SEARCHING THE HUMAN GENOME FOR SNAIL AND SLUG WITH DNA@HOME

Kristopher Zarns, Travis Desell, Archana Dhasarathy & Sergei Nechaev

Departments of Computer Science & Basic Sciences (Medical School)
University of North Dakota

http://csgrid.org/csg/dna/

September 2nd, 2015 - IEEE E-Science
DNA@HOME

DNA@Home is a volunteer computing project where anyone can volunteer their computer to aid in analyzing DNA sequences.

The project supports Windows, Linux and Apple computers, and currently has 1600 users and 3000 computers participating.

Visit us at: http://csgrid.org/csg/dna/
DNA@HOME

- Find protein binding sites using Gibbs sampling
- Use random walks (Markov chains) which result in sites distributed according to their actual probability of being the correct binding site

- Previously analyzed samples from *Mycobacterium tuberculosis* and *Yersinia pestis*
- Currently analyzing Human Genome (hg19) regions related to SNAIL and SLUG transcription factors
Why use Gibbs Sampling?

Turning a gene on by binding a protein causes new proteins to be produced, what binding sites will that activate?

Turning a gene off stops production of proteins, which other binding sites will that activate?

There are a lot of complicated relationships here.

We can use Gibbs sampling to identify binding sites without understanding these relationships.
Walks are parallel

Arrows are workunits

$S_x$ is the state at depth $x$

Workunits have fixed walk lengths

Samples are taken after Burn-In

Work is redistributed
User graphs for the last 60 days

Total users from July to September: 2100 to 3000

Active users from July to September: 800 to 1400
Graphs of host systems from July 2015 to September 2015.
USAGE STATISTICS

DNA@Home is a part of the citizen science grid, a collection of scientific applications run using BOINC

32 and 64 bit versions of Linux, OS X and Windows
~1600 volunteers
~3000 computers
5.3 teraflops in use
1300+ sequences being processed
The BOINC server is run on a virtual host providing significant cost savings over using a high performance computing cluster or supercomputing
THE CENTRAL DOGMA

- DNA => RNA => Protein
  - Protein binds to sequences near genes
  - Binding turns gene expression **on** to make RNA and then protein
  - Binding turns gene expression **off** to stop making RNA and protein
- Motifs indicate binding sites
- Snail and Slug are proteins that are **needed for organ development**
- They are also involved with the formation of some **cancerous tumors**
DATA SETS

To our knowledge, we present the largest scale use of Gibbs sampling for de novo detection of transcription factor binding sites

Test dataset size, 1000 base pairs per sequence
1. Small, 10 sequences
2. Medium, 100 sequences
3. Large, (372 for slug, 994 for snail)

Test number of motifs searched for
1. 1, 2, 3

Motif size was kept constant at 6

A sequence for claudin-7 with motif

> 'CLDN7 CHR17 7165764 7166764'

AGGTGGCTCGGAGGTGAGCCAGCAGGTGCGGGCGGCCAGAGGTGGG
GCGCACCTGAGTATATGTAGGGCGTCGGGGGCGCGGCCCGCGCC
CCGGGAGCGCGGGAGGGGGAGGCCAGGAGGCGGCAGGACAC
AGGATGACCTGACGCTACCGAAGGGACACTCACCTGAACCAGGAACTCG
TGCCCCACCCCTCCTTGGAACGTGAGGAGGAGGAGGAACGAC
CAGGGTGACGCCCTTCTGCTTGGCCCGGCTTTTGACCGCCGTGGGAG
GAGGCGGTTGCTCCTCAGGAGTCCCAAGCTCCGAGTCACCTCCAGCC
GACCTGCCCTCTCAGGAGACCCGGAGGCTCTTCTCAGGG
AGGTGCCGTGGAGCGTGCGGCGGCGGCGCCTGACCGGGCG
CGCCTTCTCTCCCGGCTTTGACCGCTGGGAGGCGGTGGCCCTCT
CTTCTGACCTCTGCTGCCTTGAGGAGGAGGAGGAGGAGGAGGAGGAGG
ATCCACAGCGGGGGAGCGCTGCGGGGCGCCCCTGACGCTGGGGCTCC
GCACCCTTGCTCCTCAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
AGGTGCCGTGGAGCGTGCGGCGGCGGCGCCTGACCGGGCG
CTGGGCTTGGACTGCCAATTGCTGTGCACTCTGTCACGCCTCACTG
GTCCTCCCTATAATCTGGCTCTCTAAGGTGGTC
OUR GOALS

