,v)=c}} Y_{vic} \geq Y_{ujc}$	We can only choose an edge-group that "contains" a chosen node.

$X_u, Y_{ujc} \in \{0,1\}$

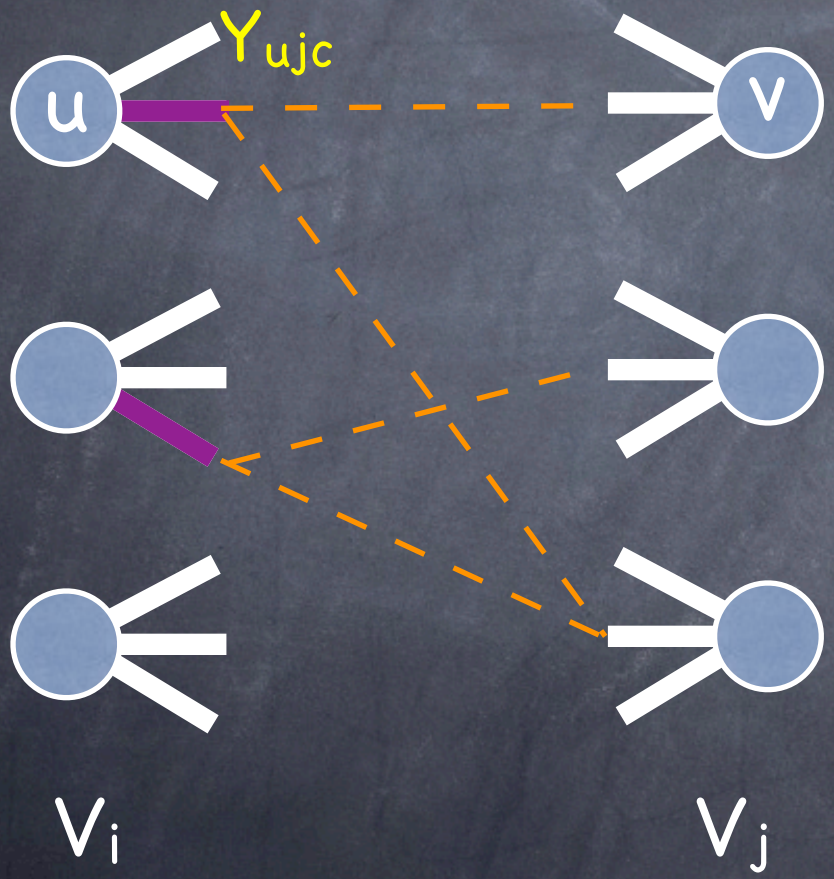
IP1 vs. IP2

- Optimum IP2 = optimum IP1.
- IP2 has a factor of
  - O(k/N) fewer variables, and
  - O(k) more constraints than IP1.

k = motif length
N = sequnce length

- k is fixed by transcription factor geometry; N will grow as longer sequences are considered.
- LP relaxation of IP2 is weaker than that of IP1.
  - ➡ Add constraints to make LP2 as tight as LP1.

Constraining Y Variables



Neighbors of a set of Y vars:

N(ujc) = {(v,i,c) : cost(u,v) = c}

N(Q) = ∪ N(ujc)

Neighbors are compatible

For every set Q: (*)

$$\sum_{(u,j,c)\in Q} Y_{ujc} \leq \sum_{(v,i,c)\in N(Q)} Y_{vic}$$

Tightening LP2

Thm. If all constraints of the form (✳) are included, then the resulting LP has the same optimum as LP1.

Thm (separation algorithm). There is a polytime algorithm to find a violated constraint of the form (✳).

⇒ Despite exponential # of constraints, new LP can be solved in polytime.

