New Computational Experiments with Binary Quadratic Optimization Models in Computational Biology

**Alexander Engau** (jointly with Gary Kochenberger)
Mathematical and Statistical Sciences, University of Colorado Denver

INFORMS INTERNATIONAL BEIJING 2012, P.R. CHINA, JUNE 24-27, 2012

On Leave at **International College Beijing** (2012-2013), **China Agricultural University**, 17 Qinghua Donglu, Haidian District, Beijing 100083, P.R. China
Binary Quadratic Programming (BQP)

Let \( Q \in \mathbb{R}^{n \times n} \) be symmetric, \( c \in \mathbb{R}^n \), and \( P \subseteq \mathbb{R}^n \) be a polyhedral set:

\[
\text{BQP: } \min_{x \in P \cap \{0, 1\}^n} x^T Q x + c^T x \quad \text{s.t.} \quad x \in P \cap \{0, 1\}^n
\]

Two common “simplifications” use linearization and quadratic convexification:

- **Standard Approach:** Let \( w_{ij} = x_i x_j \) and rewrite BQP as a mixed-integer LP:

  \[
  \min_{x \in P \cap \{0, 1\}^n} c^T x + \sum_{ij} Q_{ij} w_{ij} \quad \text{s.t.} \quad w_{ij} \geq 0, \ w_{ij} \leq x_i, \ w_{ij} \leq x_j, \ w_{ij} \geq x_i + x_j - 1
  \]

  - **Reformulation Linearization Technique (RLT, Adams-Sherali 1986):**
    \[
    a^T x = b \quad \Rightarrow \quad (a^T x)x_i = bx_i \quad \text{and} \quad (a^T x)(1 - x_i) = b(1 - x_i)
    \]

  - **Compact Linearization Technique (CLT, Glover 1975):**
    \[
    x^T Q x = \sum_i x_i \left( \sum_j Q_{ij} x_j \right) = \sum_i w_i
    \]

- **Convex Quadratic Programming Reformulation (CQPR, Hammer-Rubin 1970):**

  \[
  x^T Q x = x^T Q x + \sum_i \lambda_i (x_i^2 - x_i) = x^T (Q + \text{Diag}(\lambda)) x - \lambda^T x
  \]
### Table 3

<table>
<thead>
<tr>
<th>Problem</th>
<th>Start</th>
<th>Standard CPU</th>
<th>Value</th>
<th>Linearization RLT CPU</th>
<th>Value</th>
<th>Glover CPU</th>
<th>Value</th>
<th>No linearization (CQPR) CPU</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>cold</td>
<td>T 96</td>
<td></td>
<td>T 245</td>
<td></td>
<td>T 280</td>
<td></td>
<td>T 239</td>
<td></td>
</tr>
<tr>
<td></td>
<td>warm</td>
<td>T 156</td>
<td></td>
<td>T 17</td>
<td></td>
<td>T 255</td>
<td></td>
<td>T 239</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hot</td>
<td>T 270†</td>
<td></td>
<td>T 270†</td>
<td></td>
<td>T 279</td>
<td></td>
<td>T 270†</td>
<td></td>
</tr>
<tr>
<td>medium</td>
<td>cold</td>
<td>T none</td>
<td>M</td>
<td>none</td>
<td></td>
<td>T none</td>
<td></td>
<td>T none</td>
<td></td>
</tr>
<tr>
<td></td>
<td>warm</td>
<td>T 179†</td>
<td>M</td>
<td>none</td>
<td></td>
<td>T 179†</td>
<td></td>
<td>T 179†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hot</td>
<td>T 1,930†</td>
<td>M</td>
<td>none</td>
<td></td>
<td>T 1,930†</td>
<td></td>
<td>T 1,930†</td>
<td></td>
</tr>
<tr>
<td>large</td>
<td>cold</td>
<td>T none</td>
<td>M</td>
<td>none</td>
<td></td>
<td>T none</td>
<td></td>
<td>T none</td>
<td></td>
</tr>
<tr>
<td></td>
<td>warm</td>
<td>T 2,856†</td>
<td>M</td>
<td>none</td>
<td></td>
<td>T 2,856†</td>
<td></td>
<td>T 2,856†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hot</td>
<td>T 4,352†</td>
<td>M</td>
<td>none</td>
<td></td>
<td>T 4,352†</td>
<td></td>
<td>T 4,352†</td>
<td></td>
</tr>
</tbody>
</table>

† Same as initial value; M out of memory (16 GB); T out of time (1 hr).

