Bioinformatics Techniques Used in Diabetes Research

Robert Kelley[1] and Eric C. Rouchka[1]
TR-ULBL-2007-02

November 28, 2007

[1]University of Louisville
Speed School of Engineering
Department of Computer Engineering and Computer Science
123 JB Speed Building
Louisville, Kentucky, USA 40292

rkelley@sullivan.edu; eric.rouchka@louisville.edu

UNIVERSITY of LOUISVILLE dare to be great
Bioinformatics Laboratory Technical Report Series

Bioinformatics

Bioinformatics Techniques used in Diabetes Research

Robert Kelley[1] and Eric C. Rouchka[1,*]

[1]Department of Computer Engineering and Computer Science, University of Louisville, 123 JB Speed Building, Louisville, KY, USA

UNIVERSITY OF LOUISVILLE BIOINFORMATICS LABORATORY TECHNICAL REPORT SERIES REPORT NUMBER **TR-ULBL-2007-02**

ABSTRACT

Diabetes is an endrocinological condition that affects millions of people and can cause myriad health complications. Numerous bioinformatics tools are being used in diabetes research. Here we look at the types of tools commonly used by researchers both by category of tool e.g. sequence alignment, microarray analysis and specific tools mentioned e.g. GCG Pileup, ClustalW, etc. to get an overall picture bioinformatics and diabetes as well as the context in which such tools are employed.

1 INTRODUCTION

Diabetes is devastating disease that is characterized by high glucose levels in the blood and has been recorded in the medical literature since as early as 1500 BC [1]. The causes were not well understood until the 19th century when Paul Langerhans, a German medical student discovered what have become known as the islets of Langerhans which produce insulin. Insulin is a protein hormone that promotes the uptake of glucose by the body's cells [2]. Diabetics either do not produce enough insulin to process their intake of glucose or the body does not use the insulin efficiently enough to control glucose levels. Treating diabetes has always been difficult and prior to the 1920's, the primary treatment plan was controlling diet. In 1921, Dr. Frederick Banting discovered how to extract insulin from cattle that could be used as an injectable form for human diabetics which revolutionized the treatment of the disease and made it more manageable.

Untreated, diabetes can cause a number of health problems including, blindness, loss of circulation resulting in limb amputation, high blood pressure, heart disease, and kidney failure. According to statistics presented by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDKD), over 20 million people in the United States had diabetes as of 2005 [3]. That represents about 7 percent of the entire population! The NIDDKD further reports that 132 billion dollars were spent in 2002 on medical treatments, disability payments, work loss, and premature mortality resulting from diabetes. These statistics make clear the fact that research into diabetes and therapies to prevent or treat it are of paramount importance to the health and well being of society in general as diabetes becomes a worldwide epidemic.

Like so many other areas of medicine, computer technology has had a profound impact on diabetes research. Our goal in this paper is to look at how bioinformatics techniques have been applied specifically to diabetes research and make comparisons between them to establish a solid understanding of how bioinformatics has impacted this important field of study.

2 METHODS

In order to investigate the bioinformatics tools and methodologies used to in diabetes research, we determined it was first necessary to develop categories of tools/techniques on which we would focus. At first, this was difficult to do because we did not have a preconceived idea about how the research would be organized and how bioinformatics tools would be described or identified in the research. To get started, we ran several cursory searches using basic search terms such as *bioinformatics and diabetes(research)* through several databases to see what types of articles were returned. Since we wanted to see current practices for using bioinformatics techniques in diabetes research, we restricted our searches where possible to articles that had been published in the last five years. We then reviewed the abstracts of the results and skimmed the bodies of promising papers, and sketched some categories based on the methods sections of the reviewed papers. While there are many possible approaches to classification in this area, we were interested in simplicity and keeping the research manageable. To that end we established the following three bioinformatics tool categories in which articles could be placed.

(1) Sequence alignment projects/techniques – articles in this area cite sequence alignment as a primary tool for the research. This might include pairwise and multiple sequence alignments as well as BLAST searches.
(2) Gene Expression projects/techniques – articles in this area cite various methods to measure the expressions of genes in different organisms and conditions. Microarray analysis is also frequently mentioned in these papers.
(3) Databases and database techniques – articles in this area cite various databases that were either used to assist with research or compiled to assist other researchers with their research.