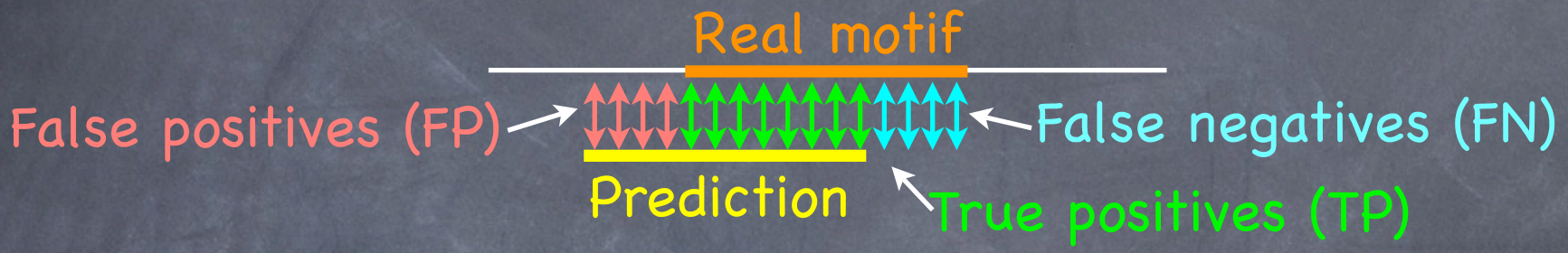
Testing Data Set

- 39 families of E. coli transcription factors from (Robison+98).
- Binding sites found via various experimental techniques.
- 3 - 20 sequences of length ≥ 300 in each family.
- Length of motif ranges: 11 to 48
- Up to 5,960 nodes in resulting graphs.

Formulation is Accurate

- Test on real data.
- Finds biologically relevant solutions.
- Compare performance to state-of-the-art probabilistic approach based on Gibbs sampling.

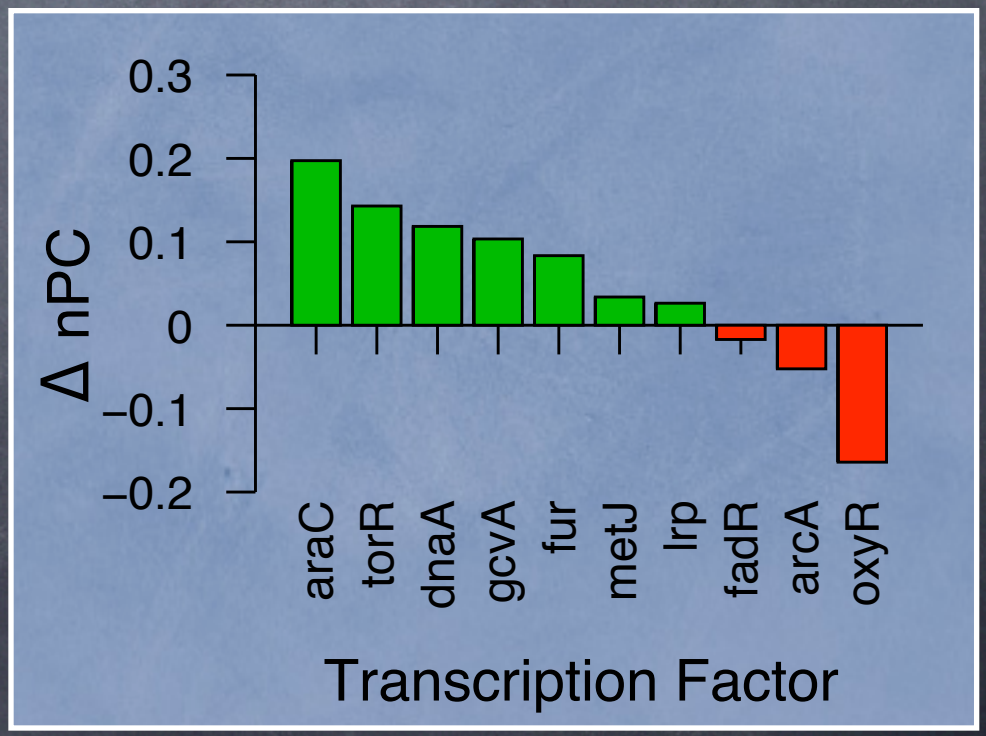
Accuracy Comparisons



Nucleotide level measure of accuracy (Pevzner&Sze,00):

