Towards Better than Human Capability in Diagnosing Prostate Cancer Using Infrared Spectroscopic Imaging

Xavier Llorà[1], Rohith Reddy[2,3], Brian Matesic[2], and Rohit Bhargava[2,3]

[1]National Center for Super Computing Applications (NCSA)

[2]Department of Bioengineering

[3]Beckman Institute for Advanced Science and Technology

University of Illinois at Urbana-Champaign, Urbana IL 61801

xllora@uiuc.edu, rkreddy2@uiuc.edu, matesic2@uiuc.edu, rxb@uiuc.edu

ABSTRACT

Cancer diagnosis is essentially a human task. Almost universally, the process requires the extraction of tissue (biopsy) and examination of its microstructure by a human. To improve diagnoses based on limited and inconsistent morphologic knowledge, a new approach has recently been proposed that uses molecular spectroscopic imaging to utilize microscopic chemical composition for diagnoses. In contrast to visible imaging, the approach results in very large data sets as each pixel contains the entire molecular vibrational spectroscopy data from all chemical species. Here, we propose data handling and analysis strategies to allow computer-based diagnosis of human prostate cancer by applying a novel genetics-based machine learning technique (NAX). We apply this technique to demonstrate both fast learning and accurate classification that, additionally, scales well with parallelization. Preliminary results demonstrate that this approach can improve current clinical practice in diagnosing prostate cancer.

Categories & Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning–Concept Learning.
I.5.4 [Pattern Recognition]: Applications.
J.3 [Life & Medical Science]: Medical Information Systems.

General Terms

Algorithms, Design, Performance, Experimentation.

Keywords

Genetics-Based Machine Learning, Learning Classifier Systems, Parallelization, Prostate Cancer.

1. INTRODUCTION

Pathologist opinion of structures in stained tissue is the definitive diagnosis for almost all cancers and provides critical input for therapy. In particular, prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men, and it is a leading cause of cancer-related death. Hence, it is, appropriately, the subject of heightened public awareness and widespread screening. If prostate-specific antigen (PSA) or digital rectal screens are abnormal, a biopsy is considered to detect or rule out cancer. Prostate tissue is extracted, or biopsied, from the patient and examined for structural alterations. The diagnosis procedure involves the removal of cells or tissues, staining them with dyes to provide visual contrast and examination under a microscope by a skilled person (pathologist).

Due to personnel, tarining, natural variability and biologic differences, the challenge in prostate cancer research and practice is to provide accurate, objective and reproducible decisions. Conventional optical microscopy followed by manual recognition has been demonstrated to be inadequate for this task. [18]. Hence, we have recently proposed developing a practical approach to this problem using chemical, rather than morphologic, imaging. [19]. In this approach, Fourier transform infrared imaging (FTIR) is employed to provide the entire vibrational spectroscopic information from every pixel of a sample's microscopy image. While the first steps of developing novel imaging and sampling technologies is now reliable, [7] the computational challenge of providing robust classification algorithms that can rapidly provide decisions remains. Due to the above advances in imaging and sampling, data from thousands of patients is available to train and validate algorithms for different disease states. While the application and type of data are unique, a further confounding factor required efficiently processing large volumes of data generated by FTIR imaging. The classification problem can be formulated as a supervised learning problem in which several million pixels (hundred of gigabytes) of accurately labeled data are available for model training and validation. The volume of tissue and (future) need for intra-operative diagnoses imply that rapid and accurate diagnoses are crucial to allow physicians to explore all possible courses of action. Under these conditions, traditional supervised learning approaches and implementations do not scale to provide diagnoses in an appropriate

GECCO'07, July 7–11, 2007, London, England, United Kingdom.
Copyright 2007 ACM 978-1-59593-697-4/07/0007 ...$5.00.

time frame. Hence, efficiently processing and learning models from gigabytes of FITR imaging data requires a careful design of the supervised learning algorithm. Moreover, the biological nature of the problem requires that such models be interpretable to provide fundamental new insight into the disease process. Genetics-based machine learning (GBML) techniques take advantage of the *"quasi embarrassing parallelism"* [17] to provide scaleable, fast, accurate, reliable, and interpretable models. In this paper we present an approach engineered to the desired solutiona and constraints of addressing this human task. A modified version of a sequential genetics-based rule learner that exploits massive parallelisms via the message passing interface (MPI) and efficient rule-matching using hardware-oriented operations is developed. We named this system `NAX` [24], and we have shown that its performance is comparable to traditional and genetics-based machine learning techniques on an array of publicly available data sets. We now show that`NAX`—taking advantage of both hardware and software parallelism—is able to provide prostate cancer diagnoses that are human-competitive. In this paper, we present preliminary results supporting this outcome.

The paper is structured as follows. Section 2 provides an overview of our approach towards computer-aided diagnoses for prostate cancer. Procedure and form of the data are summarized in section 3. `NAX` is introduced in section 4, where we describe the basic components and design decisions in this approach. In section 5 we present preliminary results indicating that the approach presented in this paper is human-competitive. Finally, section 6 summarizes some conclusions and further research.

2. PROBLEM DESCRIPTION

Prostate cancer is the most common non-skin malignancy in the western world. The American Cancer Society estimated 234,460 new cases of prostate cancer in 2006 [31]. Recognizing the public health implications of this disease, men are actively screened through digital rectal examinations and/or serum prostate specific antigen (PSA) level testing. If these screening tests are suspicious, prostate tissue is extracted, or biopsied, from the patient and examined for structural alterations. Due to imperfect screening technologies and repeated examinations, it is estimated that more than 1 million people undergo biopsies in the US alone.

