Detection of Syntenic Regions in Bacterial Genomes Through Statistical Clustering

SIEW-ANN CHEONG¹, PAUL STODGHILL², DAVID SCHNEIDER², CHRISTOPHER MYERS¹

¹Cornell Theory Center, Cornell University
²USDA/ARS, Ithaca

Interface 2007 on Systems Biology, May 24, 2007
Philadelphia, Pennsylvania
Research funded by USDA/ARS
Synteny and Evolution

- The goal of our group to understand the evolutionary roles of horizontal gene transfer and other large-scale rearrangements in shaping organization of bacterial genomes.

- Syntenic regions are homologous multigene regions in two or more genomes in which repertoire of genes are conserved, along with possible conservation of transcription direction and linear gene order.

- Detecting breaks in macrosynteny and microsynteny between closely related bacteria useful tool in unraveling mosaic structures within their genomes.
Computational Framework for Synteny Detection

I

genes
ORFs

non-overlapping blocks
statistically stationary segments

SEGMENTATION

II

sequence alignment
statistics comparison

SIMILARITY ANALYSIS

III

spatial clustering

SYNTENY ANALYSIS
The Jensen-Shannon Divergence


- For $x_1$ and $x_2$ modeled as Markov chains of order $K$ over quaternary alphabet $S = \{A, C, G, T\}$ with $S = 4$ letters, Jensen-Shannon divergence given by

$$
\Delta = \sum_{t \in S^K} \sum_{s=1}^{S} \left[ -f_{ts} \log \hat{p}_{ts} + f_{1,ts} \log \hat{p}_{1,ts} + f_{2,ts} \log \hat{p}_{2,ts} \right] \geq 0,
$$

where $t = (t_1, \ldots, t_K) \in S^K$ is shorthand notation, $f_{1,ts}, f_{2,ts}, f_{ts} = f_{1,ts} + f_{2,ts}$ are transition counts, and

$$
\hat{p}_{i,ts} = \frac{f_{i,ts}}{\sum_{s'=1}^{S} f_{i,ts'}}, \quad i = 1, 2; \quad \hat{p}_{ts} = \frac{f_{ts}}{\sum_{s'=1}^{S} f_{ts'}}
$$

are maximum-likelihood transition probabilities.
Mutations and Recombinations

Mutations and Recombinations

\[ x, N \text{ bases} \]

\[ n \ll N \text{ point mutations} \]

\[ x_1 \]

\[ n \ll N \text{ point mutations} \]

\[ x_2 \]

\[ y_1 \]

(i) deletion

\[ y_2 \]

(ii) insertion

\[ y_3 \]

(iii) reshuffle

\[ y_4 \]

(iv) inversion
Sequence Alignment and Statistics Comparison

- Without point mutations, $x_1$ and $x_2$ perfectly aligned. Also have identical $K$-mer statistics up to $K = N$.

- With $n \ll N$ point mutations, good alignment between $x_1$ and $x_2$. Very similar $K$-mer statistics up to intermediate $K$.

- Higher-order statistics strongly constrained by lower-order statistics, therefore, instead of sequence alignment, need only compare $K$-mer statistics at the few lowest $K$’s to establish homology.

- For recombination cases
  - (i) deletion and (ii) insertion, statistical similarity between $x_1$ and $y_1, y_2$ depends on segment deleted or inserted.
  - (iii) reshuffle, $K$-mer statistics between $x_1$ and $y_3$ similar.
  - (iv) inversion, complementary $K$-mer statistics of $y_4$ similar to $K$-mer statistics of $x_1$. 

$y_1$ (i) deletion

$y_2$ (ii) insertion

$y_3$ (iii) reshuffle

$y_4$ (iv) inversion
Hierarchical Single-Link Clustering

- Use Jensen-Shannon divergence as statistical distance between segments.

- Hierarchical clustering chosen because number and scales of syntenic regions not known beforehand. Hierarchical clustering tree for each $K$.

- Single-link separation between clusters:
  
  - clusters of homologous segments diffuse in statistics space, driven by random point mutations;
  
  - clusters diverging at different evolutionary times have different sizes;
  
  - two segments close together likely to have evolved from a common ancestor, even if both are far from center of homolog cluster.
Homology Within the Hierarchical Clustering Tree

increasing Jensen–Shannon divergence
Pilot Study

- Complete genomes of three *Pseudomonas syringae* strains. Plant pathogens.

<table>
<thead>
<tr>
<th>strain</th>
<th>N (Mbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC3000</td>
<td>6.4</td>
</tr>
<tr>
<td>1448A</td>
<td>5.9</td>
</tr>
<tr>
<td>B728A</td>
<td>6.1</td>
</tr>
</tbody>
</table>

- **Stage I**: Segmentations obtained using optimized recursive Jensen-Shannon segmentation scheme [Cheong *et al.*, in preparation].

- **Stage II**: Cluster at $K = 0, 1, 2, 3$ but not higher, because typical segments are only 5000 bp long.

- **Stage III**: $K = 0$ hierarchical clustering tree not sufficiently discriminating, but from $1 \leq K \leq 3$ hierarchical clustering trees identified:
  - long syntenic region between DC3000 and B728A; and
  - large number clustering events between paralogous segments containing mobile IS elements.
Syntenic Region Identified

DC3000

B728A
Feedback Between Stages

- Few long syntenic regions identified from hierarchical clustering trees, because syntenic regions segmented differently in different genomes as a result of context sensitivity problem [Cheong et al., in preparation].

- Stage III → Stage I: Work with two terminal segmentations per genome: standard (S) and fine (F). (Testing robustness of clustering to segmentation.)
  - cluster S segments against S segments: mask out syntenic regions detected at this level;
  - for remaining segments, cross cluster F segments against S segments: mask out syntenic regions detected at this level; and
  - for remaining segments, cluster F segments against F segments: identify syntenic regions detected at this level.

- Expect cross clustering to detect most, if not all, syntenic regions present.

- Hyperfine segmentation and fine-hyperfine segmentation if necessary.
Conclusions

• Described the three stages in the general framework for synteny detection.

• **Stage I: Segmentation.** Use statistically stationary segments.

• **Stage II: Similarity Analysis.**
  
  – Recasted sequence alignment problem as statistics comparison problem of statistically stationary segments.
  
  – Devised hierarchical single-link clustering of segments whose pairwise distance is their Jensen-Shannon divergence.
  
  – Explained how homologous segments cluster at small Jensen-Shannon divergence, and how syntenic blocks emerge as chains of clusters.

• **Stage III: Synteny Analysis.**
  
  – Pilot study on three *P. syringae* strains demonstrated feasibility of statistics comparison method for synteny detection.
  
  – Followup study: cross clustering between standard and fine segmentations of genomes.