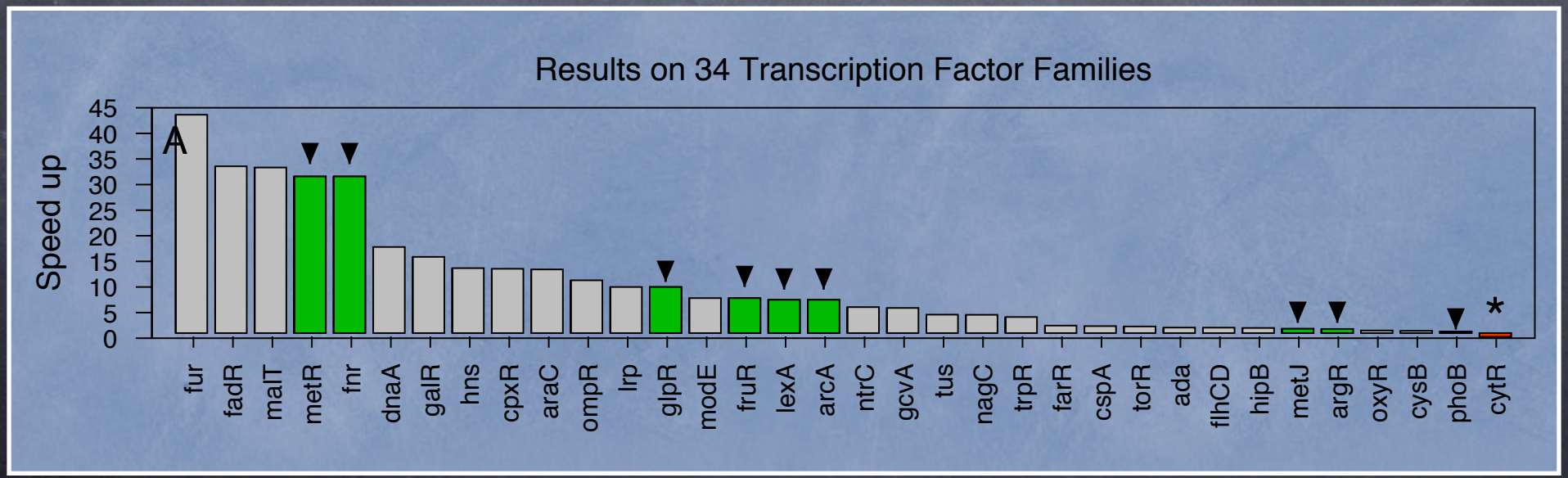
nPC = nTP / (nTP + nFN + nFP)

Compares favorably with Gibbs sampling strategy: plot our nPC minus Gibbs nPC.



Timing Comparison

- Speed up for LP2 + cutting planes to reach same objective function as LP1.
- 10x faster is not uncommon.
- 1 case where LP1 is faster.



(green indicates LP1 did not finish in ≤ 5 hours)