2.1 Prostate Cancer Diagnosis

The removal of a small section of prostate is most often accomplished by core biopsy. A needle is inserted into the tissue and several (6-23) samples are obtained from different positions. Biopsy, followed by manual examination under a microscope is the primary means to definitively diagnose prostate cancer as well as most internal cancers in the human body. Pathologists are trained to recognize patterns of disease in the architecture of tissue, local structural morphology and alterations in cell size and shape. Specific patterns of specific cell types distinguish cancerous and non-cancerous tissues. Hence, the primary task of the pathologist examining tissue for cancer is to locate foci of the cell of interest and examine them for alterations indicative of disease.

The specific cells in which cancer arises in the prostate are epithelial cells. While epithelial-origin cancers account for over 85% of all human cancers, they account for more than 95% of prostate cancers. In prostate tissue, epithelial line secretory ducts within the structural cells (collectively termed 'stroma') that allow the tissue to maintain its structure and function. Hence, a pathologist will first locate epithelial cells in a biopsy and, to examine for cancer, will mentally segment them from stroma.

Biopsy samples are prepared in a specific manner to aid in recognition of cells and disease. The sample is sliced thin ($\sim 5\mu m$ thickness), placed on a glass slide and stained with a dye to provide contrast. The most common dye is a mixture of hematoxylin and eosin ($H\&E$), which stains protein-rich regions pink and nucleic acid-rich regions blue. Empty space, lipids and carbohydrates are typically not stained and characterized by white color in images. Staining allows the pathologist to identify cells based on their nucleus and extra-nuclear regions. Patterns of the same cell type characterize structures. For example, epithelial cells arranged in a circular manner around empty space are characteristic of a duct and endothelial cells similarly arranged are characteristic of blood vessels. The empty space enclosed within a duct in pathology images is termed a lumen. The distortion of the circular pattern of epithelial cells around a lumen is characteristic of cancer.

In low severity cancers, lumens are only slightly distorted, while higher grades of cancer display a lack of lumen and simply consist of masses of epithelial cells supported by little stroma. The relative distortion and change in lumen shape is organized into a grading scheme to assess the severity of the disease, Gleason Scoring system, which is the primary measure of disease that defines diagnosis, helps direct therapy and helps predict those at danger of dying from the disease. Since prostate cancer is multi-focal and the disease quite variable, two dominant patterns of epithelial distortion are selected and each is independently graded on a scale of 1-5. The grades are then summed to provide a Gleason score ranging from 2 (low grade cancer) to 10 (maximum danger cancer). This scale has been widely used since its creation in the 1960s and currently forms the clinical standard of practice. Manual Gleason scoring, however, has severe limitations.

2.2 Limitations of Current Practice

Widespread screening for prostate cancer has resulted in a large workload of biopsied men [16], placing an increasing demand on services. Operator fatigue is well-documented and guidelines limit the workload and rate of examination of samples by a single operator (examination speed and throughput). Importantly, inter- and intra-pathologist variation complicates decision-making. The consistency in determining Gleason scores is rather poor. Intra-observer measurements show that a pathologist confirms their own score less than 50% of the time and are ±1 score no more than 80% of cases [2]. Hence, the diagnoses for $\sim 50\%$ of cases may change and may be significantly altered for $\sim 20\%$ of cases ultimately leading to changes in therapy for a patient subset [30]. The numbers are decidedly cause for concern. For example, a recent study including 15 pathologists and 537 prostate cancer patients, 70.8% of Gleason scores were shown to be inaccurate when compared with the patient's final outcome [18]. Second opinions [29] improve assessment and are cost-effective [10], not to mention their utility in mitigating the effects of healthcare costs, lost wages, morbidity, or potential litigation. In summary, the manual recognition of spatial patterns leaves much to be desired from a process perspective and has far-reaching social effects from a public health perspective.

For the reasons underlined above, there is an urgent need for high-throughput, automated and objective pathology tools. We believe that this need is best met by employing the power of computer algorithms and advanced processing to address prostate cancer diagnosis and grading.

The information content of conventionally stained images is limited, inherently non-specific and varies greatly within patient populations and processing conditions. Hence, the information derived from visible microscopy images is fundamentally limited and automated methods of analyzing stained images have failed to provide a sufficiently robust algorithm to diagnose disease. An alternative to morphology-based microscopy are molecular microscopy techniques to probe disease. Molecular technologies for disease diagnosis are an exciting venue for investigations as they promise better diagnostic capabilities through objective means and a multitude of chemicals to provide insight into the changes indicative of the disease process. In particular, spectroscopy tools allow for the measurement of many molecular species simultaneously. Spectroscopic techniques in imaging form, notably using optics, further enable the analysis to be conducted without perturbing the tissue [11]. In this manuscript, we present the analysis of prostate tissue with one such technique, Fourier transform infrared (FTIR) spectroscopic imaging.

