

Finding Molecular Signature of Prostate Cancer: An Algorithmic Approach

Patrick O. Bobbie*
Renee Reams#
Sandra Suther###
C. Perry Brown##

*Southern Polytechnic State University
Department of Computer Science
1100 S. Marietta Parkway, Marietta, Georgia, USA
pbobbie@spsu.edu

College of Pharmacy and Pharmaceutical Sciences
Basic Sciences Division
Institute of Public Health
Florida A&M University
Tallahassee, Florida, USA
[renee.reams, sandra.suther, perry.brown]@famu.edu

*Abstract*¹²³

African American men in America are 65% more likely to develop prostate cancer and are two to three times more likely to die of the disease than Caucasian men [1]. While research indicates a strong hereditary link in prostate cancer, it has been difficult to pin down the exact genes involved in this disease. Many men develop prostate cancer later in life, and some men may die before prostate cancer even develops, making it difficult to trace the disease through generations of families [2]. Current research efforts in the testing and detecting of various forms of cancer such as breast and prostate cancer have taken new directions [3]. This paper addresses the genetic predispositions or biomarkers that lead or contribute to such diseases, the application of statistical approaches to researching the biomarkers, and a proposed validation process to mitigate the problem using data mining and high-performance computing techniques.

Keywords: Prostate Cancer, Hidden-Markov Model, DNA-methylation, CG-islands, CAG-repeat

1. Introduction

Today's tools for screening and diagnosing cancer are based on rigorous mathematical models and algorithms that take the biological structure and process into consideration. Cancer-directed research efforts that used to focus on protein biomarkers are now being refocused on a process called DNA methylation, especially for genes that cause prostate cancer. Because these methyl groups attach to the gene, they really don't change the gene sequence. Biologically, the least frequently occurring dinucleotide in many genomes is the 'CG' (in the ATCG sequence) because the C is easily methylated and as the resulting methyl-C has a tendency to mutate to T [4]. However, the mutation is often suppressed in regions of genes called CG-islands in a DNA sequence, making the CG rather appear frequently in the islands. But when DNA methylation occurs and consequently has the potential to influence the gene's function or behavior, the detection of aberrant or abnormal methylation offers a promising clue to early detection of the disease's onset and process inside the CG-islands.

The question is how does one proceed to first define and find these CG-islands (targets for methylation) algorithmically or computationally? Second, can this approach be validated to test its effectiveness? Below

¹ This work was supported in part by a grant from U.S. National Institute of Health, grant # NIH/RCMI G-1203020

² This work was also supported in part by a grant from U.S. DoD, grant # W81XWH-04-1-0326

³ This work was also supported in part by a grant from U.S. NSF, grant # CISE/EIA-0219547

we present a synopsis of the Hidden Markov Model (HMM) statistical algorithm for finding such CG-islands. We propose a mapping onto a 16-node high-performance computing system for computing the likelihood of finding the biomarker in DNA datasets of African Americans and Caucasians. The ongoing experiment seeks to also validate the effectiveness of the statistical procedure.

2. Hidden Markov Models

Hidden Markov Models (HMM) are useful as a machine learning (ML) approach in bioinformatics. Typically, a ML algorithm uses training data for gaining important insights about the (often hidden) parameters or carcinogenic factors. Once the algorithm is trained, it can then be applied to a target/test data (newly found or suspected data) in order to discover, confirm, or gain knowledge about the ‘hidden’ parameters in the test data.

The larger the training data the richer the knowledge gained from the trained data; and the better the effectiveness or accuracy of the algorithm. Generally, HMM approach learns some unknown probabilistic parameters from training data and uses the parameters in the framework of dynamic programming (or other techniques) to find the best explanation for the experimental or test data.

3. HMM and CG-islands

Computationally, a HMM can be considered as an Abstract Machine (or finite-state-machine) that produces a sequence of output symbols (from a select alphabet, Σ), on given input sequence or states, Q . The machine operates in discrete steps such that at each step, the machine is in a ‘hidden’ state, from among the k possible states. At each step, the machine decides on the next step and what output symbol to produce. Thus, the machine chooses randomly from the k states and randomly from the output alphabet, Σ , at each step. The selection of the next state and the alphabet to emit are respectively characterized by some probability distributions, $A_{|Q \times |Q|}$ and $\epsilon_{|Q \times |\Sigma|}$.

Formally, a HMM is defined as:

$M(\Sigma, Q, A, \epsilon)$, where

Σ = alphabet of output symbols

Q = set of states, each of which produces an output symbol

$A = (a_{kl})$, an $|Q| \times |Q|$ matrix of transition probabilities of moving from state k to l

$\epsilon = \epsilon_k(b)$, an $|Q| \times |\Sigma|$ matrix of probabilities of emitting symbol ‘ b ’ in state k

In practice, an investigator can observe the emitted symbol (or outcome) but the state at any step is ‘hidden’ – hence the ‘Hidden’ Markov model. The goal is for the investigator (searching routine) to infer the most likely state of the HMM by analyzing the sequence of emitted symbols (or DNA sequence). An application of the HMM to the CG-islands in genomic text is to define HMM profiles, which allow comparison and alignment of each sequence against the other sequences in the dataset using dynamic programming (DP) algorithms. Thus, given a family of functionally related biological sequence, one can search for new members of the family from a database using pair-wise alignments between family members and sequences in the database. Following DP technique, this search proceeds by focusing on a subset of the database and finding a match that satisfies an ‘acceptable’ criterion and proceeding to search the rest of the database using the matched subset for the next iteration.