*To whom correspondence should be addressed.

We used a variety of tools for our search process including the database search capabilities located at:

- http://library.louisville.edu
- http://www.google.com
- http://scholar.google.com
- http://www.sciencedirect.com
- http://www.ncbi.nlm.nih.gov/.

In addition to categorizing the different types of tools and techniques used by researchers, we also considered the actual tools that were specifically mentioned in the articles. We did this to see if we could get determine if there were any programs/tools that were used more than others and how different tools were used in combination with one another.

3 RESULTS

As we indicated in the previous section, the three main categories in which we placed articles included sequence alignment techniques, gene expression techniques, and database techniques. We will discuss each section separately in the following sections, but before doing so we want to caution the reader that developing classifications of this nature are seldom straightforward. We understand that a different group of researchers might determine different categories. Moreover, with the categories we derived, we are quite aware that articles may fall into two or more categories. To keep the presentation of the articles simple, we considered the bioinformatics tool that was discussed the most (as a percentage of the paper) as the tool by which we classified the paper. Our results generally discuss the main category, although occasionally we will mention other tools if they are important or demonstrate a combined research methodology.

3.1 Sequence Alignment

Sequence alignment is often used by researchers to compare either the DNA or amino acid sequences of organisms to determine homology and generate phylogenetic relationships between them. There are actually two main types of sequence alignment, pairwise and multiple. Pairwise involves comparing two sequences to each other, while multiple sequence alignment involves aligning several sequences to each other or to a single sequence.

Without the ability to sequence DNA and proteins, much of the research generated in the last 50 years would not have been possible. Diabetes research is no exception. Until the 1950s the only way to tell if a form of insulin was usable by humans was experimentation. The sequence of the protein itself was not known and could not be compared to the sequences from different organisms. In 1953, Fred Sanger and O.E.P Thompson published their landmark paper in which they presented the entire sequence of the insulin protein, which was the first protein to be entirely sequenced [4]. In their paper they describe 20 amino acids that are chained together in two chains known as peptides. They discovered that human insulin has two chains, the A chain of 21 amino acids and the B chain, which has 30 amino acids and they are connected by disulfide bridges. The impact of this research was far-reaching not just for diabetes research, but also for genetics and sequencing research. Sanger and Thompson basically demonstrated that sequencing could be done in a reasonable amount of time for some proteins. Moreover, once sequencing was done, a synthetic could be theoretically produced, although in reality this actually took several years for insulin. Despite the fact synthesis was difficult, by analyzing the sequence of several different organisms' insulin, it was determined insulin was well conserved across organisms and any number of animals could be used to extract insulin for human subjects.

3.1.2 Traditional Sequence Alignment

Pairwise sequence alignment and multiple sequence alignment techniques are often used by researchers in biology as demonstrated by the Rao et. al in [5]. Rao's research was based on the premise that brain-derived neurotrophic factor (BDNF) controls the actions of several proteins including insulin, leptin, and gherlin and is significant to the pathobiology of type2 diabetes and obesity.

To test their hypothesis, Rao et al. located genes and proteins that are commonly present in diabetics for both homo sapiens and mus musculus (house mouse). Using multiple sequence alignment, they aligned the sequences representing BDNF, MET66, CRP, Insulin, Leptin, and Ghrelin from each organism. Using those alignment scores, they generated a phylogenetic tree using the ClustalW ver 1.83 program. After working with this smaller dataset, they then aligned additional sequences of other proteins involved in diabetes and obesity which numbered around 70 for the homo sapiens family and 59 for the mus musculus family. Results from both sets of alignment indicate there is a strong relationship between the two organisms and the related proteins that suggest BDNF could be used as a biomarker for diabetes and obesity in future studies or could be exploited as a target for drug development.

Rao et al., was a straightforward use of sequence alignment and provided some insight on how this could be used for diabetes research. However, the article was not in depth in its description of its techniques and motivations. Therefore we looked for other articles that discussed sequence alignment and diabetes. We found Dawson et al. [6], which provided us a well described study that used sequence alignment and other bioinformatics tools to analyze the function of newly discovered gene GLUT10, which they believe has a role in Type 2 diabetes.

