SnB: A STATUS REPORT

Russ Miller$^{1,2}$ and Charles M. Weeks$^2$

$^1$ Dept. of Comp. Sci., SUNY-Buffalo
$^2$ The Medical Foundation of Buffalo

{miller,weeks}@mfb.buffalo.edu
(716) 856-9600

Partial funding from NIH and NSF.
Computing from TMC, PSC, Intel, and NIH.

Principal Contributors

C.-S. Chang       G.T. DeTitta       S.M. Gallo
H.A. Hauptman     R. Jones          H. Khalak
D.A. Langs        R. Miller         S. Potter
P. Thuman         C.M. Weeks
History

1988: Hauptman develops *Minimal Function*

1989: Simulated Annealing in Reciprocal Space

1991: *Shake-and-Bake* (Parameter Shift)

1992: *SnB*

1992: Genetic Algorithms

1993: Simulated Annealing in *SnB*

1993: Tangent Formula in *SnB*
Computing Platforms

1. Intel iPSC/2 (UB)
2. Intel iPSC/860 (NIH/Intel)
3. Thinking Machines Corporation CM-2/200 (TMC/PSC)
4. Network of Sun Workstations (UB/CIT)
5. Thinking Machines Corporation CM-5 (TMC/PSC)
6. Network of SGI Workstations (MFB)
7. Cray C90 (PSC)
8. Cray T3D (PSC)
Overview

1. *Shake-and-Bake*
   - Direct Methods Optimization Approach
   - Real Space $\leftrightarrow$ Reciprocal Space
   - *Minimal Function* also as FOM
   - Sets of Random Trial Structures

2. *SnB* is the program based on *Shake-and-Bake*.

3. Phase refinement techniques considered:
   (a) parameter shift
   (b) modified global binary search
   (c) tangent formula
   (d) gradient descent
   (e) simulated annealing
   (f) genetic algorithms
Shake-and-Bake

Phase
Refinement

FFT

Map
Interpretation

Trial Structure

Structure Factors

Reciprocal Space

Real Space

Result

Structure Factors

Structural Factors
The Minimal Function

\[
R = \frac{\sum_T W_T (\cos \phi_T - I_T)^2 + \sum_Q W_Q (\cos \phi_Q - I_Q)^2}{\sum_T W_T + \sum_Q W_Q}
\]

\[
\phi_T = \phi_h + \phi_k + \phi_{-h-k}
\]

\[
\phi_Q = \phi_l + \phi_m + \phi_n + \phi_{-l-m-n}
\]

*I_T, I_Q* — Known expected value of cosine of corresponding structure invariant.

*W_T, W_Q* — Weight factor inversely proportional to the variance.
Phase Refinement by Parameter Shift

\[ R(\Phi) = 0.45 \]

\[ \phi_i \pm 180^\circ \]

\[ R(\Phi) = 0.43 \]

\[ \phi_i - 90^\circ \]

\[ R(\Phi) = 0.44 \]

\[ \phi_i + 90^\circ \]

\[ R(\Phi) = 0.47 \]
Default Parameters for $n$ Atom Structure

1. Random Atoms per Trial Structure: $n$
2. Phases: $10n$
3. Triples: $100n$
4. Quartets: 0
5. Refinement Routine: Parameter Shift
   - P1: $(\pi/2, 2, 1)$
   - Centrosymmetric: $(\pi, 1, 1)$
   - Other: $(\pi/2, 2, 3)$
6. $SnB$ Cycles: $n/2$
7. Restricted Phases: no
8. Heavy Atoms: yes
SnB Structure Determination Procedure

1. Search path: ./
2. Reflection file prefix: ph8755
3. Space group: P1
4. Cell constants:
   A: 9.1300  ALPHA: 96.7300
   B: 10.5080  BETA: 99.2900
   C: 15.0770  GAMMA: 89.0900
5. Contents of the asymmetric unit: O14,N9,C51,H83
6. Generate new invariant set: Yes
   Number of phases to use: 740
   Number of triples to use: 7400
   Number of negative quartets to use: 0
   Save invariants to file: ./ph8755.inv
7. Generate random trial structures: Yes
   Number of trials to generate: 1000
   Random number seed: 11909
   Starting atoms per trial: 74
   Save random trials to file: ./ph8755.trials
8. Trial processing information
   Number of trials to process: 64
   Beginning at trial number: 1
   Number of Shake-and-Bake cycles: 10
9. Exploit knowledge of heavy atoms: No
10. Refinement method: Parameter Shift
     Exploit knowledge of restricted phases: No
     Number of complete passes through phase set: 1
     Number of attempted phase shifts per pass: 2
         Phase shift used in pass #1: 90
11. Number of peaks to select: 74
12. Optional information storage
     Keep trace file containing Minimal Function values: Yes
     Store all final structures in file: No