4 The Model

An HMM can be defined by the following grid.

Σ (output)	$ x_1, x_2, \dots, x_n $
Q (states)	$ \pi_1, \pi_2, \dots, \pi_n $
\mathcal{E} ($p(x_i \pi_i)$)	$ \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n $
A ($p(\pi_i \rightarrow \pi_{i+1})$)	$ a_1, a_2, \dots, a_n $

In general, the search proceeds such that at each step, the transition probabilities are computed until the maximum (optimal) value, or ‘acceptable’ criterion is obtained. The following formulation results [4]:

$$P(x|\pi) = \max_{\text{forall } \pi} [p(\pi_0 \rightarrow \pi_1) \cdot \prod_{i=1}^n p(x_i | \pi_i) \cdot p(\pi_i \rightarrow \pi_{i+1}) = a_{\pi_0, \pi_1} \cdot \prod_{i=1}^n \varepsilon_{\pi_i}(x_i) \cdot a_{\pi_i, \pi_{i+1}}]$$

However, since Π is not known (as assumed in most cases), the problem becomes a Decoding Problem [5]. That is, the question becomes: What is the ‘best’ sequence of states that matches given/observed outcome? What is the best (max) $\text{Prob}(X|\Pi)$?

5 Computing System Configuration and Models

To compute the HMM probabilities, the HMM-based public domain software, *hmmcr*, is parallelized and deployed on a 16-node SGI O3800 high-performance computer for a large number of DNA datasets. The goal is to find the best classifications or clusters of CG-islands or biomarkers that have very high probability of becoming targets for methylation – thus, CG-islands which could become targets for the attachment of methyl group to the gene. Fig.1 depicts the architecture of the computational platform.

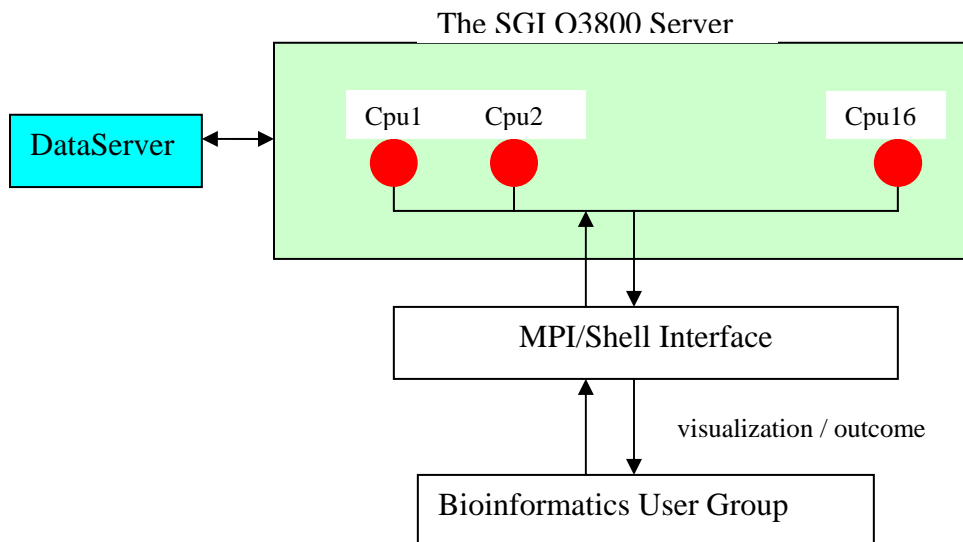


Fig 1. The HMMER System Configuration

6 The Computational Model/Process:

In the experimental setup, each CPU in the configuration (of the O3800) runs a copy of the *hmmer* code and a fragment of the DNA sequence data using the directives in the MPI/Shell interface code. The fundamental model of computing is the SPMD (Single Program, Multiple Data) since each CPU runs the same *hmmer* binaries. However, when each CPU searches a fraction of the sequence data, local ‘optimal’ values are achieved. Then, using a *tree-percolation* technique, the subsequences (or CG-islands) can be merged and re-searched/computed by a maximum of $k = n/2^i$ nodes at each i^{th} iteration, where $i=1, 2, 3, 4$; and $n=16$. Since there are 16 nodes on the O3800, there would be a maximum of k nodes in the parallel processing stages as depicted in Fig.2.