The Dawson article is extremely detailed the methodology used to isolate the human and mouse genes used in the study. Since the process of isolating genes is not the focus here, we will only briefly describe it. Using a partially identified transcript from the Sanger Center, the researchers searched NCBI for similar matches in. The information gathered from these searches were used to develop primers which were then used to amplify the GLUT10 transcript from the cDNA for both human and mouse. Here we see the importance of the NCBI databases in this type of research. Without a central repository for sequence information, many experiments would be impossible in the short term and difficult in the long run. In addition to isolating the sequences, the functional analysis of the gene was tested in Xenopus laevis subjects by removing the healthy oocytes and adding insulin to them to observe their reactions.

After the data was collected the analysis was performed. Several bioinformatics tools were employed at this stage. First, the DNA sequences were translated to amino acid sequences and the resulting sequences of residues were aligned with the programs *align* and *BOXSHADE*. Their results revealed there is a 77.4%

identity between human and mouse GLUT10. In addition, it was discovered that GLUT10 was almost the same length as GLUT9, another transporter gene in the family. Using pairwise alignment (with GCG bestfit), they determined that GLUT10 shares between 31% and 35% identity with other GLUTX genes. Further, they used cluster analysis to graphically view the relationship between GLUT10 and other members of the GLUTX family. The rest of the paper discusses cursory gene analysis performed on the cDNA sequences.

From this paper in particular, we see firsthand some of the commonly used bioinformatics tools that are engaged in diabetes research and more importantly, how they are used in concert to develop the big picture of a gene or organism.

Another example of sequence alignment is provided by [7]. In this study, Hornum et al. performed microarray analysis (which we will discuss in more detail in the next section) to determine gene expression. Using the sequences obtained from the microarray experiment, they aligned their sequences using BLAST to determine what other biomarkers might exist in the genome that might indicate susceptibility to diabetes. They found at least 10 candidate genes using their biomarkers in this experiment.

3.2 Gene Expression

Most cells in the body contain a full set of chromosomes and identical genes. However, only certain genes are actually "turned on" for different cells that give the cell type particular abilities. The genes that are "turned on" are defined as expressed genes. For many studies investigating which genes are turned on is the focus of the research. In fact, gene expression techniques were the focus of almost half the recent articles we reviewed. Techniques varied from generating genetic linkage maps to microarray analysis. In this section we will present several papers that outlined the use of different gene analysis tools.

Elbers et. al describe their inspection of genetic linkage maps to compare overlapping regions of genes that affect obesity and Type 2 diabetes. A genetic linkage map is a linear map of the positions of genes along a chromosome (see Fig 1).

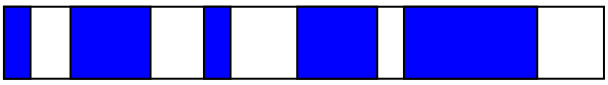

Fig. 1 – example of genetic linkage map (blue areas indicate gene locations).

Not convinced that the genetic linkage maps were that useful for their purposes however, Elbers et al. used disease gene identification tools that are designed to search chromosomes for candidate disease genes. The tools they used in their study included:

(1) Prioritizer
(2) Endeavour
(3) DGP
(4) Geneseeker
(5) G2D
(6) PandS

These tools work in different ways, typically using statistical methods to compare the relationships between genes, but the general goal of all of them is to determine which genes affect or control different diseases. In this case, the researchers were interested in the genes that were involved in both obesity and diabetes. By using all of these tools, they were able to validate the results by comparing how similar they were across tools. In total, they identified 27 candidate genes that could be further studied to determine their role in obesity and diabetes.

Another popular tool for investigating gene expression is microarray analysis. In general microarray analysis involves hybridizing mRNA molecules to the DNA templates from which they originated. By measuring the amount of mRNA bound to each site of the array, they can ascertain how genes are expressed under different conditions, in different tissues, and in different organisms. These have become important tools because several thousand genes can be expressed at one time in one experiment. This facilitates the process of gene study considerably. To analyze microarrays, several statistical techniques are used in concert including sample t tests, log2 ratios, normalization, and ANOVA tests.