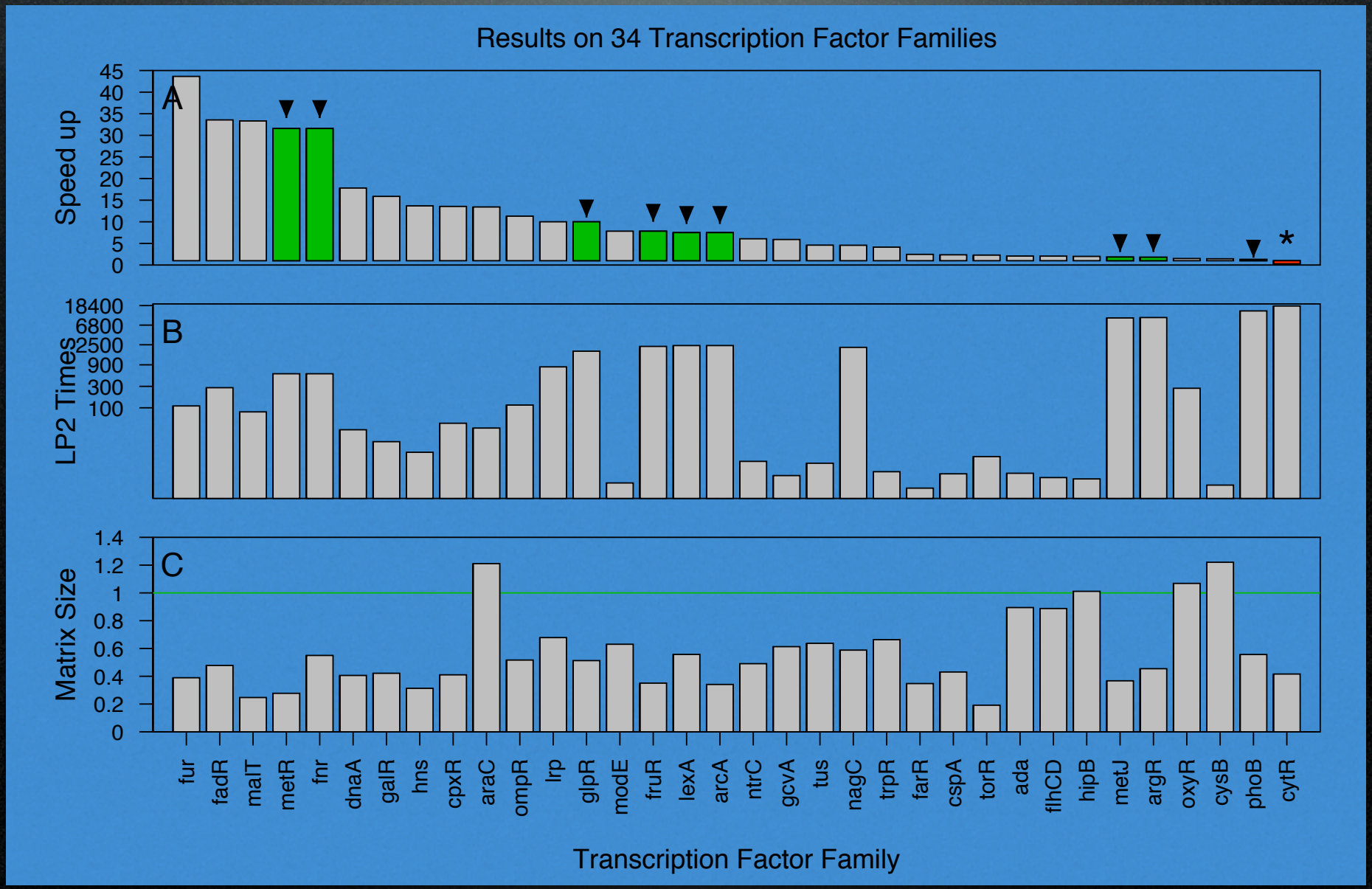
Conclusion

- Able to find provably optimal solutions to real transcription factor binding site discovery problems.
- Large speed up by using bounded number of objective function costs.
- Finds as good or better motifs than other motif-finding approaches.
- Open: how to use triangle inequality and overlapping windows to further shrink IP.

Acknowledgments

- Co-authors: Mona Singh, Elena Zaslavsky
- Additional funds provided by:
  - Center for Bioinformatics and Computational Biology, University of Maryland, College Park
    http://www.cbcb.umd.edu
  - Program in Integrative Information, Computer and Application Sciences (PICASso), Princeton University
    http://www.cs.princeton.edu/picasso/

Times & Sizes


Results on 34 Transcription Factor Families
Speed up
45
40
35
30
25
20
15
10
5
0
A
LP2 Times
18400
6800
2500
900
300
100
B
Matrix Size
1.4
1.2
1
0.8
0.6
0.4
0.2
0
C
fur
fadR
malT
metR
fnr
dnaA
galR
hns
cpxR
araC
ompR
lrp
glpR
modE
fruR
lexA
arcA
ntrC
gcvA
tus
nagC
trpR
farR
cspA
torR
ada
flhCD
hipB
metJ
argR
oxyR
cysB
phoB
cytR
Transcription Factor Family