Bioinformatics in Drug Discovery for Tuberculosis

Thomas R. Ioerger
Department of Computer Science and Engineering
Texas A&M University
Tuberculosis

- caused by *Mycobacterium tuberculosis*
- 2-3 million deaths/year
- standard DOTS chemotherapy:
  - 2 months:
    isoniazid+rifampicin+ethambutol+
    pyrazinamide
  - 4 months: isoniazid+rifampicin

- H37Rv genome sequence (Cole, 1998) - 4.4 Mb, 3989 genes
Drug Resistance

- mono-resistant (INH)
- MDR (INH+RIF) “multi-drug resistant”, 440,000 cases/yr
- XDR (INH+RIF+streptomycin+fluoroquinolone) “extensively drug resistant”, 25,000 cases/yr
- prevalence increasing especially in South Africa, Russia/Eastern Europe, Asia
- causes: poor compliance? inadequate detection? higher transmissibility? HIV co-infection?

Survival of patients in South Africa with drug-resistant TB. (Gandhi, 2010)
Discovering New Drug Targets

• known drug targets:
  – isoniazid inhibits mycolic-acid biosynthesis (InhA)
  – rifampicin inhibits RNA polymerase (transcription)
  – trimethoprim, sulfamethoxazole, para-aminosalicylate inhibit folate biosynthesis (DHFR, DHPS, ThyA)

• persistence factors???

• an approach to discovering new targets
  – find inhibitor with high-throughput screening
  – select resistant mutants
  – use whole-genome sequencing to identify gene with mutations
high-diversity library with ~100,000 drug-like compounds

whole-genome sequencing

high-throughput screening

pick resistant colonies

alignment with SNP

structure with drug bound
Challenges of Whole-genome Sequencing

- **Illumina Genome Analyzer II**
  - sequence 8 strains in 24 hours
  - ~10 million “short reads” (50 bp each)
  - paired-end data
- **comparative genome assembly**
  - map reads onto a reference genome (like H37Rv)
  - look for differences:
  - “SNPs” – single nucleotide polymorphisms
  - “indels” – insertions/deletions
- **problems**
  - data quality (base call errors, GC-rich regions)
  - ambiguity (due to repetitive sequences)
  - indels (initially look like SNPs)
Figure 2. Histogram of coverage at 87 sites for CDC1551_{CSU} vs. CDC1551 which had an indel of 1-3 bp. The mean coverage over the whole genome was 70.4x.
Local Contig-Building Heuristic Search Algorithm

a) Depth First Search – susceptible to repeats

Key ideas:
1. pick a read upstream of indel site as “start” for contig and a read downstream as a “goal” node
2. use hash table to find candidate overlapping reads to extend contig
3. prioritize partial contigs by a combination of length of consensus and number of reads

BuildContig(site i, reads R, reference sequence G)
// wrapper routine that initializes parameters for LCB
rup = Δ
rdown = Δ
make initial contig consisting only of rup
PQ = (C)
return LCB(PQ, rdown, R)

LCB(priority queue PQ, target read rgoal, reads R)
// recursive routine that implements best-first search
C = PQ.pop() // select partial contig with minimum heuristic score H(C)
for each read s ∈ S
    extendContig(C, s)
if s = rgoal, return C’ // found read matching downstream; done
H(C’) = |seq(C’)| - |reads(C’)| // calculate heuristic score
PQ.insert(C’, H(C’)) // insert contig in sorted order
return LCB(PQ, rdown, R)

Figure 6. The LCB algorithm for building contigs.