Reece et al. describe their use of microarray analysis investigate how maternal diabetes can affect the developing embryo [8]. Since the physical process of developing a microarray is not the focus here, we will concern ourselves with statistical tools used for the analysis. The article specifically states the researchers used the 1-sample t test on log2 ratios to determine significant differences between expression levels. In addition to traditional statistical methods, Reece also used Pathway Assist (http://www.ariadnegenomics.com/products/pathway-studio/), a program designed to visually present the pathways between cell birth and death.

Mazzarelli et al. describe their use of Loess normalization to analyze pancreatic tissue clones [9]. In addition, they used the student t test to compare the microarray data to qRT-PCR data. While this article did not name diabetes specifically, the focus was on pancreatic tissue and islets, which are responsible for insulin production.

It is interesting to mention, that diabetes seems to be used often as prototype disease for developing and validating bioinformatics approaches. This is certainly the case with Collins et al., who write about high throughput biomarker discovery [10]. The goal for their study was to search for biomarkers that might indicate a disease. For their test case, they searched for genes that are known to make organisms susceptible to Type 1 diabetes . By locating these markers, they hope to determine what other factors might be responsible for a person developing diabetes who possesses the susceptibility genes as having the gene alone does not indicate a person will develop the disease. Using traditional microarray experiments to generate data, they used t-tests to compare the data from T1D patients with those who were autoantibody positive (susceptible to develop diabetes, but do not currently have diabetes). The first part of their experiment involved comparing single markers in the different populations, which led to univariate statistical analysis. They reasoned however that using multiple markers would yield more conclusive results, so they employed several additional statistical tools in the discriminant analysis area including, parametric and nonparametric tests (kernel based and K nearest neighbor).

Another tool used for relating proteins to the genes that encode them is transcript mapping which was a technique used by Fossey et al. [11]. In this technique, the researchers used the

greedy search algorithm to determine the best Hamiltonian path through the gene markers which helped them connect genes with proteins they encode. The result was a 6-Mb map of the diabetes linked areas on Chromosome 20.

3.3 Data Mining/Databases

While sequence alignment and microarray analysis techniques are important in diabetes and other medical research, the storage and retrieval of biology related data is key to the exploitation of the data acquired during that research. Supporting that argument, several of the papers we reviewed dealt specifically with the topic of databases that were used or developed for research. In this section we will look at those articles to see how diabetes research is impacted by existence and development of databases.

Perhaps the most common interaction with databases that researchers have is using them. We have already seen several examples of databases (BLAST, genome database, NCBI resources) that have been used during the normal course of research. They are often used to supplement data that cannot be reasonably collected directly by the researchers as was the case for Craig et al. in [12]. In their study, they screened several exon regions from the DNA of 48 African Americans and 48 Caucasians from which they identified 21 single nucleotide polymorphisms. In addition to the 21 SNPs they observed, they used public databases (NCBI) to located other SNPs on which they could perform their study. By drawing from other sources, they were able to combine their data with a wider resource to get a better picture of the binding protein in which they were interested. Without such resources, this type research can be difficult and possibly less convincing because research completed with smaller datasets can be suspect depending on the context of the study.

Diabetes is a disease that has both genetic and environmental causes. While much of the research performed by biologists today focus on the genetics or biological environmental data, there is some argument for using historical data to establish patterns of disease pathobiology. Humphreys et al. did this in [13]. Data collected in written form by the Center for Population Economics at the University of Chicago regarding Civil War veterans, were digitized for convenient study. Prior to the digitization project, searching the data and deriving meaningful statistics was difficult to do, if not impossible for some research projects. The dataset contains over 35,000 records on white soldiers and 6000 records on black soldiers culled from data that was originally housed at the National Archives in Washington D.C. The medical records consisted of applications for benefits on which the results of a physical examination was recorded. To determine if the soldiers had diabetes, they searched the field that was specifically designed to code for endocrine diseases and searched the comment fields for the words diabetes, glucose, or sugar.

The methods used in this study demonstrate the importance of using every method available for researching diabetes. While the biology of the disease was not studied, patterns of diagnosis and treatments as well as lifespans and related problems could be established. For instance, they discovered that the incidence of diabetes was lower in blacks than whites, although not by much. However, that situation has significantly changed today and now African-American men are much more likely to develop diabetes than caucasian men. This finding underscores the impact exercise has on diabetes because most of the black veterans had jobs that involved daily physical labor, and were therefore in better physical shape overall than their white counterparts.