Unix CPU Times and Best Values Found for QBP of MSA
Let $S^n = \{ X \in \mathbb{R}^{n \times n} : X = X^T \}$ be the set of symmetric matrices and define:

- $S^n_+ = \{ X \in S^n : u^T X u \geq 0 \text{ for all } u \in \mathbb{R}^n \}$ semidefinite matrices
- $C^n = \{ X \in S^n : u^T X u \geq 0 \text{ for all } u \geq 0 \}$ ($S^n_+ \subset C^n$) copositive matrices
- $C^n_+ = \{ X \in S^n : X = VV^T \text{ for } V \geq 0 \in \mathbb{R}^{n \times m} \}$ completely positive matrices
- $S^n_{++} = \{ X \in S^n_+ : X \geq 0 \}$ ($C^n_+ \subset S^n_{++} \subset S^n_+$) doubly nonnegative matrices

Theorem (Burer 2009)

Let $P = \{ x \in \mathbb{R}^n : a_j^T x = b_j \}$ and $\mathcal{P} = \{ X \in \mathbb{R}^{n \times n} : a_j^T X a_j = b_j^2 \}$. Then:

$$\min_{x \in P \cap \{0,1\}^n} \left\{ x^T Q x + c^T x \right\} = \min_{(X,x) \in \mathcal{P} \times P} \left\{ \text{trace}(QX) + c^T x : \begin{pmatrix} 1 & x^T \\ x & X \end{pmatrix} \in C^n_{+1} \right\}. $$

Semidefinite programming (SDP) and doubly nonnegative (DNN) relaxations approximate $C^n_+$ by $S^n_+$ or $S^n_{++}$ (much current research on cone hierarchies).
Semidefinite Programming

Let \( S_+^n = \{ X \in \mathbb{R}^{n \times n} : X = X^T \succeq 0 \} \) and \( Q \bullet X = \text{trace}(QX) = \sum Q_{ij}X_{ij} \):

\[
\text{SDP:} \quad \min \quad Q \bullet X \quad \text{s.t.} \quad X \in \mathcal{P} \cap S_+^n
\]

- **Relaxation 1**: In BQP, relax \( X = xx^T \) to \( X \succeq 0 \) and drop that \( \text{rank}(X) = 1 \)
  - Constraint “quadratification”: \( \mathcal{P} = \{ X : \text{diag}(X) \in P \} \), \((a^T x - b)^2 = 0\), RLT, etc.

- **Relaxation 2**: In BQP, relax \( X - xx^T = 0 \) to \( X - xx^T \succeq 0 \) and \( \text{diag}(X) = x \)
  - We can use the Schur complement to write \( X - xx^T \succeq 0 \) as \( \begin{pmatrix} 1 & x^T \\ x & X \end{pmatrix} \succeq 0 \).

**Schur Complement Lemma**

Let \( M = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \) be symmetric, and let \( A \) be invertible. Then

\[
M \succeq 0 \quad \text{if and only if} \quad A \succeq 0 \quad \text{and} \quad C - B^T A^{-1} B \succeq 0.
\]

For DNN, replace \( S_+^n \) by \( S_{++}^n = \{ X \in S_+^n : X \geq 0 \} \):

\[
\text{DNN:} \quad \min \quad Q \bullet X \quad \text{s.t.} \quad X \in \mathcal{P} \cap S_{++}^n
\]

Compared to (basic) SDP, DNN has \( O(n^2) \) additional inequality constraints.
Solving SDP/DNN Relaxations

There are several algorithms for solving semidefinite optimization problems:

- **Interior-point methods**: generally efficient and robust, but limited in problem size due to second-order nature (CSDP, DSDP, SeDuMi, SDPT3, SDPA).
- **(Spectral) bundle methods**: very efficient for some SDPs that are formulated as eigenvalue problems, less robust for general problems (ConicBundle).
- **Augmented Lagrangian and alternating direction methods**: similar behavior as interior-point methods, but can be specialized for large-scale problems.