2.3 Molecular Imaging

Infrared spectroscopy is a classical technique for measuring the chemical composition of specimens. At specific frequencies, the vibrational modes of molecules are resonant with the frequency of infrared light. By monitoring all frequencies in the region, a pattern of absorption can be created. This pattern, or spectrum, is characteristic of the chemical composition and is hypothesized to contain information that will help determine the cell type and disease state of the tissue. Recently, FTIR spectroscopy has been developed in an imaging sense. Hence, The data are similar to optical microscopy. The first difference is that no external dyes are needed and the contrast in images can be directly obtained from the chemical composition of the tissue. The second is that each pixel in the visible image contains RGB values but in IR imaging contains several thousand values across a bandwidth $(2000 - 14000nm)$ that is $\sim$ 40 times larger than the visible spectrum $(400 - 700nm)$ [7].

3. DATA AND METHODOLOGY

3.1 Experimental Details

Prostate tissues were obtained from Cooperative Human Tissue Network for the tissue array research program (TARP) laboratory. Using these tissues, tissue microarrays were prepared using a Beecher automated tissue arrayer containing a video overlap system and $0.6mm$ needles. Appropriate institutional review board and National Institutes of Health (USA) guidelines for the protection of human subjects were followed. $5\mu m$ sections of tissue were floated on an infrared transmissive optical window for FTIR spectroscopic imaging. Another $5\mu m$ section obtained from the same point on the tissue specimen was observed using traditional microscopy for comparison. Expert pathologists determined the tissue classification using these microscopy samples by staining with $H\&E$. Pathologists' classification were used as the 'gold standard' for comparison with the results from the methods mentioned in this paper.

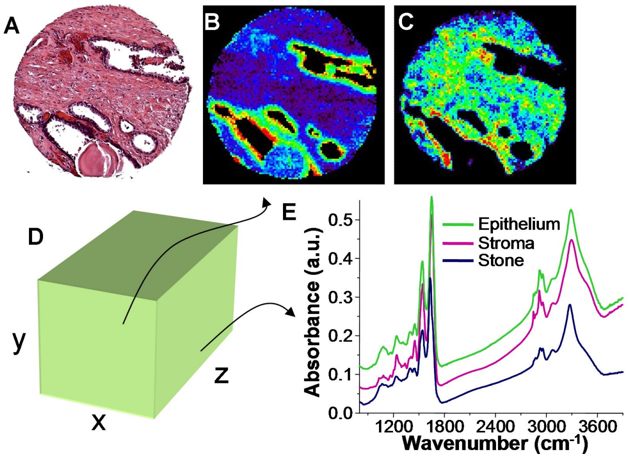


Figure 1: Conventional Staining and Automated Recognition by Chemical Imaging. (A) Typical H&E stained sample, in which structures are deduced from experience by a human. Highlights of specific regions in the manner of H&E is possible using FTIR imaging without stains. (B) Absorption at 1080 cm-1 commonly attributed to nucleic acids and (C) to proteins of the stroma. The data obtained is 3 dimensional (D) from which spectra (E) or images at specific spectral features may be plotted.

Tissues were analyzed using a Michelson interferometer attached to a microscope (Perkin-Elmer Spotlight 300) in transmission mode at a resolution of $4cm^{-1}$ The sample was then raster scanned to obtain images of the entire specimen. Typical specimen size is $600\mu m \times 600\mu m$ with each pixel being $6.25\mu m \times 6.25\mu m$ on the sample plane. Spectra are composed of $1,641$ sample points of the spectral range $4,000 - 720cm^{-1}$. Data acquisition using these techniques required 40 minutes per cylindrical core of the tissue microarray to yield a root mean square signal to noise ratio of $500 : 1$. A typical array was composed of approximately 2.5 million pixels and required 40 GB of storage space.

The data obtained from FTIR imaging is three-dimensional. The $x-$ and $y-$dimensions locate pixels on the tissue-sample plane. The z-dimension values compose the IR spectrum for the corresponding pixel. The spectra can be analyzed to determine what type of tissue (epithelium, stroma, or muscle) the specimen is as well as whether the tissue is malignant or benign. We have developed this technology to provide data from tissue in minutes and employ a high-throughput sampling strategy using Tissue Microarrays (TMA) to obtain data.[19] Samples from multiple tissues, from multiple patients and multiple clinical settings are included in the data set to maximize the sampling of natural variability and ensure the development of robust analysis algorithms. These high-throughput imaging and microarray technologies combine to provide very large data sets—see Figure 1. A typical single core consists of 300×300 pixels on the $x - y$ plane with 1641 bands on the z-axis. A tissue microarray consists of several hundred such cores and analysis of such large datasets (typically, tens of GB) is computationally expensive.

3.2 Data Format

Each pixel's z-dimension contains a spectrum characteristic of the chemical composition of that region of the specimen. Certain spectral quantities provide measures of chemistry. For example, the height of each feature is proportional to its abundance, the peak position is associated with the vibrational identity and peak shape often reflects the multitude of environments around the molecule. Therefore, differences in spectral characteristics can be used in classification and these exact spectral features are termed 'metrics'. For example, the ratio of absorbance of the spectral peak at $1080cm^{-1}$ to the spectral peak at $1545cm^{-1}$ is commonly used to distinguish epithelial from stromal cells. Trained spectroscopists determine these metrics based upon examination of spectral patterns. Hence, the reduction of ull spectra to descriptive metrics forms an intelligent dimensionality reduction strategy. Genetic algorithms form decision rules based upon these metrics to classify pixels by tissue type. Furthermore, the transparency of the genetic algorithms allows the scientist to correlate specific rules to biological features (tissue type and cancer classification) via metrics based upon spectral characteristics.

