Theme A: Watermarking for Provenance

Goal

- To design and develop novel computational methods based on genomic watermarking for robust and high resolution provenance for produce.

Technical Challenges

- Using intrinsic genomic variability, combined with combinatorial pooling for watermarking
- Designing extrinsic barcodes for robust and accurate watermarking

Computational Challenges

Extrinsic Barcodes

- Designing optimal barcodes that allow for rapid and inexpensive detection, traceback, and audit of food sources
- Minimizing barcode cost by minimizing perturbations to non-coding regions of DNA
- Robust coding through to distribution of codes over disparate parts of the DNA
- Inexpensive detection through shallow reads and mapping, which in turn relates to the repeat complexity of selected regions of the genome
- Accurate traceback using barcodes that are maximally distant from each other

Computational Challenges

Intrinsic Barcodes

- Use of intrinsic genomic features (the most prominent being simple sequence repeats (SSRs)) for unique signatures.
- Combining SSRs with combinatorial pooling to achieve desired level of specificity.
- Significant additional challenges in modeling, deconvolution, and sampling.

Technical Approach

Extrinsic Barcode Requirements

- Induce silent mutations, i.e., no changes to the phenotype
- Minimize change to DNA, so that the scheme is practical and cost-effective
- Chance of a random occurrence of the watermark is low
- Minimize the impact of recombination and/or cross-pollination

Uniqueness Properties of Watermarks

- The uniqueness property of a sequence watermark implies that *if the watermark is embedded as a subsequence in a particular genomic region, the probability of observing this watermark by chance is very low.*
- The watermark $\mathcal{W}$ occurs as a subsequence in text T if

$$T_{i_1} = w_1, \; T_{i_2} = w_2, \; \ldots, \; T_{i_m} = w_m.$$

  with additional distance constraints that $i_{j+1} - i_j \leq d_j$.
- The $I = (i_1, \ldots, i_m)$-tuple is called a position and $\mathcal{D} = (d_1, \ldots, d_m)$ constitutes the constraints.

Uniqueness Properties: Existing Results

Let $O_n(\mathcal{W})$ be the number of occurrences of watermark $\mathcal{W}$ in T.

Mean and Variance (IID)

$$\begin{aligned} \mathbf{E}[O_n(\mathcal{W})] &= nP(\mathcal{W}) \prod_{i=1}^{m} d_i + O(1), \\ \mathbf{Var}[O_n(\mathcal{W})] &= n\sigma^2(\mathcal{W}) + O(1) \text{ where } \sigma^2(\mathcal{W}) \text{ can be computed explicitly.} \end{aligned}$$

Central Limit Theorem (IID)

$$\Pr\left\{\frac{O_n - \mathbf{E}[O_n(\mathcal{W})]}{\sigma(\mathcal{W})\sqrt{n}} \leq x\right\} \sim \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

Large Deviations (IID)

We have a local large deviation $\Pr\{O_n(\mathcal{W}) = a\mathbf{E}[O_n]\} \sim \frac{1}{\sigma_a\sqrt{2\pi n}} e^{-nI(a)+\theta_a}$ where $I(a)$ can be explicitly computed, and θ_a is a known constant.

Uniqueness: Unresolved Questions

- DNA sequences are commonly modeled as Markovian sources. Extensions of our results to Markovian (and more general) sources pose interesting questions with wide applicability.
- Existing results consider distance constrained subsequences. The constraints in our applications correspond to synonymous substitution and/or to specific regions of the DNA sequence. Modeling these constraints and deriving associated results is unresolved.
- The subsequences (watermark characters) in our application are constrained (synonymour substitutions). Analyses for this constrained class of watermarks is an open question.

Finding Thresholds for Watermarks

- If false identification of watermarks is to be avoided, the problem is one of finding a threshold: $\alpha_0 = \alpha_0(\mathcal{W}; n, \beta)$ such that (say) $P(O_n(\mathcal{W}) > \alpha_{th}) \leq \beta (= 10^{-5})$.
- In the context of our previous result(s), it follows that

$$\alpha_{th} = nP(\mathcal{W}) + x_0(\beta)\sigma(\mathcal{W})\sqrt{n}, \quad \beta = \frac{1}{\sqrt{2\pi}} \int_{x_0}^{\infty} e^{-t^2/2} dt \sim \frac{1}{x_0} e^{-x_0^2/2}.$$

- We will derive results for Markov models, with real-world constraints for reliable thresholds.

Constructing Minimal Watermarks

- To construct a minimal watermark, for a given β, we find α_{th} such that

$$P(O_n(\mathcal{W}) > \alpha_{th}) \leq \beta, \quad \text{where} \quad \alpha_{th} = nP(\mathcal{W}) \prod_i d_i + x_0 \sigma(\mathcal{W}) \sqrt{n}$$

  where α_{th} and x_0 are defined on the previous slide.
- To answer this, we need to solve the following problem

$$\arg \min_{m, d_i} P(O_n(\mathcal{W}) > \alpha_{th}) \rightarrow 0.$$

- This problem is hard in the general case. We will use a formulation based on a De Bruijn graph abstraction for solving this problem.

Intrinsic Watermarks Using SSRs

- In the short sequence repeat problem, we ask: what is the minimal set of short sequence repeats (e.g., if you have a sequence ”ATA” repeating 10 times), that this is highly likely to be unique.
- We can solve this problem using the same machinery as above, in particular, the de Bruijn graph.
- We can also generalize the subsequence problem to sets of subsequences. In this case we have a set of words, say $\mathcal{W} = \{w^1, \ldots, w^N\}$, and we ask how many times their occur as subsequences.
- This analytical machinery provides us with the tools for characterizing the use of SSRs as watermarks.

Robustness of Watermarks

- In the robustness problem, we seek a solution to the following interesting problem: given a watermark, what is the number of characters one would have to flip to erase the watermark.
- This is formulated as a deletion channel problem or as the trace reconstruction problem.
- A deletion channel with parameter κ, a deletion vector, takes a sequence $x := x_1^n = x_1 \cdots x_n$ where $x_i \in \mathcal{A} = \{A, T, C, G\}$ as input and deletes each symbol in the sequence independently with probability determined by κ.
- The trace reconstruction problem is related to the deletion problem. We ask, how many copies N_n of the output deletion channel we need to see, until we can reconstruct the input sequence with high probability.

Robustness of Watermarks

- We will develop practical solutions to the robustness question in the context of DNA sequences.
- Consider w^1 as a watermark that is unique. That is,

$$P(O_n(w^1) > \alpha_{th}^1) \to 0.$$

- We will seek a word w^2 as an output of a deletion channel such that

$$P(O_n(w^2) > \alpha_{th}^1) > \beta > 0.$$

- We will derive techniques for characterizing this probability for Markov sources, for random, as well as adversarial perturbations, as well as modeling biological processes such as recombinations, inversions, and point mutations.