The data gathered in the Humphreys study was based on historical data that was incomplete and not completely homogenous. There were after all, over 17,000 whites and only around 1700 blacks reflected in the study, Moreover this was a homogenous population of only Army veterans, a group which might have particular attributes not found in other populations thus leading to potential biased research results. Other modern database registries suffer from the same problem; the data contained therein is only from a specific population, which as mentioned before can lead to research results that may be biased towards that population. To tackle that problem, Zgibor et al. performed a study at the University of Pittsburgh Medical Center that focused on validating heterogeneous data on diabetics [14]. For this study an enormous amount of data was searched including 46,082,941 lab reports, 233,292,544 medical records, and 9,351,415 medical record abstracts which represented around 2 million patients. As with the Humphreys study, they searched for data indicative of diabetes, however, they were able to be more precise because they had better clinical data with which to work. For instance, they were able to use actual ICD-9 codes, A1c results, and diabetes medication keywords to determine whether or not a patient was diabetic.

The purpose of the study was to validate the data and ensure that they were able to draw a heterogeneous dataset from which further diabetes studies could be done that would not be biased towards a single population. They validated their results by actually pulling charts and comparing data from the charts versus data acquired from their database. Using the Students t-test and Pearson's Chi square test, they compared the two datasets to determine the level of difference between them. The results indicated the registry information they wanted to assemble was valid and could be used to derive datasets on the whole population or subpopulations with reasonable certainty that all relevant data would be extracted with little extraneous data included. This registry is the first of its kind because it contains heterogeneous data and could be much more useful for general studies on diabetes.

3.3.1 Important Diabetes Information Sources

We have established in the previous section that diabetes datasets are essential to modern diabetes research. As a result, some databases that are available to researchers bear some discussion. While the list of databases here is not exhaustive, it represents several that are mentioned in the literature or that contain unique information.

Resource:	**National Diabetes Information Clearinghouse**
Address:	http://diabetes.niddk.nih.gov/about/
Scope:	Online articles
Audience:	Patients Health care professionals General public
Description:	The NIDDK provides this resource that is targeted more to the general population than diabetes researchers. However, the site does contain statistics on diabetes incidence in the United States that was gathered by the National Center

for Health Statistics that could be useful for research. In addition, the site contains a section on Clinical Trials that are currently in progress at the NIDDK.

Resource: **American Diabetes Association**
Address: http://www.diabetes.org
Scope: Articles and research database
Audience: General public
Health care professionals and scientists
Description: The American Diabetes Association's website is probably the premiere website for information about diabetes for the general public. It contains not only information about the disease, but also treatment options, recipes, and exercise routines. For diabetes researchers, the site contains a database of research supported by the ADA and other organizations. While these are not the actual research results since the studies are ongoing, they provide a good overview what is currently happening in research. The site also provides a list of research articles that have been recently published on diabetes. Although access from ADA is not possible, the site directs the researcher to publisher's site so the article can be purchase or searched for in an academic library's resources. Like the NIDDK site, ADA provides information about clinical trials and what is involved in joining a clinical trial.

Resource: **Juvenile Diabetes Research Foundation**
Address: http://www.jdrf.org
Scope: Articles about life with diabetes
Research opportunities for scientists
Publications for both the general public and scientists offered for sale.
Audience: General public
Health care providers and scientists
Description: The Juvenile Diabetes Research Foundation is a non-profit research organization dedicated to finding a cure for diabetes. Much of the site is targeted to the general public as well as fund-raising since it is a non-profit. For the professional researcher, the most useful information JDRF provides is about grant opportunities they provide. Researchers can apply for grants through the site as well as monitor grants they have been awarded.

Resource: **Jackson Laboratory Type 1 Diabetes Mouse Repository**
Address: http://www.jax.org/resources/index.html
Scope: Mouse Strain Information
Mouse Service Information
Audience: Professional researchers/scientists
Description: The Jackson Laboratory is a non-profit organization dedicated to biomedical research. It provides access to myriad data on the genetics, genomics and biology of the laboratory mouse as well as performs research in several disease areas including diabetes. For the professional diabetes researcher this site provides one of the most comprehensive datasets we have seen and it essential for those doing laboratory research with mice.

