Genome Sequencing

Machine Learning for Big Data

Seminar by
Karishma Agrawal
(U10CO013)

Guided by
Dr. M.A. Zaveri
Associate Professor, COED, SVNIT
Agenda

- Genome
- Impact on Statistics
- Dependency Structure of Genome
- Hidden Markov Model
- Trivial HMM Parallelization
- Trivial HMM Parallelization with Sequence Partitioning
- Genome Sequencing
- Genetic Disease
- Evolution
- Conclusion
Genomes

- Your body is made up of about one hundred, million, million cells.
- Each of these cells has a complete set of instructions about how to make your cells, their components and their components' components. This set of instructions is your genome.
- Your genome is made of a chemical called DNA.
- Genome would contain 23 chromosomes.
- Hence, the DNA contains over 3.2 billion base pairs, total which fits into a cell nucleus the size of a pinpoint.
Coding and Non Coding DNA

- The content of the human genome is commonly divided into coding and noncoding DNA sequences.
- **Coding DNA** is defined as those sequences that can be transcribed into mRNA and translated into proteins during the human life cycle; these sequences occupy only a small fraction of the genome (<2%).
- **Noncoding DNA** is made up of all of those sequences (ca. 98% of the genome) that are not used to encode proteins.
- Because non-coding DNA greatly outnumbers coding DNA, the concept of the sequenced genome has become a more focused analytical concept than the classical concept of the DNA-coding gene.
Human Genomes

- In 2000 the first draft of the Human Genome was reported
  - Took 10 years to complete
- In 2012 the UK government announced plans to sequence 100,000 Genomes
  - Takes around 2 days to sequence a genome
- Alongside this UK Biobank is storing a detailed record of molecular features in blood, and outcome data (phenotypes) on 500,000 Individuals
- Within 3 to 5 years it will be routine to have all cancers sequenced in the UK, as well as the patient DNA
Impact on Statistics

- Such developments will have a huge impact on statistics and machine learning
- We will require new methods that can
  - scale to Big data
  - deal with heterogeneity and loosely structured data objects on different scales (genomes to eHealth records)
  - provide robust inference (under known model misspecification)
  - on mixed variable types
- To do this we will also need to exploit advances in computer hardware to allow us to develop increasingly richer classes of models
  - where the hardware structure is included within the design stage of the method
  - e.g. MapReduce, GPGPUs, ....
Dependence Structure in Genomes

The genome exhibits complex dependence. These dependence structures are a product of the interplay between cell division processes.

- Mutations introduce new variations into a population of individuals (i.e. a pool of genes).
- Recombination shuffle the genomes between each generation. Each recombination event splits parental chromosomes and combines them to produce child chromosomes. Recombination introduce independence between positions along the genome.

To formally model population and sequential dependence requires graphical models whose structures prevent computation.

- Hence we require simplifying models that capture the major sources of dependence and allow for computation.
- Perhaps the most important simplifying structure in statistics is the notion of Markov conditional independence.
What is Hidden Markov Model?

Hidden Markov models (HMMs) are a class of statistical models for sequential data with an underlying hidden structure.

HMMs are defined by

- A number of unobserved “hidden” states \( S = \{1, \ldots, N\} \) and Markov transition probabilities \( \Pr(S_{t+1}|S_t) \) that defines rules for state transitions.
- A possible emission or observations \( Y_i \) and emission probabilities \( \Pr(Y_i|S_i) \) for each state.
- An initial distribution \( \pi = \Pr(S_1) \)

The state sequence \( S \) then forms a Markov Chain:

- Such that the future \( S_{t+1}, \ldots, S_T \) does not depend on the past \( S_1, \ldots, S_{t-1} \), given the present \( S_t \).
Hidden Markov Models (HMMs)

- The Markov property is key to the success of HMMs. Dependencies are represented as edges and conditional independences as missing edges in the graph representation.

Figure: HMM depicted as a directed graphical model for observation $e_1, \ldots, e_T$.

- The joint distribution for HMMs is written as:

$$\Pr(y_1, \ldots, y_T, s_1, \ldots, s_T) = \Pr(s_1)\Pr(y_1|s_1)\prod_{t=2}^{T}\Pr(y_t|s_t)\Pr(s_t|s_{t-1})$$
There are **three basic problems** which can be efficiently solved using HMM.

- **Evaluation Problem** - how do we compute the probability of an observation $Y = \{Y_1, \ldots, Y_T\}$ given a parameterized HMM?
- **Decoding Problem** - how do we find the optimal state sequence $S = \{S_1, \ldots, S_T\}$ corresponding to an observation given a parameterized HMM?
- How do we estimate the **model parameters** to maximize the probability of an observation sequence?

**Solutions**

- Using the **Forward Algorithm** we can compute the probability of an observation, thus solving the 1\textsuperscript{st} problem.
- The **Viterbi Algorithm** finds the most probable (MAP) state sequence, maximizing the $\hat{s} = \arg \max_s \Pr(S_{1\ldots t} | Y_{1\ldots T})$
- The **Forward- Backward Algorithm** computes the posterior marginal probabilities $\Pr(S_t | Y_1, \ldots Y_T)$ for each state at every $t$.