4. APPROACH

In this section we review related work on the GBML community, highlighting previous efforts to deal with large data sets. We also present the motivation and techniques that lead to the design of NAX. Special attention is paid to the description of the hardware and software techniques used, as well as to the design of a scalable GBML algorithm.

4.1 Related Background

Bernadó, Llorà & Garrell [6] presented a first empirical comparison between genetics-based machine learning techniques (GBML) and traditional machine learning approached. The authors reported that GBML techniques were able to perform as well as traditional techniques. Later on, Bacardit & Butz [3] repeated the analysis again obtaining similar results. Most of the experiments presented on both papers were conducted using publicly available data sets provided by the *University of California at Irvine* repository [28]. Most of the data sets are defined over tens of features and up to few thousands of records. However, a key property of GBML approaches is its intrinsic massive parallelism and scalability properties. Cantú-Paz [8] presented how efficient and accurate genetics algorithms could be assembled, and Llorà [21] presented how such algorithms can be efficiently used as machine learning and data mining techniques.

GBML techniques require evaluating candidate solutions against the original data set matching the candidate solutions (e.g. rules, decision trees, prototypes) against all the instances in the data set. Regardless of the GBML flavor used, Llorà & Sastry [25] showed that as the problem grows, the matching process governs the execution time. For small data sets (teens of attributes and few thousands of records) the matching process takes more than 85% of the overall execution time marginalizing the contribution of the other genetic operators. This number easily passes 99% when we move to data sets with few hundreds of attributes and few hundred thousands of records. Such results emphasize one unique facet of GBML approaches: scalability via exploiting massive parallelism. More than 99% of the time required is spent on evaluated candidate solutions. Each solution evaluation is independent of each other and, hence, it can be computed in parallel. Moreover, the evaluation process can also be parallelized further on large data sets by splitting and distributing the data across the computational resources. A detailed description of the parallelization alternatives of GBML techniques can be found elsewhere [21].

Currently available off-the-shelf GBML methods and software distributions [5, 20] do not usually target dealing with very large data sets. Three different works need to be mentioned here. Flockhart [12] proposed and implemented GA-MINER, one of the earliest effort to create data mining systems based on GBML systems that scale across symmetric multi-processors and massively parallel multiprocessors. The work review different encoding and parallelization schemes and conducted proper scalability studies. Llorà [21] explored how fine-grained parallel genetic algorithms could become efficient models for data mining. Theoretical analysis of performance and scalability were developed and validated with proper simulations. Recently, Llorà & Sastry [25] explored how current hardware can be efficiently used to speed up the required matching of solutions against the data set. These three approaches are the basis of the incremental rule learning proposed in the next section to approach very large data sets—such as the prostate tissue classification one.

4.2 The Road to Tractability

NAX evolves, one at a time, maximally general and maximally accurate rules. Then, the covered instance are removed and another rule is added to the previously stored one, forming a decision list. This process continues until no uncovered instances are left. Llorà, Sastry & Goldberg [26] showed that maximally general and maximally accurate rules [32] could also be evolved using Pittsburgh-style learning classifier systems. Later, Llorà, Sastry & Goldberg [27] showed that competent genetic algorithms [15] evolve such rules quickly, reliably, and accurately. From these early works, it can be inferred that approaching real-world problems, such as the prostate tissue classification and cancer diagnosis, using GBML techniques may produce the desired byproduct: proper scalability. We discuss next efficient implementation techniques to deal with very large data sets using NAX [24].

4.3 Exploiting the Hardware

Recently, multimedia and scientific applications have pushed CPU manufactures to include support for vector instruction sets again in their processors. Both applications areas require heavy calculations based on vector arithmetic. Simple vector operations such as *add* or *product* are repeated over and over. During 80s and 90s supercomputers, such as Cray machines, were able to issue hardware instructions that took care of basic vector operations. A more constrained scheme, however, has made its way into general-purpose processors thanks to the push of multimedia and scientific applications. Main chip manufactures—

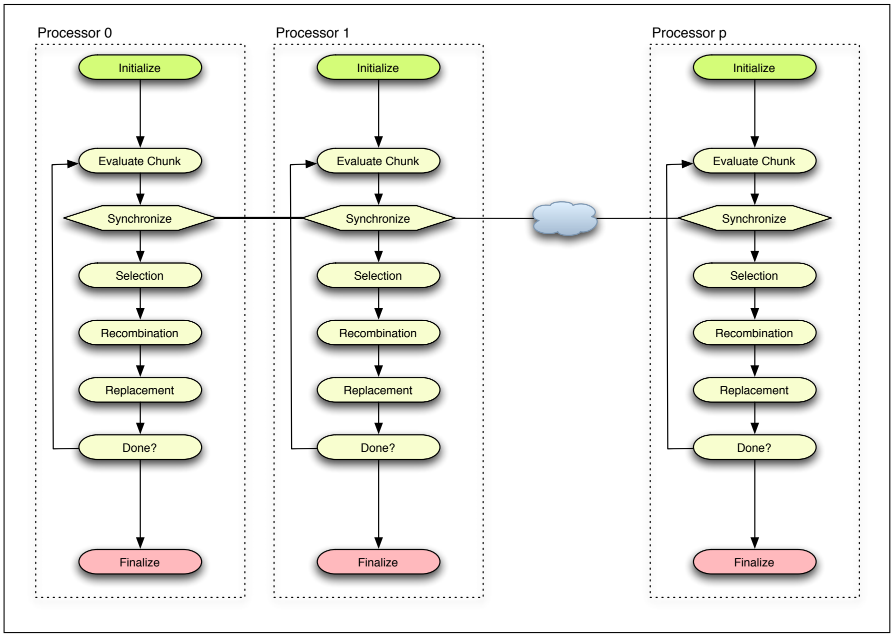


Figure 2: This figure illustrates the parallel model implemented. Each processor is running an identical `NAX` algorithm. They only differ in the portion of the population being evaluated. The population is treated as collection of chunks where each processor evaluates its own assigned chunk sharing the fitness of these individuals with the rest of processors. This approach minimizes communication cost.