We recently developed a new class of interior-point cutting-plane methods that can “efficiently” handle large classes of known inequalities and cutting planes:

- integrates a cutting-plane scheme within an infeasible path-following IPM;
- uses a set of indicators that predict the violation of non-active inequalities;
- adds selected inequalities dynamically using IPM warm-starting techniques.

The method is described in E. et al. 2012 (Optimization Methods and Software) with a new polynomial convergence proof given in E.-Anjos (GERAD 2011-44).
Problems in molecular biology are often related to protein similarity and folding:

- How does the structural shape of a protein determine its (dis-)function?
- How is its folded structure determined by its linear chain of amino acids?
  - Proteins are biochemical compounds formed from a sequence of amino acids.
  - Amino acids are molecules formed by triples of nucleotides from DNA via mRNA.
  - Messenger Ribonucleic Acids (mRNA) are molecules “transcribed” from DNA.
  - Deoxyribonucleic Acid (DNA) is a double helix of base pairs adenine-thymine (A-T) and cytosine-guanine (C-G) attached to a sugar-phosphate backbone.
Secondary Structure of DNA and Amino Acids

**Diagram:**
- Adenine (5' end)
- Thymine
- Phosphate-deoxyribose backbone
- Cytosine (3' end)
- Guanine (5' end)

**Twenty-One Amino Acids**

A. Amino Acids with Electrically Charged Side Chains
   - Arginine (R)
   - Histidine (His)
   - Lysine (K)
   - Aspartic Acid (Asp)
   - Glutamic Acid (Glu)

B. Amino Acids with Polar Uncharged Side Chains
   - Serine (Ser)
   - Threonine (Thr)
   - Asparagine (Asn)
   - Glutamine (Gln)

C. Special Cases
   - Cysteine (Cys)
   - Selenocysteine (Sec)
   - Glycine (Gly)
   - Proline (Pro)

D. Amino Acids with Hydrophobic Side Chain
   - Alanine (Ala)
   - Valine (Val)
   - Isoleucine (Ile)
   - Leucine (Leu)
   - Methionine (Met)
   - Phenylalanine (Phe)
   - Tyrosine (Tyr)
   - Tryptophan (Trp)
Tertiary Structure of DNA and Amino Acids
Protein Folding

A protein’s native state is that folded shape with the minimum of free energy.

- **Bryngelson’s Energy Landscape Theory**: proteins seek minimal frustration.
- **Anfinsen’s Dogma**: a protein’s native state is determined only by its amino acid sequence (at given environmental conditions, a protein’s structure is the unique, stable and kinetically accessible minimum of the free energy).
- **Levinthal Paradox**: the number of possible conformations is very large.

- **Experimental methods**: mutation studies, dual polarisation interferometry, circular dichroism and nuclear magnetic resonance (NMR) spectroscopy.
- **Computational methods**: modeling/simulation of minimal energy landscape.
Rotamer Assignment (RoA)

Protein structure depends on the torsion angles of its side-chain conformation:

- **Isomers** are compounds with the same molecular but different structural formulas, that can be interconverted by rotations about its single bonds.
- **Rotamers** are isomers whose rotations are restricted by energy barriers.

The following model is given by Fung et al. (07) and Forrester-Greenberg (08):

Minimize

\[ \sum_{i} \sum_{r \in R_i} \left( E_{ir} x_{ir} + \sum_{j>i} \sum_{t \in R_j} E_{irjt} x_{ir} x_{jt} \right) \]

subject to

\[ \sum_{r \in R_i} x_{ir} = 1 \text{ for all } i = 1, \ldots, n \]

\[ x_{ir} \in \{0, 1\} \text{ for all } i \text{ and } r \in R_i \]

- **Sets**
  - \( i = 1, \ldots, n \) set of rotamer sites
  - \( R_i \) set of possible rotamers at \( i \)

- **Data Parameters**
  - \( E_{ir} \) energy barrier of \( r \in R_i \) at \( i \)
  - \( E_{irjt} \) energy between \((r, t)\) at \((i, j)\)