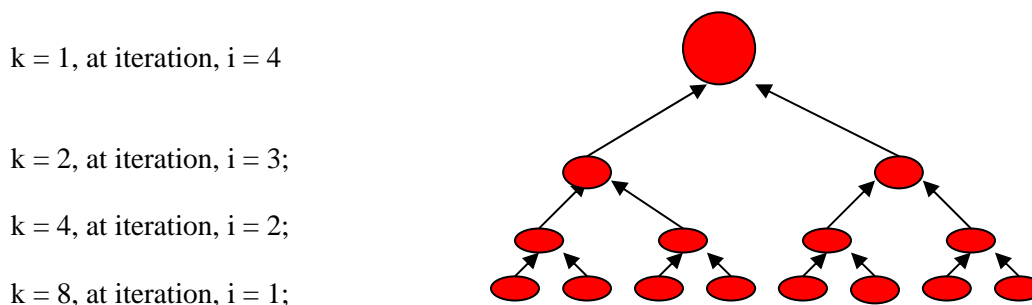


Fig. 2: *Tree-Topology of the Parallel Computation*

After each iteration, partial results are percolated up (the underlying tree) until the top level (or iteration 4) when the master CPU produces the maximum probability value (or ‘acceptance’ criterion). The resultant cluster will yield the corresponding subsequences that contain the most significant biomarkers.

7 Experimental & Validation Process

Research has shown that the “CAG repeat” polymorphisms in the androgen receptor gene are the underlying cause of prostate health disparity in African American males [1]. The CAG repeat androgen receptor polymorphism is associated with increased risk for prostate cancer and influences disease outcome and resistance to therapies in patients with short (17 or fewer) CAG repeat length. Short CAG repeat length on the androgen receptor gene is associated with prostate cancer in African American men and possibly with higher stage cancers. To this end, we are currently using input datasets on “CAG repeat” polymorphisms in the androgen receptor gene, which are cited in MEDLINE. The experiment is intended to strongly validate our methodology. The extrapolation of the CG-islands model to ‘CAG repeat’ polymorphisms model in the androgen receptor follows naturally since both subsequences are derived from DNA sequences.

In the analysis, it is hypothesized that unique androgen receptor (AR) molecular signatures exists in prostate tumor specimen obtained from African American men; and that the ‘CAG repeat’ is a contributing biomarker. An antithesis of this hypothesis is the absence or less prevalence of the AR signatures in prostate tumor specimen obtained from Caucasian men. Thus, polymorphisms in the androgen receptor found in African American men could be potentially important in explaining the increased risk of prostate cancer in African American men.

As an outcome of the experimental studies, it is expected that global gene expression profiles of prostate cancer in African American and Caucasian men will be obtained. The information learned from these global profiles will help us discern if there are differences in the molecular signature of prostate cancer for Caucasian men versus African American males. Coupling literature reviews on microarray analysis of prostate tumors from white and blacks with the results from our data mining approach will have far-reaching results. In that the results could yield recommendations that impact the development of new

pharmacogenomic therapies/strategies for effective prevention and successful management of prostate cancer, particularly in African American men.

8 Conclusion

In this paper, we have described algorithmic and statistical approaches for finding biomarkers that could be potential factors that cause prostate cancer among African American men. It is the goal that applying computational techniques to large DNA datasets on population that has a significant genetic predisposition will also help validate these techniques. In particular, the focus has been on the application of HMM – a statistical approach – and associated software on high-performance computing environment. With this approach, an extensive analysis on large datasets could be done. The validation process is also intended to strengthen the computational process, and to support the conclusion that polymorphisms in the androgen receptor found in African American men could be potentially important in explaining the increased risk of prostate cancer in African American men.

As a benefit to public health issues, finding clusters of genes that elucidate possible mechanisms/etiologies such as nutrition, environment, and genes leading to prostate cancer in the general population is a desirable process. A comparative study of literature review could further indicate the interplay of genetic, behavioral, and environmental factors that dictate prostate cancer susceptibility; and also benefit patient education and public health prevention programs tailored to the needs of the populations at risk.

References

- [1] Sanderson M, Coker AL, Logan P, Zheng W, Fadden M. K., Lifestyle and prostate cancer among older African-American and Caucasian men in South Carolina. *Cancer Causes and Control*, Vol. 15, 2004, pp. 647-655.
- [2] *U-M researchers seek answers for African-Americans at risk for prostate cancer*. July 1, 2003. <http://www.med.umich.edu/opm/newspage/2003/africanamericanprostate.htm>
- [3] Lok C, Better Cancer Detection: New tests could catch the disease earlier, *Technology Review: MIT's Magazine of Innovation*, Vol. 108, No. 10, October 2005, pp. 21-22.
- [4] Pevzner P. A., Computational Molecular Biology: An Algorithmic Approach, MIT Press, Cambridge, MA, 2000, 314 p.
- [5] Viterbi, A., Error bounds for convolutional codes and an asymptotically optimal coding algorithm, *IEEE Transactions on Information Theory*, Vol. 13, 1967, pp. 260-269.