DIMACS Technical Report 2004-37

August 2004

Report on DIMACS Working Group Meeting: Data Mining and Epidemiology

March 18-19, 2004

by

James Abello[1,2]

Graham Cormode[3,4]

Center for Discrete Mathematics and Computer Science

Rutgers University, Piscataway NJ

[1] Permanent Member, DIMACS
[2] Abello@dimacs.rutgers.edu
[3] DIMACS Postdoctoral Fellow
[4] graham@dimacs.rutgers.edu

Abstract

Epidemiology is an observational science that concerns itself with finding and explaining patterns of health and disease in populations, usually of humans, but also populations of animals, insects and plants. Data mining is an active area of research interested in finding algorithms for describing latent patterns in often very large data sets. This Working Group had the objective of fostering collaboration between these two disciplines. In March of 2004 it organized a two-day meeting at DIMACS to bring these two groups together in a format designed to initiate such collaborations.

Report on DIMACS Working Group Meeting:

Data Mining and Epidemiology

March 18-19, 2004

Introduction

Epidemiology is an observational science that concerns itself with finding and explaining patterns of health and disease in populations, usually of humans, but also populations of animals, insects and plants. Data mining is an active area of research interested in finding algorithms for describing latent patterns in often very large data sets. This Working Group had the objective of fostering collaboration between these two disciplines. In March of 2004 it organized a two-day meeting at DIMACS to bring these two groups together in a format designed to initiate such collaborations.

Meeting Format

Each day there were a number of presentations from epidemiologists, statisticians and computer scientists. They gave presentations on methods current in today's epidemiological practice. These included descriptive, molecular and veterinary epidemiology; propensity score techniques, association rules and graph theoretical methods; pharmacovigilance, infectious crop disease and case-control surveillance.

The goal was to identify some fruitful areas for applying new techniques from theoretical computer science, discrete mathematics, and statistics and to allow new collaborations among participants to be initiated. This workshop brought together with epidemiologists experts interested in exploring data mining and algorithmic techniques potentially useful in epidemiology from a range of viewpoints. The aim was to identify fundamental epidemiological problems that can benefit from efficient computational, statistical and mathematical models that can aid in the processing and understanding of combined epidemiological, and immunological data.

There were 37 participants from academia, industry and the US Forestry Service. Slides of the presentations are posted at

http://dimacs.rutgers.edu/Workshops/WGDataMining/material/ .

Working Subgroups

Out of 8 suggested topics for discussion the participants assembled themselves on 5 focused working subgroups. These were:

- **Descriptive and Analytical Epidemiology**
- **Surveillance and Epidemic Detection**
- **Text Data Mining**
- **Biostatistics, Molecular and Genetic Epidemiology**
- **Spatial and non-human Epidemiology**

Each group was asked to do the following:

(a) Produce a list of concrete epidemiological questions (3 to 5) and corresponding data mining or statistical techniques that are suitable to address them.

(b) Explain the concrete problem in some detail and the corresponding computational challenges. This may include interface or visualization tools that the group envisions as helpful to epidemiological data analysis.

(c) Describe the impact a computational solution will have on the original epidemiological question.

Please see the findings of the working subgroups which follow this preface.

Concluding Remarks

One of the concrete outcomes of this working group is the preparation of an AMS-DIMACS volume with chapters contributed by presenters at this meeting and beyond. In many cases, these will expand upon the presented material, and will lay the groundwork for further explorations of the interface between Data Mining and Epidemiology. It was clearly recognized during this meeting that Data Mining technologies have a lot to offer to the Epidemiological community. This cannot come to fruition without further concerted efforts including specific pilot projects with real data sets and specific questions to answer. This Working Group has made the first steps in this direction, and we look forward to seeing further progress.

Working Subgroup on Descriptive and Analytical Epidemiology

James Abello (DIMACS)
Dave Ozonoff (Boston University)
Alex Pogel (New Mexico State University)
Greg Ridgeway (Rand Corporation)
Dona Schneider (Rutgers University)

Introduction

Epidemiology is concerned with patterns in populations, whether those patterns are encoded in descriptions or in causal associations between particular features such as exposures and disease outcomes (Shannon, Schneider). The essence of data mining techniques is also to find patterns in data using automated or machine assisted techniques (Madigan). At the same time, conventional methods of data analysis using statistical techniques show limitations in certain situations now more commonly encountered in epidemiology (e.g., massive data sets or very small data sets). We identified several typical and important epidemiological objectives potentially amenable to data mining techniques in the workshop presentations.

Concrete questions

Epidemiological questions and data mining or statistical techniques suitable to address them

1. Discovering patterns in massive data sets
(*e.g., micro-array data, data routinely collected for other purposes such as billing of patient encounters, Medicaid data, workers' compensation claims*).