b) Breadth-first Search – produces “thin” contigs

ACGGTTTGGCTACAGCATCTGGTGCACACAG
CGGACATGATCTGGTCAGCATACGACACTGAG
TCGGTTATGTGGCAGCAAGCAGACTGAGAAC
TCATACATGACTGGTGCACCAATGTGAGACT
AGCTTGAGATGGTGGTGCACCAATGTGAGACT
GACATCAGCATCTGGTGCACCAATGTGAGACT
TCGATCAGCATCTGGTGCACCAATGTGAGACT
AGCAGATGATCTGGTCAGCATACGACACTGAG
TCGGTTATGTGGCAGCAAGCAGACTGAGAAC
ACGGTTTGGCTACAGCATCTGGTGCACACAG
CGGACATGATCTGGTCAGCATACGACACTGAG
TCGGTTATGTGGCAGCAAGCAGACTGAGAAC
TCATACATGACTGGTGCACCAATGTGAGACT
AGCTTGAGATGGTGGTGCACCAATGTGAGACT
GACATCAGCATCTGGTGCACCAATGTGAGACT
TCGATCAGCATCTGGTGCACCAATGTGAGACT
AGCAGATGATCTGGTCAGCATACGACACTGAG
TCGGTTATGTGGCAGCAAGCAGACTGAGAAC
ACGGTTTGGCTACAGCATCTGGTGCACACAG
CGGACATGATCTGGTCAGCATACGACACTGAG
TCGGTTATGTGGCAGCAAGCAGACTGAGAAC
TCATACATGACTGGTGCACCAATGTGAGACT
AGCTTGAGATGGTGGTGCACCAATGTGAGACT
Figure 6. The LCB algorithm for building contigs.
Table 1: Recovery of 200 indels artificially introduced into H37RvCO genome. Each category represents 20 random mutations made in the reference genome, and the results report the number of those mutations re-discovered by mapping a set of reads against the modified genome and applying the LCB algorithm at sites with an indel signal of $\omega > 0.4$.

<table>
<thead>
<tr>
<th>indel length</th>
<th>example insertion</th>
<th>insertions recovered</th>
<th>example deletion</th>
<th>deletions recovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>111222: +G</td>
<td>20/20</td>
<td>4241489: -T</td>
<td>18/20</td>
</tr>
<tr>
<td>2</td>
<td>3867592: +GA</td>
<td>19/20</td>
<td>1544866: -GG</td>
<td>19/20</td>
</tr>
<tr>
<td>3</td>
<td>229382: +TGG</td>
<td>19/20</td>
<td>2321014: -GCT</td>
<td>18/20</td>
</tr>
<tr>
<td>5</td>
<td>3736341: +GCAAC</td>
<td>17/20</td>
<td>386791: -TGCAA</td>
<td>19/20</td>
</tr>
<tr>
<td>10</td>
<td>1081130: +TCAGACCAGA</td>
<td>18/20</td>
<td>3265177: -CGCAGCCGC</td>
<td>18/20</td>
</tr>
</tbody>
</table>
Table 2. Identifying indels in laboratory strains of H37Rv.