IBM, Intel, and AMD—have introduced vector instruction sets—Altivec, SSE3, and 3DNow$^+$—that allow performing vector operations over packs of 128 bits by hardware. We will focus on a subset of instructions that are able to deal with floating point vectors. This subset of instructions to implemented by hardware vector operations against groups of four floating-point numbers. These instructions are the basis of the fast rule matching mechanism proposed.

Our set of rules seek both to correctly classify the prostate data set and provide biological insight into the rules. All the attributes of the domain are real-value and the conditions of the rules need to be able to express conditions in a $\Re^n$ spaces. We use a rule encoding similar to the one proposed by Wilson [33] and widely used in the GBML community. Rules express the conjunction of tests across attributes. Each test can be defined in multiple fashions, but without loss of generality, we pick a simple interval based one. A simple example of and *if-then* rule, could be expressed as follows:

$$1.0 \leq a_0 \leq 2.3 \wedge \cdots \wedge 10.0 \leq a_n \leq 23 \rightarrow c_1 \qquad (1)$$

Where the condition is the conjunction of the different attribute tests, as introduced earlier, and the condition is the predicting class. We also allow a special condition—`don't care`—which always returns `true` to allow generalized to rules evolve. The rule below illustrates an example of a generalized rule.

$$1.0 \leq a_0 \leq 2.3 \wedge -3.0 \leq a_3 \leq 2 \longrightarrow c_1 \qquad (2)$$

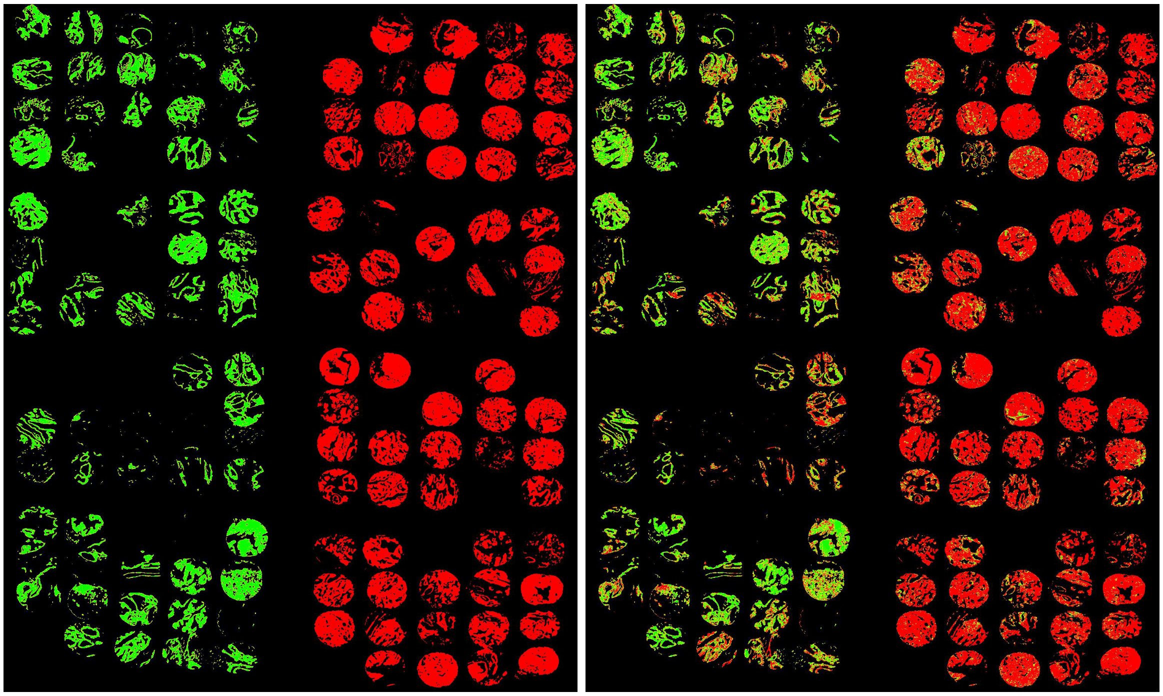
All attributes except a_0 and a_3 were marked as `don't care`.

Matching a rule requires performing the individual tests before the final *and* condition can be computed. Vector instruction sets can help improve the performance of this process by performing four tests at once. Actually, this process can be regarded as four parallel running pipelines. The process can be improved further by stopping the matching process when any one test fails. The code implemented assumes that the two vectors containing the upper and lower bounds are provided and records are stored in a two dimensional matrix. As also shown elsewhere [25], exploiting the hardware available can speed between 3 and 3.5 times the matching process[24].