Extremely large (in epidemiologic terms of reference) data sets present special problems for traditional methods of epidemiologic analysis. The advent of inexpensive and massive data storage technologies and the compilation of many routinely recorded data sources, like Medicaid or other insurance records or microarray data has stimulated inquiry into new techniques for analyzing these data, since conventional statistical techniques are frequently inappropriate (Madigan, Shannon). For example, the huge populations lead to the identification of many associations as "unlikely" under the null hypothesis if conventional criteria are used (e.g., $p < .05$) but where the usual corrections, like the Bonferroni technique, would produce draconian results and potentially lead to the elimination of many interesting associations (Shannon). Shannon, Rucsinski, and Imielinski discussed several techniques that might be useful in this instance. Since data mining developed in the context of machine learning and have frequently been applied to large, routinely collected data sets (e.g., in marketing) this would seem a fruitful area for application.

2. Discover patterns or associations in very small (too small to use large sample statistics) data sets

The "dual" of the large data set problem (i) is the very small data set, where the large sample approximations of conventional statistical techniques also break down. Such situations are frequently encountered in practice where relatively small populations are exposed and the (important) public health question is whether they have suffered any harm. A typical example might be the contamination of a well that serves a neighborhood-sized population. Data mining techniques, although originally developed to find patterns in large data sets are also amenable to finding them in small ones. Two papers presented approaches to this problem (Ozonoff/Pogel, Abello) using results from lattice theory and graph theory, respectively. Both techniques also provide novel methods for visualizing aspects of data structure that might be used heuristically for later follow-up (cf. 4, below).

3. Causal inference from observational studies that take advantage of data mining techniques to make adjustments between otherwise incomparable groups

Much of epidemiology concerns discerning whether a certain exposure is a cause of an outcome. While the randomized controlled experiment is the gold standard, for many exposures and outcomes of interest randomization simply is not possible. This includes assessing the effect of exposure of smoking, water contamination, smog, race, sex, and policies on outcomes such as contraction of disease, survival, access to health care, and employment. As a result, any efforts to assess the effect of an exposure must rely on laboratory studies (analysis with "hard science" or experiments on laboratory animals) or observational data. Existing methods for assessing the causal effect of an exposure from observational data include covariate adjustment, instrumental variable methods, and propensity scores.

Covariate adjustment: Covariate adjustment involves fitting a predictive model of the outcome of interest (y) from exposure to a "treatment" (t) and potential confounders ($\mathbf{x}$). These models have the form $y = f(t, \mathbf{x}) + \varepsilon$. The data mining and machine learning communities have generated numerous methods, in addition to the rather standard linear models, for estimating $f(t, \mathbf{x})$. If the method offers a good estimate of $f(t, \mathbf{x})$ and $\mathbf{x}$ captures all of the important confounders, variables associated with both the treatment and the outcome, then we can utilize that estimate to evaluate the causal effect of t.

Instrumental variables: Every so often an observational study contains a variable that acts as a "pseudo-randomizer." For example, to evaluate a child-care program for two child families making less than $20,000, one could compare two child families making $20,000 to $21,000 with those making $19,000 to $20,000. One could consider that families might fall into either one of these groups by chance and will likely be similar on either side of the $20,000 boundary. These instances represent a very small fraction of observational studies, but when possible instrumental variable methods offer nice causal effect estimates. Angrist and Levy (1999) uses instrumental variables for assessing the causal effect of class size on test scores. Grogger and Ridgeway (2004) use variation in natural

lighting to assess the causal effect of race identification in police deciding which vehicles to stop.

J. Angrist and V. Levy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement," Quarterly Journal of Economics, 533-575.

J. Grogger and G. Ridgeway (2004, submitted). "Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness."

Propensity score methods: Ideally we would like to compare subjects exposed to a treatment to control subjects that are identical to treatment subjects in all ways except exposure to treatment. Rosenbaum and Rubin (1983) defined the propensity score, $p(\mathbf{x})$, as the probability that a subject with features $\mathbf{x}$ is in the treatment group. They showed that it is sufficient to match on the unidimensional $p(\mathbf{x})$ rather than the multidimensional $\mathbf{x}$. The challenging task is to estimate $p(\mathbf{x})$ from the observed data where x may consist of continuous, ordinal, nominal, and missing values. Fitting such models has been the focus of data mining and machine learning researchers for the last decade. Merging the statistical techniques involving propensity scores and the data mining/machine learning techniques for estimating the propensity scores is a promising convergence of the two fields. A successful merger of the methods and fields can solve many important applied epidemiology problems. McCaffrey et al (2004) and Morral et al (2004) describe one method for combining propensity score and machine learning techniques for assessing the causal effect of a particular drug treatment program.

P.R. Rosenbaum and D.B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects," Biometrika, 70, 41–55.

D. McCaffrey, G. Ridgeway, and A. Morral (2004, submitted). "Propensity Score Estimation with Boosted Regression for Evaluating Adolescent Substance Abuse Treatment"

A.R. Morral, D.F. McCaffrey, and G. Ridgeway (2004, to appear). "Effectiveness of Community-Based Treatment for Substance Abusing Adolescents: 12-month Outcomes From A Case-Control Evaluation of a Phoenix Academy," Psychology of Addictive Behaviors.