- **Optimization Variables**
  - \( x_{ir} = \begin{cases} 1 & \text{if } r \text{ is assigned to site } i \\ 0 & \text{otherwise} \end{cases} \)
RoA Test Problems and Results

We used all test problems from Forrester-Greenberg (2008) based on energy data from the molecular dynamics simulation at Sandia National Laboratories.

| Protein | Sites | Rotamers ($\sum R_i$) | Binary Vars | Search Space ($\prod |R_i|$) |
|---------|-------|------------------------|-------------|---------------------------|
| 1aboFull | 10    | 1,546                  | 1,546       | 3.418 $10^{21}$          |
| 1bbzFull | 10    | 1,614                  | 785         | 4.602 $10^{21}$          |
| 1ddvFull | 6     | 1,016                  | 1,016       | 9.111 $10^{21}$          |

The following table gives the minimum energy, our SDP and DNN bounds with number of inequalities (nnc), times (cpu), and approximation qualities (approx):

<table>
<thead>
<tr>
<th>Protein</th>
<th>Energy</th>
<th>SDP</th>
<th>cpu</th>
<th>approx</th>
<th>DNN</th>
<th>nnc</th>
<th>cpu</th>
<th>approx</th>
</tr>
</thead>
<tbody>
<tr>
<td>1abo</td>
<td>-58.3</td>
<td>-1.0e+7</td>
<td>22.0</td>
<td>1.9e-3</td>
<td>-61.0</td>
<td>3,405</td>
<td>333.8</td>
<td>4.1e-1</td>
</tr>
<tr>
<td>1bbz</td>
<td>-58.5</td>
<td>-1.4e+7</td>
<td>2.7</td>
<td>9.7e-3</td>
<td>-61.4</td>
<td>3,257</td>
<td>40.9</td>
<td>1.7e-1</td>
</tr>
<tr>
<td>1ddv</td>
<td>-34.8</td>
<td>-3.7e+7</td>
<td>4.3</td>
<td>1.2e-3</td>
<td>-246.0</td>
<td>5,000</td>
<td>78.0</td>
<td>5.6e-3</td>
</tr>
</tbody>
</table>

- The basic SDP relaxation (without nonnegativity constraints) is very weak.
- The DNN relaxation with selective addition of inequalities works quite well.
Contact Map Overlap

Protein structure can also be described by contact map graphs $G = (V, E)$.

- The (ordered) vertex set $V$ represents the (linear) sequence of amino acids.
- The edge set $E$ indicates which amino acids are close (distance $\leq 4.5$ Å).
- The contact map overlap of two protein graphs $(G_1, G_2)$ is their largest isomorphic subgraph that preserves the order of their backbones.

\[
\begin{align*}
\text{Max } & \sum_{(i,k) \in E_1, (j,l) \in E_2} x_{ij} x_{kl} \\
& \quad \text{subject to} \\
& \sum_{j \in V_2} x_{ij} \leq 1 \quad \text{for all } i \in V_1 \\
& \sum_{i \in V_1} x_{ij} \leq 1 \quad \text{for all } j \in V_2 \\
& x_{ij} + x_{kl} \leq 1 \quad \text{for } 1 \leq i < k < |V_1|, 1 \leq l < j < |V_2| \\
& x_{ij} \in \{0, 1\} \quad \text{for all } i \in V_1, j \in V_2 \\
\end{align*}
\]
CMO Test Problems and Results

The following data was obtained from the Protein Data Bank (www.pdb.org).

| Protein 1 | $|V_1|$ | $|E_1|$ | Protein 2 | $|V_2|$ | $|E_2|$ | Binary Variables |
|-----------|-------|-------|-----------|-------|-------|-----------------|
| 1avy      | 58    | 56    | 1f22      | 48    | 55    | 2,784           |
| 1f22      | 48    | 55    | 1f22      | 48    | 55    | 2,304           |
| 1qr8      | 54    | 75    | 1qr9      | 55    | 72    | 2,970           |
| 8msi      | 58    | 108   | 9msi      | 59    | 112   | 3,422           |

- **New problem**: There are even more crossing than nonnegativity constraints.
- **Current strategy**: We also used the crossing constraints as cutting planes.