<table>
<thead>
<tr>
<th>coordinate</th>
<th>indel</th>
<th>ω score</th>
<th>cov before</th>
<th>cov after</th>
<th>ω score</th>
<th>cov before</th>
<th>cov after</th>
<th>ω score</th>
<th>cov before</th>
<th>cov after</th>
<th>ω score</th>
<th>cov before</th>
<th>cov after</th>
</tr>
</thead>
<tbody>
<tr>
<td>131177</td>
<td>4-G</td>
<td>0.141</td>
<td>19</td>
<td>94</td>
<td>0.001</td>
<td>22</td>
<td>70</td>
<td>0.827</td>
<td>5</td>
<td>32</td>
<td>0.442</td>
<td>12</td>
<td>52</td>
</tr>
<tr>
<td>234497</td>
<td>4-T</td>
<td>0.466</td>
<td>11</td>
<td>52</td>
<td>0.610</td>
<td>7</td>
<td>50</td>
<td>0.335</td>
<td>2</td>
<td>21</td>
<td>0.873</td>
<td>9</td>
<td>52</td>
</tr>
<tr>
<td>424323</td>
<td>4-C</td>
<td>0.999</td>
<td>14</td>
<td>74</td>
<td>0.914</td>
<td>14</td>
<td>56</td>
<td>0.758</td>
<td>3</td>
<td>19</td>
<td>0.017</td>
<td>19</td>
<td>60</td>
</tr>
<tr>
<td>1010207</td>
<td>4-G</td>
<td>0.942</td>
<td>15</td>
<td>76</td>
<td>0.999</td>
<td>8</td>
<td>58</td>
<td>0.758</td>
<td>4</td>
<td>43</td>
<td>0.917</td>
<td>12</td>
<td>54</td>
</tr>
<tr>
<td>1168718</td>
<td>4-T</td>
<td>0.097</td>
<td>18</td>
<td>69</td>
<td>0.010</td>
<td>19</td>
<td>53</td>
<td>0.441</td>
<td>8</td>
<td>27</td>
<td>0.262</td>
<td>15</td>
<td>52</td>
</tr>
<tr>
<td>1313339</td>
<td>4-C</td>
<td>0.735</td>
<td>11</td>
<td>49</td>
<td>0.917</td>
<td>11</td>
<td>41</td>
<td>0.565</td>
<td>2</td>
<td>14</td>
<td>1.000</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>1780588</td>
<td>4-G</td>
<td>0.424</td>
<td>9</td>
<td>9</td>
<td>0.680</td>
<td>7</td>
<td>30</td>
<td>0.478</td>
<td>2</td>
<td>2</td>
<td>0.897</td>
<td>12</td>
<td>38</td>
</tr>
<tr>
<td>2207592</td>
<td>4-C</td>
<td>0.993</td>
<td>16</td>
<td>113</td>
<td>0.993</td>
<td>12</td>
<td>85</td>
<td>0.895</td>
<td>3</td>
<td>32</td>
<td>0.854</td>
<td>10</td>
<td>72</td>
</tr>
<tr>
<td>2523208</td>
<td>4-CGG</td>
<td>0.756</td>
<td>13</td>
<td>33</td>
<td>0.775</td>
<td>7</td>
<td>33</td>
<td>0.098</td>
<td>6</td>
<td>20</td>
<td>0.703</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td>2525727</td>
<td>4-G</td>
<td>0.825</td>
<td>16</td>
<td>50</td>
<td>0.441</td>
<td>13</td>
<td>54</td>
<td>0.342</td>
<td>7</td>
<td>20</td>
<td>0.985</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>3580637</td>
<td>4-T</td>
<td>0.982</td>
<td>17</td>
<td>81</td>
<td>0.977</td>
<td>11</td>
<td>62</td>
<td>0.875</td>
<td>6</td>
<td>26</td>
<td>0.993</td>
<td>9</td>
<td>49</td>
</tr>
<tr>
<td>3590687</td>
<td>4-C</td>
<td>1.000</td>
<td>6</td>
<td>47</td>
<td>1.000</td>
<td>5</td>
<td>42</td>
<td>0.528</td>
<td>2</td>
<td>25</td>
<td>0.905</td>
<td>13</td>
<td>47</td>
</tr>
<tr>
<td>3862474</td>
<td>4-A</td>
<td>0.991</td>
<td>6</td>
<td>28</td>
<td>0.939</td>
<td>7</td>
<td>23</td>
<td>0.615</td>
<td>2</td>
<td>13</td>
<td>0.857</td>
<td>10</td>
<td>54</td>
</tr>
<tr>
<td>4095002</td>
<td>4-G</td>
<td>0.145</td>
<td>2</td>
<td>8</td>
<td>0.590</td>
<td>1</td>
<td>19</td>
<td>0.814</td>
<td>4</td>
<td>21</td>
<td>0.916</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>4400663</td>
<td>4-C</td>
<td>0.522</td>
<td>12</td>
<td>44</td>
<td>0.467</td>
<td>7</td>
<td>44</td>
<td>0.761</td>
<td>2</td>
<td>20</td>
<td>0.681</td>
<td>8</td>
<td>54</td>
</tr>
</tbody>
</table>

(Ioerger et al., 2010, *Journal of Bacteriology*)
Coumarin analogs

- collaboration with Sarah Stanley in Deb Hung’s lab (MIT)
- minimum inhibitory concentrations (MICs): B) 25μM, E) 12.5μM
- 4 resistant mutants selected in *M. tuberculosis* H37Rv
- 36 bp paired-end sequencing

<table>
<thead>
<tr>
<th>strain</th>
<th>cov.</th>
<th>SNPs</th>
<th>fadD32</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2</td>
<td>65.7x</td>
<td><strong>Rv0223c</strong>: V75G, <strong>Rv0538</strong>: T417R, <strong>recB</strong>: T312P, G311A, <strong>Rv1278</strong>: T282P, <strong>Rv1751</strong>: V354F, <strong>Rv2828</strong> (T141R), <strong>ppsC</strong>: 1 bp del in Gly1340, <strong>Rv3645</strong>: V149F</td>
<td>E120A (gcg)</td>
</tr>
<tr>
<td>B3</td>
<td>45.6x</td>
<td><strong>Rv1158c</strong>: del aa107-131 <strong>ppsA</strong>: H955P</td>
<td>F291L</td>
</tr>
<tr>
<td>E2</td>
<td>53.1x</td>
<td><strong>ppsC</strong>: 1bp del in Gly1340</td>
<td>E120V (gtg)</td>
</tr>
<tr>
<td>E1</td>
<td>51.7x</td>
<td>none</td>
<td>E120G (ggg)</td>
</tr>
</tbody>
</table>

- All four strains had a mutation in *fadD32*: an acyl-CoA synthetase involved in mycolic acid biosynthesis
- mutations validated to shift MICs by 25-30x in culture
Table 1: Drug-resistance mutations for inhibitors found by whole-genome sequencing as part of Integrated Methods for Tuberculosis program.