Would you like to make any changes [n]:
Recognizing a Solution

Structure Name: ph8755

Number of Atoms: 74
Number of trials processed: 64
Lowest Rmin value: 0.244
Highest Rmin value: 0.471

Rmin range     Num in range

0.244 to 0.255  3 ***
0.256 to 0.267  0
0.268 to 0.279  0
0.280 to 0.291  0
0.292 to 0.303  0
0.304 to 0.315  0
0.316 to 0.327  1 *
0.328 to 0.339  2 **
0.340 to 0.351  2 **
0.352 to 0.363  5 *****
0.364 to 0.375  8 ********
0.376 to 0.387 10 **********
0.388 to 0.399  8 ********
0.400 to 0.411  4 ****
0.412 to 0.423 10 **********
0.424 to 0.435  4 ****
0.436 to 0.447  5 *****
0.448 to 0.459  1 *
0.460 to 0.471  1 *

Hit RETURN go back to main menu.
## Previously Known Structures

<table>
<thead>
<tr>
<th>Structure</th>
<th>Atoms</th>
<th>Space Group</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostaglandin $E_2$</td>
<td>25</td>
<td>$P1$</td>
<td>20%</td>
</tr>
<tr>
<td>Prostaglandin $F_1 \beta$</td>
<td>25</td>
<td>$C2$</td>
<td>8</td>
</tr>
<tr>
<td>Aldosterone</td>
<td>27</td>
<td>$P2_1$</td>
<td>14</td>
</tr>
<tr>
<td>9α-Methoxycortisol</td>
<td>28</td>
<td>$P2_1 2_1 2_1$</td>
<td>16</td>
</tr>
<tr>
<td>AZET</td>
<td>48</td>
<td>$Pca2_1$</td>
<td>21</td>
</tr>
<tr>
<td>Tetrahymanol</td>
<td>63</td>
<td>$P2_1$</td>
<td>3</td>
</tr>
<tr>
<td>APAPA</td>
<td>69</td>
<td>$P4_1 2_1 2$</td>
<td>1</td>
</tr>
<tr>
<td>Antibiotic A204A</td>
<td>71</td>
<td>$C2$</td>
<td>4</td>
</tr>
<tr>
<td>Emerimycin</td>
<td>74</td>
<td>$P1$</td>
<td>91</td>
</tr>
<tr>
<td>Isoleucinomycin</td>
<td>84</td>
<td>$P2_1 2_1 2_1$</td>
<td>12</td>
</tr>
<tr>
<td>Meso-valinomycin</td>
<td>84</td>
<td>$P1$</td>
<td>1</td>
</tr>
<tr>
<td>Synthetic Pyrrole</td>
<td>96</td>
<td>$P1$</td>
<td>33</td>
</tr>
<tr>
<td>Sea Urchin Peptide</td>
<td>115</td>
<td>$P2_1 2_1 2_1$</td>
<td>1</td>
</tr>
<tr>
<td>Hexadeca Isoleucinomycin</td>
<td>127</td>
<td>$P2_1 2_1 2_1$</td>
<td>2</td>
</tr>
<tr>
<td>Cholesterol Butanoate</td>
<td>132</td>
<td>$P2_1$</td>
<td>3</td>
</tr>
<tr>
<td>Gramicidin A</td>
<td>317</td>
<td>$P2_1 2_1 2_1$</td>
<td>0.3</td>
</tr>
<tr>
<td>Crambin</td>
<td>400</td>
<td>$P2_1$</td>
<td>4</td>
</tr>
<tr>
<td>Rubredoxin</td>
<td>500</td>
<td>$P2_1$</td>
<td>2.7</td>
</tr>
</tbody>
</table>
## Previously Unknown Structures