Resource: **EPConDB PancChip [15]**
Address: http://www.betacell.org/resources/data/epcondb/
Scope: Gene expression data for genes expressed in the pancreas.
Audience: Professional researchers/scientists
Description: The Endocrine Pancreas Consortium Database is sponsored by the Beta Cell Biology Consortium originally established by the National Institute of Diabetes and Digestive and Kidney Diseases in 2001. It contains information about genes that expressed in the cells of the pancreas and their transcriptional regulation. It uses the Database of Transcribe Sequences, which is a human and mouse transcript index for it transcription database. It can present results obtained from microarray data. It can be used to find similar genes to a gene of interest and search results provide access to other databases and tools such as BLAT, Genecard, and Entrez for further information.

Resource: **T1Dbase [16]**
Address: http://T1DBase.org
Scope:
Audience:
Description: The T1Dbase Database collects information from various sources including the Beta Cell Gene Bank, BLAT, and others. It also provides software tools for analyzing genome data. Unlike the EpconDB, this database aggregates information from more than just microarray experiments. Some of the specific tools it provides includes:

(1) Gbrowse – an open source genome browser
(2) Gene Dossier – a tool that provides a way to view data on genes from several datasets at once
(3) T1Dmart – a query tool for finding biomarker and genotyping data
(4) The Microarray Viewer – provides gene expression data from over twenty experiments
(5) The Tissue Expression Viewer – shows gene expression across a range of tissues
(6) Cytoscape – visualization tool for viewing and analyzing biological networks

Of all of the databases we reviewed, this database seems to be the most comprehensive and encompassing for diabetes research and seems

similar in scope to NCBI's tools in that it seeks to provide everything in one interface. In fact, we spent more time on this site looking at its features than any other because there was so much to look at!

3.4 Summary

Tables 1 – 3 summarize the results of investigation of bioinformatics tools used by various researchers. Not surprisingly many of the papers cited the use of some sort of database during some phase of their research. Except for the articles that were specifically written about T1Dbase and EPConDB, none of the other mentioned using these tools. We were a little surprised by this at first because both have been around for at least a couple of years. We surmised since they were new and it takes time to collect enough data and interfaces to other data sources to be useful this might account for that situation. We expect to see more references to both of those tools in future diabetes research papers because they offer convenient and direct access to only diabetes data which can help reduce the information overload experienced by researchers in this area.

We also observed that much diabetes research focuses on the studying gene expression. From Table 1 we see five articles enumerated microarray analysis as a primary tool for analysis. In addition, a chief subject of the T1Dbase and EPConDB databases are gene expression tools. This is probably because now that much genome sequencing has been done, researchers are moving to the next phase which involves making sense of the genes that are located on the genome and how they are expressed under different conditions and across different organisms.

Paper	Sequence Alignment	Cluster/Phylogenetic Tree Building	Gene Expression/Microarray Analysis	Database Searches/ Database Assembly
[5]	X	X		
[6]	X	X	X	X
[7]	X			X
[8]			X	
[9]			X	X
[10]			X	
[11]			X	
[12]				X
[13]				X
[14]				X
[15]				X
[16]				X
Totals	**3**	**2**	**5**	**8**

Table 1 – Bioinformatics techniques used in diabetes research

Table 2 shows the many different tools that were mentioned in the articles we reviewed (this does not include the databases we already discussed in Section 3.3.1. Again, we did not expect the results we received. Not a single tool was mentioned more than once across our review of the literature except ClustalW/X which are related programs. Particularly in the area of gene expression and microarray analysis, we expected to see a recurrence of commonly used tools, but we did not. In fact, in most cases, the researchers did not indicate the tools they used for microarray expression analysis. It is assumed any number of tools like Excel, Stata, SPSS or other statistical packages could have been used and were therefore not considered worthy of mention in the analysis section. The tools that were mentioned included a variety of packages including sequence alignment tools, disease gene identification tools, cluster analysis tools and graphical display tools.