The following table gives maximum overlap, our SDP and DNN bounds with number of inequalities (nnc), times (cpu), and approximation qualities (approx):

<table>
<thead>
<tr>
<th>Protein Pair</th>
<th>CMO</th>
<th>SDP</th>
<th>cpu</th>
<th>approx</th>
<th>DNN</th>
<th>nnc</th>
<th>cpu</th>
<th>approx</th>
</tr>
</thead>
<tbody>
<tr>
<td>1avy-1f22</td>
<td>21</td>
<td>76</td>
<td>265</td>
<td>4.8e-3</td>
<td>45</td>
<td>5,000</td>
<td>3,672</td>
<td>3.3e-2</td>
</tr>
<tr>
<td>1f22-1f22</td>
<td>55</td>
<td>79</td>
<td>157</td>
<td>3.7e-1</td>
<td>55</td>
<td>4,372</td>
<td>1,100</td>
<td>9.7e-1</td>
</tr>
<tr>
<td>1qr8-1qr9</td>
<td>61</td>
<td>133</td>
<td>384</td>
<td>8.1e-2</td>
<td>64</td>
<td>5,000</td>
<td>5,922</td>
<td>4.2e-1</td>
</tr>
<tr>
<td>8msi-9msi</td>
<td>105</td>
<td>242</td>
<td>589</td>
<td>1.2e-3</td>
<td>114</td>
<td>5,000</td>
<td>8,812</td>
<td>5.2e-2</td>
</tr>
</tbody>
</table>

- **Good bounds, but challenged by high computation time and inequality limit.**
Comparison with Heuristic Approach based on Tabu Search

Kochenberger et al. (2004, 2008) model any BQP as unconstrained BQP:

- **Linear constraints** in \( \{0, 1\} \)-variables are **added to objective as penalty**.
- **Example**: Set packing constraint of the form \( \sum_{i=1}^{n} x_i \leq 1 \) are added as

\[
P \left( \sum_{i=1}^{n-1} x_i \sum_{j=i+1}^{n} x_j \right) = P \left( \sum_{i=1}^{n-1} x_i x_{i+1} + \sum_{i=1}^{n-2} x_i x_{i+2} + \sum_{i=1}^{n-3} x_i x_{i+3} + \ldots + x_1 x_n \right)
\]

- All CMO constraints (assignment and crossing) are set-packing constraints!
- Unconstrained BQP is solved using a basic tabu search (Glover et al. 1999)

The following table repeats times for SDP and DNN relaxation, and gives time and number of iterations for above \( xQx \) **reformulation** solved by tabu search:

<table>
<thead>
<tr>
<th>Protein Pair</th>
<th>BinVars</th>
<th>SDP</th>
<th>DNN</th>
<th>XQX</th>
<th>iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1avy-1f22</td>
<td>2,784</td>
<td>265</td>
<td>3,672</td>
<td>139</td>
<td>1,366,057</td>
</tr>
<tr>
<td>1f22-1f22</td>
<td>2,304</td>
<td>157</td>
<td>1,100</td>
<td>61</td>
<td>39,681</td>
</tr>
<tr>
<td>1qr8-1qr9</td>
<td>2,970</td>
<td>384</td>
<td>5,922</td>
<td>172</td>
<td>1,622,560</td>
</tr>
<tr>
<td>8msi-9msi</td>
<td>3,422</td>
<td>589</td>
<td>8,812</td>
<td>217</td>
<td>1,802,857</td>
</tr>
</tbody>
</table>

- The XQX approach was able to find optimal solutions for all four instances.
- Computation times are much shorter despite the large number of iterations.
Conclusions

We discussed two models in computational biology that make stimulating and challenging test problems for BQP relaxation and reformulation approaches:

1. Semidefinite relaxations can be strengthened relatively easily by the doubly nonnegative relaxation (using nonnegativity constraints as cutting planes).

2. Fractional solutions tend to clearly indicate good feasible integer solutions (existing randomized rounding strategies further close the optimality gap).

3. Linear relaxations and quadratic convex reformulations are computationally dominated by quadratic unconstrained reformulation solved by tabu search.

Thanks to Gary Kochenberger, and interest in new (local) collaborators!

Thank you! (Email: alexander.engau@udenver.edu)