All three algorithms have computation cost $O(N^2T)$, so linear in sequence length $T$. 
Efficient HMM Algorithms with Dynamic Programming

Figure: A single computation step of HMM algorithms.
Each step in the forward recursion means filling in a cell of a dynamic programming table:

\[
\alpha(s_t) = Pr(y_t|s_t) \sum_{s_{t-1} \in S} Pr(s_t|s_{t-1}) \alpha(s_{t-1})
\]
Trivial HMM Parallelization

The parallelization of HMM algorithms is straightforward (in theory) over multivariate emissions and over the state-space:

- calculations corresponding to different observations are parallelizable and is perfectly suited even for distributed computation.
- the standard HMM algorithms all repeat the same operations for each state $s_t$. Moreover, the calculation of the $(s_t)$ values are independent and hence the calculations are suitable for parallelization.
HMM Parallelization with Sequence Partitioning

Genetic datasets are generally large and the length of sequences is much greater than the state space ($T \gg N$).

The algorithm works by partitioning the sequence into blocks:

- assume the sequence is partitioned into $B$ blocks each of length $T_b$
- the starting (e.g. $\alpha(s_{kT_b})$) values are not available for each block at the beginning of each recursion
- hence run the algorithm $N$-times, from each possible state, each time conditioning on a different starting state. Naturally these sets of $N$ conditional runs may be also run in parallel.
- when the computation is done for all blocks, they can be merged sequentially updating each conditional run with its corresponding starting value
Forward-Backward algorithm, serial and parallel versions

Figure 2. The traditional forward algorithm, as described by Rabiner [1]. The rectangles represent matrices and vectors. The black lines denote dependencies. The top row is the $C_t$ matrices.

Figure 3. Using parallel reduction on the forward algorithm. The rectangles represent matrices and vectors. The black lines denote dependencies, while the horizontal dotted ones denote synchronization. The top row is the $C_t$ matrices.

Figure: Serial and Parallel representation of FB algorithm (Neilsen & Sand, 2011).
Genome Sequencing

The International Human Genome Project used a top down approach:

- They made a map of the whole genome and divided the chromosomes up between the centers.
- Because sequencing machines could only read a few hundred DNA letters at a time, the sequencing centers broke the chromosome down into tinier parts.
- The ends of these tiny pieces were then sequenced so that the information would built up to finally give researchers the complete sequence of entire chromosome.
DNA extraction

DNA fragmentation

Clone into Vectors

Transform bacteria, grow, isolate vector DNA

Sequence the library

Assemble contiguous fragments
Why is Genome Sequencing Important?

- The human genome has a surprising tendency to grow, shrink or otherwise rearrange itself. This so-called structural variation is the cause of 'genomic disorders' but also provides the raw material needed by evolution.
- But in finding out more about how genomic rearrangements occur, scientists are clarifying diagnoses and opening up the prospect of new therapies. At the same time, they are shedding light on that most extraordinary of processes - the evolution of humankind.
Genetic Diseases

- Genetic diseases often arise from simple changes to the coding region of a gene-altering the protein made by that gene. The disease arises because the protein does not work as it should do.
- Some genomic conditions also affect coding regions. A translocation, for example, can end up fusing genes together, creating an entirely new protein. More often, genes are lost entirely, if a chunk of chromosome is deleted.
- But sometimes a genomic rearrangement can have more subtle effects, leaving a coding sequence unchanged - so a normal protein is produced - but altering where and when it is made.
- Moreover, copy number variation has been found to contribute to several common diseases, including Parkinson's disease, Alzheimer's disease and autoimmune disorders such as rheumatoid arthritis, type 1 diabetes and Crohn's disease. How exactly it contributes to common disease is unclear but it is clearly significant.
- Swapping an A for a T in a gene for hemoglobin - the protein in our blood that carries oxygen around the body - causes a serious disease called sickle cell anemia.
Cure for Genetic Diseases?

- Although there are no specific treatments for most genomic disorders, *diagnosis* is still valuable. It can allow doctors to predict how children are likely to develop and which health problems they are likely to face, so symptoms can be managed better. It can also enable families to prepare more effectively for the future. There is also a *psychological benefit* for families to have an identified cause of their children's problems.

- There is an *entirely new approach*, in theory applicable to any genetic condition in which the product of a single mutant gene interferes with normal developmental processes. If the *mutant protein can be eliminated*, the remaining normal gene may be sufficient to direct normal development.
Copy number variation is particularly important to evolution as the duplication of genes is thought to be a key driver of evolutionary change - the original gene can maintain its function while the duplicated version takes on a new role. Indeed, many gene families exist, the result of ancient gene duplications.

One intriguing finding is that several genes amplified in humans code for parts of the centrosome, a key structure in cell division. Mutations in several centrosomal proteins cause microcephaly - very small brain size - and changes in a related protein have been linked to mental retardation. Such findings hint at an evolutionary link between cell division, brain size and the growth of human cognitive powers.
Conclusion

- Medical genetics and genomics will produce **vast data sets** over the next few years
- We need statistical methods that can scale to handle
  - dimension
  - Heterogeneity
  - model misspecification
- To do so, exploiting parallel computation within the algorithm design will be key, both for model development and model fitting
Q&A?

Thank You