Generally speaking, most researchers seemed to use several unrelated tools in their research. However, in one case, several bioinformatics and statistical tools that were used by the researchers appeared in one package. Cho et al. outline their use of CiphergenExprress which is one such tool [17]. CiphergenExpress includes principle component analysis (PCA), heat maps, receiver-operator characteristics (ROC), scatter plots, and Box and Whisker plots. At first glance we were surprised at this finding. However, upon further consideration it occurred to us that unified commercial tools are probably outside of the financial reach of many labs. Most of the other tools mentioned are free tools that are sponsored by the lab that originally created them or that have been freely distributed for academic use.

Program Name	Number of Times Mentioned
Align	**1**
BOXSHADE	**1**
CGC Pileup	**1**
ClustalW	**1**
ClustalX	**1**
Prioritizer	**1**
Endeavour	**1**
DGP	**1**
Geneseeker	**1**
G2D	**1**
PandS	**1**
Pathway Assist	**1**
Graphmap	**1**
CipherGen Express	**1**

Table 2 – Bioinformatics software used in diabetes research

Not every article mentioned the statistical tools employed in the analysis of the data, although several did. For the most part, researchers used simple statistics tools for analysis. Table 3 summarizes the statistical tools that were specifically mentioned in the

articles we reviewed. For instance, the T-test was often mentioned as most analyses were performed against two different datasets. In at least one case however, the application of multivariate statistical techniques was necessary and a package that was able to do principle component analysis (CipherGen Express) was used. Other basic techniques that were mentioned were scatter plots and the Pearson Chi-square test. We were not surprised by these results because often the goal of statistical analysis is to determine the significant different between two populations which does not necessarily involve complicated tests. For large studies principle component analysis might be necessary if the researchers need to look at many dimensions of the data at one time. But in the case of microarray analysis, researchers are generally comparing gene expression under two different conditions which vitiates the need for complicated statistics and tools.

Statistical Tool	Number of Times Mentioned
T-tests	**5**
K nearest neighbor	**1**
Logistic regression	**2**
PCA	**1**
Heat maps	**1**
ROC	**1**
Scatter Plots	**1**
Box and Wisker Plots	**1**
Chi Square Test	**1**

Table 3 – statistical techniques used in diabetes research

It is quite clear from our research then, that interaction with databases is the only common thread that runs through the majority of the research we reviewed. The importance of continued development and exploitation of such resources cannot be overstated. In addition to database resources, numerous statistical techniques are employed to analyze data, most notably the t-test and then in general relatively straight forward and simply statistical tests are used for analysis. Also a large number of different programs are used in diabetes research, although none of those that were mentioned seem to be used universally used by the majority of researchers. Future research in this area might involve developing a diabetes research toolset that researchers in this field use so results can be more easily compared across studies.

ACKNOWLEDGEMENTS

This technical report is an extension of a paper written by RK to fulfill a requirement for CECS 660 Introduction to Bioinformatics during the summer of 2007. RK would like to thank Christy Bogard for her assistance with some of the topics covered early in the course while I was getting my bearings in this area. ER is supported by NIH-NCRR grant P20RR16481 and NIH-NIEHS grant P30ES014443-01A1. The contents of this manuscript are solely the responsibility of the authors and do not represent the official views of NCRR, NIEHS, or NIH.

REFERENCES

[1] D. Brar, "The History of Diabetes." vol. 2007: INTERNATIONAL ISLET TRANSPLANT REGISTRY 2007.

[2] Dictionary of Biology. London: Constable and Robinson Ltd., 2005.

[3] "Total Prevalence of Diabetes in the United States, All Ages 2005," National Institute of Diabetes and Digestive and Kidney Diseases

[4] F. Sanger and E. O. P. Thompson, "The amino acid sequence in the glycyl chain of insulin," Biochemistry Journal, vol. 53, pp. 353-366, 1953.

[5] A. A. Rao, G. R. Sridhar, B. Srinivas, and U. N. Das, "Bioinformatics analysis of functional protein sequences reveals a role for brain-derived neurotrophic factor in obesity and type 2 diabetes mellitus," Medical Hypotheses, vol. In Press, Corrected Proof.

[6] P. A. Dawson, J. C. Mychaleckyj, S. C. Fossey, S. J. Mihic, A. L. Craddock, and D. W. Bowden, "Sequence and Functional Analysis of GLUT10: A Glucose Transporter in the Type 2 Diabetes-Linked Region of Chromosome 20q12-13.1," Molecular Genetics and Metabolism, vol. 74, pp. 186-199, 2001.