4.4 Massive Parallelism

Since most of the time is spent on the evaluation of candidate rules when dealing with large data sets, our next goal was to find a parallelization model that could take advantage of this feature. Due to the embarrassing parallelism model [17] for rule evaluation, we designed a coarse-grain parallel model for distributing the evaluation load. Cantú-Paz [8] proposed several schemes, showing the importance of the trade off between computation time and time spent communicating. When designing the parallel model, we focused on minimizing the communication cost. Usually, a feasible solution could be a master/slave one—the computation time is much larger than the communication one. However, GBML approaches tend to use rather large populations, forcing us to send rules to the evaluation slaves and collect the resulting fitness. This scheme also increments sequential instructions that cannot be parallelized, reducing the overall speedup of the parallel implementation as a result of Ambdhals law [1].

To minimize communication cost, each processor runs identical `NAX` algorithms—all seeded in the same manner, and, hence performing the same genetic operations. They only differ in the portion of the population being evaluated. Thus, the population is treated as collection of chunks where each processor evaluates its own assigned chunk, sharing the fitness of the individuals in its chunk with the rest of processors. in this manner fitness can be encapsulated and broadcasted, maximizing the occupation of the underlying packing frames used by the network infrastructure. Moreover,



(a) Original labeled array

(b) Automatically classified array

Figure 3: This figure on the left-hand side presents the original labeled data contained in the P80 array. The figure on the right-hand side presents the reconstructed image based on the predictions issued by the the rule set evolved by `NAX`. Green represent non cancerous tissue spots; red represent malignant tissue spots.

this approach also removes the need for sending the actual rules back and forth between processors—as a master/slave approach would require—thus, maintaining the communication to the bare minimum—namely, the fitness. Figure 2 presents a conceptual scheme of the parallel architecture of `NAX`.

To implement the model presented in Figure 2, we used `C` and the *open message passing interface* (openMPI) implementation [13]. Each processor computes which individuals are assigned to it. Then it computes the fitness and, finally, it broadcasts the computed fitness. The rest of the process is unchanged. Except for the cooperative evaluation, all the processors generate the same evolutionary trace.

4.5 Lists of Maximally General and Maximally Accurate Rules

One main characteristic of the so-called Pittsburgh-style learning classifier systems—a particular type of GBML—is that the individuals encode a rule set [14, 22, 15]. Thus evolutionary mechanisms directly recombine one rule set against another one. For classification tasks of moderate complexity, the rule sets are not large. For complex problems, however, the potential number of rules required to ensure accurate classification may use prohibitively large amounts of memory. The requirements increase even further in the presence of noise [23]. Hence, this family of GBML techniques works very well on moderate complexity problems [6, 3], but needs to be modified for complex and large data sets.

A sequential rule learning approach may alleviate the requirements by evolving only one rule at a time, hence, reducing the memory requirements [9, 4]. This allows maintaining relatively small memory footprints that makes feasible processing large data sets. However, an incremental approach to the construction of the rule set requires paying special attention to the way rules are evolved. For each run of the genetic algorithm, we would like to obtain a maximally general and maximally accurate rule, that is, a rule that covers the maximum number of examples without making mistakes [32]. `NAX` (our proposed incremental rule learner) evolves maximally general and maximally accurate rules by computing the *accuracy* (α) and the *error* (ε) of a rule [26]. In a Pittsburgh-style classifier, the *accuracy* may be computed as the proportion of overall examples correctly classified, and the *error* is the proportion of incorrect classifications issued. Once the *accuracy* and *error* of a rule are known, the fitness can be computed as follows.

$$f(r) = \alpha(r) \cdot \varepsilon(r)^{\gamma} \tag{3}$$

where γ is the error penalization coefficient. We have set γ to 18 to guarantee that the evolutionary process will produce maximally general and maximally accurate solutions. Further details may be found elsewhere [24]. The above fitness measure favors rules with a good classification accuracy and a low error, or maximally general and maximally accurate rules. By increasing γ, we can bias the search towards correct rules. This is an important element because assembling a rule set based on accurate rules guarantees the overall performance of the assembled rule set. `NAX`'s efficient implementation of the evolutionary process is based on the techniques described using hardware acceleration—section 4.3—and coarse-grain parallelism—section 4.4. The genetic algorithm used was a modified version of the *simple genetic algorithm* [14] using tournament selection ($s = 4$), one point crossover, and mutation based on generating new random boundary elements.

5. RESULTS

NAX has shown competitiveness in evolving rule sets that perform as accurately as the ones evolved by other genetics-based machine learning and non-evolutionary machine learning techniques. However, NAXs key element is the ability to deal with large data sets. In this paper, we present preliminary results towards evolving a model capable of correctly classifying pixels as cancerous or non-cancerous. The original array of spots is presented in figure 3(a). Each spot corresponds to a different biopsy sample from a patient. The pixels present in each spot correspond to the epithelial tissue of the biopsy, we supress all other tissue types with a prior classification filter based on Bayesian Likelihood.[7] Each pixel of a spot is defined by 93 different metrics extracted from the processed infrared spectra—as described in section 3. Finally, each pixel in the array was labeled with the diagnostic class provided by a human pathologist. Figure 3(a) presents in green all the non-cancerous pixels while red identifies cancerous ones.