<table>
<thead>
<tr>
<th>Structure</th>
<th>Atoms</th>
<th>Space Group</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zomepirac (Weeks/Miller)</td>
<td>24</td>
<td>$P2_1/c$</td>
<td>2%</td>
</tr>
<tr>
<td>Tolmetin Sodium Salt (Weeks/Miller)</td>
<td>40</td>
<td>$P\bar{1}$</td>
<td>1%</td>
</tr>
<tr>
<td>Boc-Gln-D-EtA-Hyp-Ala-Phol (Ciszak/Smith)</td>
<td>47</td>
<td>$P2_1\bar{2}1\bar{2}1$</td>
<td>9%</td>
</tr>
<tr>
<td>7-KetoCholesterol (McCourt/Dorset)</td>
<td>58</td>
<td>$P2_1$</td>
<td>1%</td>
</tr>
<tr>
<td>Full-Retro Valinomycin (Langs)</td>
<td>78</td>
<td>$P1$</td>
<td>1%</td>
</tr>
<tr>
<td>Cholesteryl Myristate / Cholesteryl Pentadecanoate (McCourt/Dorset)</td>
<td>87</td>
<td>$C2$</td>
<td>1%</td>
</tr>
<tr>
<td>Ternatin_E (Miller/Langs)</td>
<td>105</td>
<td>$P2_12_12_1$</td>
<td>1%</td>
</tr>
<tr>
<td>Ternatin_D (Miller/Langs)</td>
<td>110</td>
<td>$P2_12_12_1$</td>
<td>2%</td>
</tr>
</tbody>
</table>
## Electron Diffraction Data (Known)

(with D. Dorset and M. McCourt)

<table>
<thead>
<tr>
<th>Structure</th>
<th>Atoms</th>
<th>Space Group</th>
<th>Success Rate</th>
<th>3-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copper Chloride</td>
<td>5</td>
<td>$P2_1/m$</td>
<td>21%</td>
<td>52%</td>
</tr>
<tr>
<td>Poly-Butene-1</td>
<td>8</td>
<td>$P2_12_12_1$</td>
<td>28%</td>
<td>37%</td>
</tr>
<tr>
<td>Poly-$\epsilon$-Caprolactone</td>
<td>8</td>
<td>$P2_12_12_1$</td>
<td>41%</td>
<td>15%</td>
</tr>
<tr>
<td>Copper Perchloro-phthalocyanine</td>
<td>16</td>
<td>$Cmm2$</td>
<td>Yes</td>
<td>0%</td>
</tr>
<tr>
<td>$n$-Paraffin</td>
<td>36</td>
<td>$Pca2_1$</td>
<td>Yes</td>
<td>50%</td>
</tr>
</tbody>
</table>
SnB – Tangent Refinement

1. Traditional Tangent Formula

2. Hull-Irwin Weights
   - Improvement for Small Structures
   - Worse for Larger Structures

3. Feedback

4. Optimal SnB Cycles: \( n/5 \)

5. Optimal Iterations: 1 or 2
## Cost-Effectiveness

<table>
<thead>
<tr>
<th>Structure</th>
<th>Atoms</th>
<th>S.G.</th>
<th>T.R.</th>
<th>P.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nordihydroquaiaretic Acid</td>
<td>22</td>
<td>$P2_1/c$</td>
<td>213.48</td>
<td>123.15</td>
</tr>
<tr>
<td>Prostaglandin $E_2$</td>
<td>25</td>
<td>$P1$</td>
<td>194.60</td>
<td>80.70</td>
</tr>
<tr>
<td>9α-Methoxycortisol</td>
<td>28</td>
<td>$P2_12_12_1$</td>
<td>101.44</td>
<td>37.43</td>
</tr>
<tr>
<td>3β-hydroxy-16-methyl-5,16-pregnadien-20-one</td>
<td>48</td>
<td>$P1$</td>
<td>405.3</td>
<td>209.3</td>
</tr>
<tr>
<td>Tetrahymanol</td>
<td>63</td>
<td>$P2_1$</td>
<td>3.01</td>
<td>2.91</td>
</tr>
<tr>
<td>Emerimycin</td>
<td>74</td>
<td>$P1$</td>
<td>65.51</td>
<td>46.41</td>
</tr>
<tr>
<td>Isoleucinomycin</td>
<td>84</td>
<td>$P2_12_12_1$</td>
<td>4.10</td>
<td>1.88</td>
</tr>
<tr>
<td>Meso-valinomycin</td>
<td>84</td>
<td>$P1$</td>
<td>1.03</td>
<td>0.67</td>
</tr>
<tr>
<td>Synthetic Pyrrole</td>
<td>96</td>
<td>$P1$</td>
<td>12.09</td>
<td>9.27</td>
</tr>
<tr>
<td>Ternatin</td>
<td>105</td>
<td>$P2_12_12_1$</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>Hexadeca Isoleucinomycin</td>
<td>127</td>
<td>$P2_12_12_1$</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>Cholesterol Butanoate</td>
<td>132</td>
<td>$P2_1$</td>
<td>0.43</td>
<td>0.30</td>
</tr>
<tr>
<td>Gramicidin A</td>
<td>317</td>
<td>$P2_12_12_1$</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Crambin</td>
<td>400</td>
<td>$P2_1$</td>
<td>0.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Genetic Algorithms