• Approximate the source distribution of motifs
• Detect burn-in and convergence
  • Are larger dataset better?
  • Does number of motifs searched affect the outcome?
• Can parallel walks can be used to gather data?
• Will volunteer computing be sufficient, can the sampler scale?
• Will we need high performance computing for post processing?
• Do our results have biological importance?
## DATA SET RESULTS

### Burn-In and Convergence under 20K steps

<table>
<thead>
<tr>
<th>SLUG</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Motif</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>2 Motif</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>3 Motif</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>SNAIL</td>
<td>Small</td>
<td>Medium</td>
<td>Large</td>
</tr>
<tr>
<td>1 Motif</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>2 Motif</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>3 Motif</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>
KOLMOGOROV SMIRNOV (KS)

• Plot cumulative distributions
• Find largest difference
• Compare distribution shape
• Our goal is to approximate the source distribution
• The test finds the probability that consecutive portions of the walk share a common source distribution
INTERPRETING KS
BURN-IN AND CONVERGENCE

- Top - distance log scale, $10^{-3}$ to $10^{0}$
- Bottom – same source probability

 Slug Medium 2 Motifs

- Mean
- Min
- Max
- Standard deviation
KS RESULTS
SLUG 3 MOTIFS, SMALL VS MEDIUM

Size affects burn-in steps and convergence stability
KS RESULTS SMALL, 1 MOTIF VS 2

Number of motifs searched for affects stability

3 Motif small data set size images display similar results to 2 motif
KS RESULTS
SLUG LARGE 1, 2, 3 MOTIFS

Slug is most stable with 3 motifs or possibly an odd number of motifs.
KS RESULTS
SNAIL LARGE 1, 2, 3 MOTIFS

Snail is most stable with 2 motifs or possibly an even number of motifs.
INTERWALK METHOD

- Comparison of results across 1000 walks at every 10,000 steps
- Best case, Slug, small, 1 motif $\approx 20$GB of data
- Worst case, Snail, large, 3 motifs $\approx 500$GB of data
- The extreme data to compute and data to transmit ratios rule out BOINC
- C++ and MPI on a 32 node Beowulf HPC
- Compiled and run with mvapich2-x network optimization
- Node specification
  - Red hat enterprise Linux 6.2
  - Dual quad-core e5-2643 INTEL processors 3.3 ghz
  - 64GB of 1600 MHz ram
  - 2 mirrored RAID 146GB 15KRPM SAS drives
  - Private 56gigabit InfiniBand FDR 1-1 network
INTERWALK DISTANCE GRAPH

- Sample 100 of 1000 walks
- Find distance between every pair
- Calculate at every 10,000 steps
- This graph is Slug, 1 motif, medium size

Look for results of:
- Varying motif number
- Varying dataset size
INTERWALK – VARYING MOTIF NUMBERS

Slug Large 1 motif

Slug Large 2 motifs

Slug Large 3 motifs

Note the distance Y scale

Increasing the number of motifs causes each walk to contain multiple solutions

The improved convergence with multiple motifs suggest multiple motifs are present
INTERWALK - VARYING SIZE

Small dataset does not indicate convergence
Medium dataset converges at around 130k steps
Large dataset converges immediately
WHAT WERE WE LOOKING FOR?
WHAT DID WE FIND?