Our goal with the initial experiments here was to demonstrate the usefulness of the proposed approach to computer-aided diagnosis. Our current experimental efforts are planning mass experimentation on several tissue arrays using the Tungsten cluster at the National Center for Supercomputing Applications. These initial experiments were conducted on a dual core Intel Xeon 2.8GHz Linux computer with 1Gb of RAM. NAX was run using both processors. The training time to obtain a model describing all the data took less than ten hours—indicating that very competitive training times can be achieved by just using more processors. The obtained model was able to correctly classify $> 99.99\%$ of the training pixels correctly. However, these results do not illustrate the generalization capabilities of the models evolved by NAX. Hence, we ran a series of ten-fold stratified cross-validation runs [34] to measure generalization and test performance of the evolved models. It is important to mention that tools such as WEKA [34] and other off-the-shelf data miners were not able to handle the volume of data required to evolve a model— either due to the large memory footprint required or by not being able to provide an accurate model in a feasible time period. The results of the cross-validation experiments using NAX correctly classified 87.34% of validation pixels. Such results are more than encouraging, because they show a human-competitive computer-aided diagnosis system is possible. Another interesting property is that a few rules classify a large number of pixels—see Figure 4. Such a result is interesting for the interpretability of the model, since a small number of rules have a great expressiveness, and hence may provide valuable biological insight. Most importantly, they allow us to classify tissue accurately. Subsequent to this pixel level classification, each circular spot in figure 3 was assigned as malignant or benign based on the majority of pixels of he class in the sample. We were able to accurately classify 68 of 69 malignant spots and 70 of 71 benign spots in this manner. While human accuracy is difficult to quantify due to the variation between persons,a generally accepted anecdotal figure is about 5% error rates. The preliminary results we demonstrate here could potentially reduce that five-fold to about 1%, providing a solution to this real-world problem by a combination of novel spectroscopy and advanced machine learning.

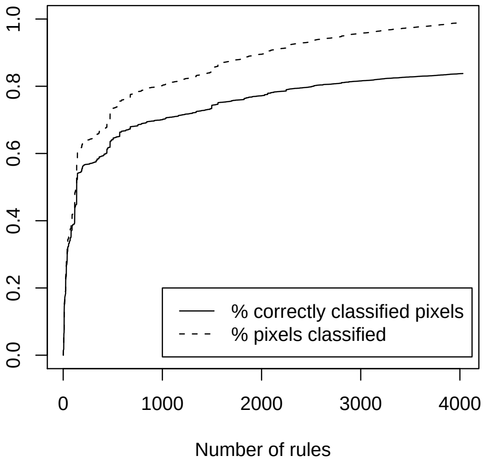


Figure 4: Performance of the evolved model as a function of the number of rules used.

6. CONCLUSION

In this manuscript, we present the application of advanced genetics-based machine learning algorithms to a real-world problem of large scope, namely, the diagnosis of prostate cancer. As opposed to subjective human recognition of disease in tissue using light microscopy, we employed a chemical microscopy approach that required extensive computation but provided a decision without human input. Our development of a learning algorithm based on maximally general and maximally accurate rules was scalable to very large data sets and parallelized to provide learning and classification speed advantages. The algorithm was able to classify a majority of pixels correctly, resulting in overall error rates that were comparable to human examination, the current gold standard of care.

7. ACKNOWLEDGMENTS

This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant FA9550-06-1-0370, the National Science Foundation under grant ISS-02-09199 at the National Center for Supercomputing Applications, UIUC. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation thereon.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research, the National Science Foundation, or the U.S. Government. We would like to thank Dr. Stephen M. Hewitt, MD PhD, who provided us the samples and expert pathology consultation. The research is also supported by grants to RB from the Research Board, UIUC; the Department of Defense through the Prostate Cancer Research Program and by the National Center for Supercomputing Applications, under the auspices of the NCSA/UIUC faculty fellows program.

8. REFERENCES

[1] G. Amdahl. Validity of the single processor approach to achieving large-scale computing capabilities. In *Proceedings of the American Federation of Information Processing Societies Conference (AFIPS)*, volume 30, pages 483–485. AFIPS, 1967.

[2] M. Amin, D. Grignon, P. Humphrey, and J. Srigley. *Gleason Grading of Prostate Cancer: A Contemporary Approach*. Lippincott Williams & Wilkins: Philadelphia, 2004.

[3] J. Bacardit and M. Butz. *Advances at the frontier of Learning Classifier Systems (Volume I)*, chapter Data Mining in Learning Classifier Systems: Comparing XCS with GAssist, page in press. Springer-Verlag, 2006.

[4] J. Bacardit and N. Krasnogor. Biohel: Bioinformatics-oriented hierarchical evolutionary learning. Nottingham eprints, University of Nottingham, 2006.

[5] A. Barry and J. Drugowitsch. LCSWeb: the LCS wiki, 1997. http://lcsweb.cs.bath.ac.uk/.

[6] E. Bernadó, X. Llorà, and J. Garrell. *Advances in Learning Classifier Systems: 4th International Workshop (IWLCS 2001)*, chapter XCS and GALE: a Comparative Study of Two Learning Classifier Systems with Six Other Learning Algorithms on Classification Tasks, pages 115–132. Springer Berlin / Heidelberg, July 2001.

[7] R. Bhargava, D. Fernandez, S. Hewitt, and I. Levin. High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data. *Biochemica et Biophisica Acta*, pages 830–845, 2006.