[7] L. Hornum and H. Markholst, "A Sequence-Ready PAC Contig of a 550-kb Region on Rat Chromosome 4 Including the Diabetes Susceptibility Gene Lyp," Genomics, vol. 69, pp. 305-313, 2000.

[8] E. A. Reece, I. Ji, Y.-K. Wu, and Z. Zhao, "Characterization of differential gene expression profiles in diabetic embryopathy using DNA microarray analysis," American Journal of Obstetrics and Gynecology, vol. 195, pp. 1075-1080, 2006.

[9] J. M. Mazzarelli, P. White, R. Gorski, J. Brestelli, D. F. Pinney, A. Arsenlis, A. Katokhin, O. Belova, V. Bogdanova, E. Elisafenko, M. Gubina, L. Nizolenko, P. Perelman, M. Puzakov, A. Shilov, V. Trifonoff, N. Vorobjeva, N. Kolchanov, K. H. Kaestner, and J. C. J. Stoeckert, "Novel genes identified by manual annotation and microarray expression analysis in the pancreas," Genomics, vol. 88, pp. 752-761, 2006.

[10] C. D. Collins, S. Purohit, R. H. Podolsky, H. S. Zhao, D. Schatz, S. E. Eckenrode, P. Yang, D. Hopkins, A. Muir, M. Hoffman, R. A. McIndoe, M. Rewers, and J. X. She, "The application of genomic and proteomic technologies in predictive, preventive and personalized medicine," Vascular Pharmacology, vol. 45, pp. 258-267, 2006.

[11] S. C. Fossey, J. C. Mychaleckyj, J. K. Pendleton, J. R. Snyder, J. T. Bensen, S. Hirakawa, S. S. Rich, B. I. Freedman, and D. W. Bowden, "A High-Resolution 6.0-Megabase Transcript Map of the Type 2 Diabetes Susceptibility Region on Human Chromosome 20," Genomics, vol. 76, pp. 45-57, 2001.

[12] R. L. Craig, W. S. Chu, and S. C. Elbein, "Retinol binding protein 4 as a candidate gene for type 2 diabetes and prediabetic intermediate traits," Molecular Genetics and Metabolism, vol. 90, pp. 338-344, 2007.

[13] M. Humphreys, P. Costanzo, K. L. Haynie, T. Ostbye, I. Boly, D. Belsky, and F. Sloan, "Racial disparities in diabetes a century ago: Evidence

from the pension files of US Civil War veterans," Social Science & Medicine, vol. 64, pp. 1766-1775, 2007.

[14] J. C. Zgibor, T. J. Orchard, M. Saul, G. Piatt, K. Ruppert, A. Stewart, and L. M. Siminerio, "Developing and validating a diabetes database in a large health system," Diabetes Research and Clinical Practice, vol. 75, pp. 313-319, 2007.

[15] J. M. Mazzarelli, J. Brestelli, R. K. Gorski, J. Liu, E. Manduchi, D. F. Pinney, J. Schug, P. White, K. H. Kaestner, and C. J. S. Jr., "EPConDB: a web resource for gene expression related to pancreatic development, beta-cell function and diabetes," Nucleic Acids Research, vol. 35, pp. 751-755, 2007.

[16] E. M. Hulbert, L. J. Smink, E. C. Adlem, J. E. Allen, D. B. Burdick, O. S. Burren, C. C. Cavnor, G. E. Dolman, D. Flamez, K. F. Friery, B. C. Healy, S. A. Killcoyne, B. Kutlu, H. Schuilenburg, N. M. Walker, J. Mychaleckyj, D. L. Eizirik, L. S. Wicker, J. A. Todd, and N. Goodman, "T1DBase: integration and presentation of complex data for type1 diabetes research," Nucleic Acids Research, vol. 35, pp. 742-746, 2007.

[17] W. C. S. Cho, T.-T. Yip, W.-S. Chung, S. K. W. Lee, A. W. N. Leung, C. H. K. Cheng, and K. K. M. Yue, "Altered expression of serum protein in ginsenoside Re-treated diabetic rats detected by SELDI-TOF MS," Journal of Ethnopharmacology, vol. 108, pp. 272-279, 2006.