Initialize $P$, a population of $n$ randomly generated individuals.

Evaluate $F(i)$, $\forall i \in P$, where $F$ is the pre-specified fitness function.

while (another generation is required) do

Selection: Create $P_s'$ from $P$ based on $F$.

Crossover: Randomly mate members of $P_s'$, interchanging a contiguous substring for each of the $\frac{n}{2}$ such pairs, to create $P_c'$.

Mutation: From $P_c'$, randomly flip bits in a subset of the $n$ members to create the next generation $P$.

Evaluation: Compute $F(i)$, $\forall i \in P$.

end while
Genetic Algorithms

- Parallel Genetic Algorithms
- Demes (isolated subpopulations)
  Evaluation $\rightarrow$ Exchange $\rightarrow$
  Selection $\rightarrow$ Crossover $\rightarrow$ Mutation
- Combination of Fine- and Coarse-Grained
- Achieve Convergence
- Extremely Poor Performance
- Low Cost-Effectiveness
Simulated Annealing

begin

Initialize:
configuration \( x = x_0 \)
cooling parameter \( c = c_0 \)
objective function \( C = C_0 \)
cooling step \( m = 0 \)

repeat

repeat

Generate new configuration \( x'_i \) from \( x_i \)
Compute \( C'_i \)
if \( (C'_i - C_i < 0) \) then accept else
  if \( e^{(C_i - C'_i)/c} > random[0,1] \) then accept

if accept then \( x_i = x'_i \)
until Equilibrium

\( c_{m+1} = f(c_m) \)
\( m = m + 1 \)

until Convergence

end
Simulated Annealing

- Initial Temperature vs. Structure Size
- Rate of Cooling is Fixed
- Results Encouraging

<table>
<thead>
<tr>
<th>Structure</th>
<th>Atoms</th>
<th>S.G.</th>
<th>S.A.</th>
<th>P.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>9α-Methoxycortisol</td>
<td>28</td>
<td>$P2_{1}2_{1}2_{1}$</td>
<td>75.5</td>
<td>29.7</td>
</tr>
<tr>
<td>Emerimycin</td>
<td>74</td>
<td>$P_{1}$</td>
<td>69.8</td>
<td>38.3</td>
</tr>
<tr>
<td>Isoleucinomycin</td>
<td>84</td>
<td>$P2_{1}2_{1}2_{1}$</td>
<td>3.5</td>
<td>1.7</td>
</tr>
</tbody>
</table>
Fragment Recycling

CBT: 132 atoms, $P2_1$, 20-atom fragment
Base: 20-atom random trials yield 2.5% success

<table>
<thead>
<tr>
<th>RMS Å</th>
<th>Tangent Formula (RANTAN)</th>
<th>Shake-and-Bake ($SnB$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deformed RMS Model</td>
<td>Deformed RMS Model</td>
</tr>
<tr>
<td>0.0</td>
<td>19.80</td>
<td>100</td>
</tr>
<tr>
<td>0.1</td>
<td>7.20</td>
<td>86</td>
</tr>
<tr>
<td>0.15</td>
<td>2.20</td>
<td>50</td>
</tr>
<tr>
<td>0.2</td>
<td>0.15</td>
<td>12</td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Availability of $SnB$

- Cray C90 at Pittsburgh Supercomputing Center

- Cray T3D at Pittsburgh Supercomputing Center

- Thinking Machines Corporation CM-5

- Molecular Structures Corporation
Concluding Remarks

- Available / Unix

- Success Rate: 100% (6/94)

- Manual / Help Screens

- Parameter Tuning

- Space Group Tuning (e.g., P1)

- Evaluating Simulated Annealing

- Evaluating Tangent Formula

- Electron Diffraction Data