[8] E. Cantú-Paz. *Efficient and Accurate Parallel Genetic Algorithms*. Kluwer Academic Publishers, 2000.

[9] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena. *Genetic Fuzzy Systems. Evolutionary tuning and learning of fuzzy knowledge bases*. World Scientific, 2001.

[10] J. Epstein, P. Walsh, and F. Sanfilippo. Clinical and Cost Impact of Second-opinion Pathology: Review of Prostate Biopsies Prior to Radical Prostatectomy. *American Journal of Surgical Pathology*, 20:851–857, 1996.

[11] D. Fernandez, R. Bhargava, S. Hewitt, and I. Levin. Infrared spectroscopic imaging for histopathologic recognition. *Nature Biotechnology*, 23(4):469–474, 2005.

[12] I. Flockhart. GA-MINER: parallel data mining with hierarchical genetic algorithms (final report). Technical Report Technical Report EPCCAIKMS-GA-MINER-REPORT 1.0, University of Edinburgh, 1995.

[13] E. Gabriel, G. Fagg, G. Bosilca, T. Angskun, J. Dongarra, J. Squyres, V. Sahay, P. Kambadur, B. Barrett, A. Lumsdaine, R. Castain, D. Daniel, R. Graham, and T. Woodall. Open MPI: Goals, concept, and design of a next generation MPI implementation. In *Proceedings of the 11th European PVM/MPI Users' Group Meeting*. Springer, 2004.

[14] D. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional., 1989.

[15] D. Goldberg. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Springer, 2002.

[16] S. J. Jacobsen, S. K. Katusic, E. J. Bergstralh, J. E. Oesterling, O. Del, G. G. Klee, C. G. Chute, and M. M. Lieber. Incidence of prostate cancer diagnosis in the eras before and after serum prostate-specific antigen testing. *JAMA*, 274:1445–1449, 1995.

[17] V. Kumar, A. Grama, A. Gupta, and G. Karpis. *Introduction to Parallel Computing: Design and Analysis of Parallel Algorithms*. Benjamin-Cummings Publishing Company, 1994.

[18] J.-B. Lattouf and F. Saad. Gleason score on biopsy: is it reliable for predcting the final grade on pathology? *BJU International*, 90:694–699, 2002.

[19] I. Levin and R. Bhargava. Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. *Annual Review of Physical Chemistry*, 56:429–474, 2005.

[20] X. Llorà. Learning Classifier Systems and other Genetics-Based Machine Learning Blog, 2006. http://www-illigal.ge.uiuc.edu/lcs-n-gbml/.

[21] X. Llorà. *Genetics-Based Machine Learning using Fine-grained Parallelism for Data Mining*. PhD thesis, Enginyeria i Arquitectura La Salle. Ramon Llull University, Barcelona, Catalonia, European Union, February, 2002.

[22] X. Llorà and J. Garrell. Knowledge-independent data mining with fine-grained parallel evolutionary algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'2001)*, pages 461–468. Morgan Kaufmann Publishers, 2001.

[23] X. Llorà and D. Goldberg. Bounding the effect of noise in multiobjective learning classifier systems. *Evolutionary Computation Journal*, 11(3):279–298, 2003.

[24] X. Llorà, A. Priya, and R. Bhargava. Observer-invariant histopathology using genetics-based machine learning. Technical report, Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign (IlliGAL TR No 200627), 2006.

[25] X. Llorà and K. Sastry. Fast rule matching for learning classifier systems via vector instructions. In *Proceedings of the 2006 Genetic and Evolutionary Computation Conference*, pages 1513–1520. ACM Press, 2006.

[26] X. Llorà, K. Sastry, and D. Goldberg. The compact classifier system: Motivation, analysis and first results. In *Proceedings of the Congress on Evolutionary Computation*, volume 1, pages 596–603. IEEE press, 2005. (Also as IlliGAL TR No. 2005019).

[27] X. Llorà, K. Sastry, D. Goldberg, and L. de la Ossa. The χ-ary extended compact classifier system: Linkage learning in Pittsburgh LCS. In *Advances at the frontier of Learning Classifier Systems (Volume II)*, page in preparation. Springer, 2007. IlliGAL report no. 2006015.

[28] C. J. Merz and P. M. Murphy. UCI repository for machine learning data-bases, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository. html.

[29] W. Murphy, I. Rivera-Ramirez, L. Luciani, and Z. Wajsman. Second opinion of anatomical pathology: A complex issue not easily reduced to matters of right and wrong. *J. Urol*, 165:1957–1959, 2001.

[30] J. Nguyen, D. Schultz, A. Renshaw, R. Vollmer, W. Welch, K. Cote, and A. D'Amico. The impact of pathology review on treatment recommendations for patients with adenocarcinoma of the prostate. *Urologic Oncology: Seminars and Original Investigations*, 22:295–299, 2004.

[31] A. C. Society. How Many Men Get Prostate Cancer?, 2006. `http://www.cancer.org/docroot/CRI/content/CRI_2_2_1X_How_many_men_get_prostate_cancer_36.asp?rnav=cri`.

[32] S. Wilson. Classifier fitness based on accuracy. *Evolutionary Computation*, 3(2):149–175, 1995.

[33] S. Wilson. Get real! XCS with continuous-valued inputs. *Lecture Notes in Computer Science*, 1813:209–219, 2000.

[34] I. H. Witten and E. Frank. *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA., 2000.