- Verify E-box motifs, CAGGTG and CACCTG
- These motifs are known to be present in the sequence set
- We wanted to verify that our approximate distribution highlighted these motifs
- We found motifs which overlapped the E-box motif
- Some of the E-box related motifs we found were known and predicted targets
- Others will need to be verified and may provide new therapeutic targets
## E-BOX MOTIF RESULTS

### SNAIL

<table>
<thead>
<tr>
<th>Count</th>
<th>Hit %</th>
<th>Gene</th>
<th>Motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>383</td>
<td>15.59</td>
<td>SGK3</td>
<td>caggtGGAGGGaccccc</td>
</tr>
<tr>
<td>385</td>
<td>22.89</td>
<td>RABAC1</td>
<td>cacctGGAGGGcttgcc</td>
</tr>
<tr>
<td>418</td>
<td>27.45</td>
<td>FAM195A</td>
<td>caggtGGAGGGccggc</td>
</tr>
<tr>
<td>459</td>
<td>13.39</td>
<td>RALGAPA2</td>
<td>caggtGGAAAAGataag</td>
</tr>
<tr>
<td>470</td>
<td>11.01</td>
<td>MYL7</td>
<td>cacctGGGAGAccgct</td>
</tr>
<tr>
<td>501</td>
<td>20.36</td>
<td>SLC22A17</td>
<td>caggtGGAGGGagggg</td>
</tr>
<tr>
<td>891</td>
<td>19.14</td>
<td>ESRP2</td>
<td>cacctGGGAAAAgggga</td>
</tr>
<tr>
<td>930</td>
<td>15.4</td>
<td>STX3</td>
<td>cacctGGGAAGGcgctc</td>
</tr>
<tr>
<td>208</td>
<td>26.6</td>
<td>TXNRD2</td>
<td>cacctGGGAAGGggggc</td>
</tr>
</tbody>
</table>

### SLUG

<table>
<thead>
<tr>
<th>Count</th>
<th>Hit %</th>
<th>Gene</th>
<th>Motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>423</td>
<td>13.27</td>
<td>FAM136A</td>
<td>cacctGCCGCCGggtg</td>
</tr>
<tr>
<td>423</td>
<td>13.69</td>
<td>WWOX</td>
<td>caggtGCCTCCGCtggg</td>
</tr>
<tr>
<td>423</td>
<td>14.33</td>
<td>TMEM116,ERP29</td>
<td>caggtGCCGCCGggtg</td>
</tr>
<tr>
<td>423</td>
<td>15.29</td>
<td>GNS</td>
<td>caggtGGCGGGGggtg</td>
</tr>
<tr>
<td>423</td>
<td>16.4</td>
<td>MYD88</td>
<td>caggtGGCGGGGggtg</td>
</tr>
<tr>
<td>423</td>
<td>19.38</td>
<td>BST2</td>
<td>caggtGGCGGGGggtg</td>
</tr>
<tr>
<td>423</td>
<td>19.94</td>
<td>COQ9,CIAPIN1</td>
<td>cagctGCCGCCGCCcggg</td>
</tr>
<tr>
<td>695</td>
<td>10.86</td>
<td>ZNF57</td>
<td>caggtGGGAAGGagggggg</td>
</tr>
<tr>
<td>962</td>
<td>13.67</td>
<td>PIN1</td>
<td>caggtGGGAAGGagggggg</td>
</tr>
</tbody>
</table>
CONCLUSION

• We are able to approximate the source distribution of motifs

• We have shown that we can detect burn-in and convergence
  • We verified that larger datasets are better for detecting these
  • We verified that the number of motifs searched affects outcome

• We have shown that parallel walks can be used to gather data

• We have shown that a mix of volunteer computing and high performance computing is ideal for our use case

• We have verified that our results have biological importance

• To our knowledge we have demonstrated the largest use case of Gibbs sampling for de novo detection of transcription factor binding sites
QUESTIONS?

• Thank you to our volunteers who make it all possible

• Thanks to Adam Burkholder of the National Institutes of Health for his work on the ChIP-Seq workflow

• Our work is supported by
  National institute of health, 5p20gm104360, to N.S. and A.D.
  University of North Dakota school of medicine faculty seed grant to K.Z., T.S., N.S., and A.D.
  University of North Dakota faculty seed grant. UND's office of Research Development and Compliance
  University of North Dakota research development and compliance travel grant to K.Z.

• We use BOINC: https://boinc.berkeley.edu/

• Visit us at: http://csgrid.org/csg/dna/

• Contact Kristoher Zarns at kzarns@gmail.com with questions