<table>
<thead>
<tr>
<th>compound (class)</th>
<th>parental strain</th>
<th># of mutants sequenced</th>
<th>depth of coverage</th>
<th>shared mutations (number of strains)</th>
<th>confirmed by recombineering?</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMTB-001 (quinolinedione)</td>
<td>H37Rv</td>
<td>1</td>
<td>144x</td>
<td>(scrubbed)</td>
<td>yes</td>
</tr>
<tr>
<td>IMTB-002 (thiadiazole)</td>
<td>M. smeg mc2 155</td>
<td>2</td>
<td>80-99x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMTB-003 (thiadiazole)</td>
<td>M. smeg mc2 155</td>
<td>2</td>
<td>93-95x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMTB-008 (coumarin)</td>
<td>H37Rv</td>
<td>2</td>
<td>46-66x</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>IMTB-009 (coumarin)</td>
<td>H37Rv</td>
<td>2</td>
<td>52-53x</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>IMTB-014 (adenosine analog)</td>
<td>H37Rv</td>
<td>3</td>
<td>130-150x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMTB-026</td>
<td>H37Rv</td>
<td>4</td>
<td>23-32x</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>IMTB-027</td>
<td>H37Rv</td>
<td>4</td>
<td>26-28x</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>IMTB-028</td>
<td>H37Rv</td>
<td>4</td>
<td>151-165x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMTB-029</td>
<td>H37Rv</td>
<td>4</td>
<td>159-174x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HET0016 (oximealdehyde)</td>
<td>H37Rv</td>
<td>4</td>
<td>114-133x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Mycobacterial Genome Database

Coverage Statistics (more detailed sequencing information on each individual site)

Coverage Statistics (tabular)

Generate catalog of polymorphic loci

Jump to Location/Gene: [Submit (examples: Rv0340, katG, 1057728)]

H37Rv [4262831-4262880]

H37Rv [4262881-4262930]

H37Rv [4262931-4262980]
Other projects in Collaboration with the Sacchettini lab

- Sequencing genomes of MDR and XDR clinical isolates from KwaZulu-Natal, South Africa (Ioerger et al, 2009, PLoS ONE)
- Target Identification for *M. tuberculosis*
  - funded by Bill and Melinda Gates Foundation
- Target Identification in *Staphylococcus aureus* and *Pseudomonas aeruginosa*
  - funded by NIH
- TB Structural Genomics Consortium
  - funded by NIH
  - solving structures of many essential proteins/drug targets
  - also sequencing genomes of *M. abscessus, M. fortuitum, M. thermoresistible*...
- discovery of malate synthase inhibitors
  - funded by TB GlobalAlliance
  - compound we designed has efficacy in preclinical trials (acute-phase mouse model), tested at GlaxoSmithKline
- high-throughput screening of mouse stem-cell knock-out cell lines that reduce infectivity of pathogens: rabies virus, botulism, *Brucella*
  - funding from DHS-DTRA
  - collaboration with Deeann Wallis (TAMU), TIGM
Acknowledgements

• Dr. James Sacchettini (Dept. Biochem/Biophys, TAMU)
• Dr. Inna Krieger, Gulcin Gulten, Yushan Chen, Kika Chavez (mutant selection)
• Dr. Xiaohua Chen (runs the sequencer)
• Dr. Joel Freundlich (Rutgers; medicinal chemistry)
• my group: Krishna Ganesula, Yicheng Feng, Purvaja Narayamaswamy, Michael DeJesus (sequence analysis)
• colleagues in the TB community: (David Sherman, SBRI; Eric Rubin, Harvard; Deb Hung, MIT; Bill Jacobs, Einstein College Med., New York; Valerie Mizrahi, South Africa...