4. How to find interesting associations present in the data that are fruitful for follow-up

(for use in the construction of more elaborate models or for detailed validation)?

When data is sufficiently small (or is sampled to be so), one hypothesis generation method involves viewing the concept lattice. This allows the analyst to gather a variety of insights regarding the data. First, through the support of the concepts taken separately, all possible n-way attribute value combinations (this includes all values less than n) are presented, giving a global one-dimensional frequency table. Second, through the edges between concepts (the edges of the Hasse diagram of the concept lattice), the lattice

describes the classes of equal confidence values of association rules expressed through the lower vertex of the ordered pair. Third, each diamond (with bottom at A-and-B, middle vertices at A and at B, respectively, and top at U, the universe of subjects) represents a 2x2 contingency table extracted from the universal *n*-way contingency table.

Current work by Pogel, Ozonoff, and others aims to extend the expressive power of the concept lattice and to create epidemiologically focused manipulation methods that enhance the usability of the lattice.

When data has a moderate size, other approaches are employed before considering any use of the concept lattice. This is because a weak point of lattice-centered data analysis is that the lattice is usually viewed on a computer monitor (hand-calculations are only reasonable in the smallest of examples; see LatDrawWin by Ralph Freese and Concept Explorer by Serhiy Yevtushenko for some automated lattice drawing tools) and the size of the lattices quickly exceeds the available number of screen pixels.
More generally, the concept lattice computation simply generates too large a number of concepts to easily manage. These complexities concerns create a need for some control to be exercised with regard to how many concepts are computed at a time. This leads us to examine decomposition methods to apply to the given input binary relation that is the usual initial datum for the concept lattice construction. In particular, we use several graph decomposition methods including breadth first search, graph cores and graph cuts. We are building a body of theoretical results that describe the relationship between the data decomposition and the corresponding formal concept lattice. An important aspect of these methods is that each refers to some (inexpensively) computed structural aspect of the input data, gathered via one of the decomposition methods.

When the data is massive, more data-driven overview methods are necessary. For example, localized density computations and searches for quasi-cliques in sparse data can yield subgraphs that are sufficiently small to begin to examine with semantic interpretation. Again, the graph decomposition methods described above (BFS, cores and cuts) can be applied. Remaining challenges when massive data is present are to determine epidemiological criteria that inform the decomposition methods and to maintain awareness of the analytic effect of the sampling activity that results when we dismiss a larger portion of the data in order to reduce it to a usable, viewable level.

A variety of references are in the list below.

Agrawal, R., Imielinski, T. and A. Swami, A., Mining association rules between sets of items in large databases, in ACM SIGMOD Intl. Conf. Management of Data, May 1993.

J. Abello, A. Pogel, L. Miller, Graph Partitions and Formal Concept Lattices, Submitted to Journal of Universal Computer Science, Under Revision.

J. Abello, J. Korn: MGV: A System for Visualizing Massive Multidigraphs, IEEE Transactions on Visualization and Computer Graphics {\bf 8} No. 1, January-March 2002.

J. Abello, J. Vitter (eds): External Memory Algorithms, Vol 50 of the AMS-DIMACS Series in Discrete Mathematics and Theorethical Computer Science, 1999.

J. Abello, M. Resende, and S. Sudarsky: Massive Quasi-Clique Detection, In Proceedings of Latinoamerican Informatics, Springer Verlag LNCS, May 2002.

A. Berry and A. Sigayret: Representing a concept lattice by a graph, Proceedings of Discrete Maths and Data Mining Workshop, 2nd SIAM Conference on Data Mining (SDM'02), Arlington (VA), April 2002.

A. Berry and A. Sigayret: Obtaining and maintaining polynomial-sized concept lattices, Proceedings of Workshop FCAKDD (Formal Concept Analysis for Knowledge Discovery in Data bases), ECCAI 02.

For an extensive reference list of FCA application papers, see http://www.mathematik.tu-darmstadt.de/ags/ag1/Literatur/literatur_en.html.

Birkhoff, G., Lattice Theory, American Mathematical Society, Providence, R.I., 1st edition, 1940.

Ganter, B. and Wille, R., Formal Concept Analysis: Mathematical Foundations, Springer, NY, 1999. ISBN 3-540-62771-5

Freese, R. LatDrawWin, a lattice drawing applet, www.math.hawaii.edu/~ralph/LatDraw.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N. and Lakhal, L.,Computing iceberg concept lattices with Titanic, Data and Knowledge Engineering (Elsevier), 42 (2002), pp. 189-222.

Wille, R., Why Can Concept Lattices Support Knowledge Discovery in Databases?, Technische Universitat Darmstadt, Fachbereich Mathematik, Preprint Nr. 2158, June 2001. Available at http://wwwbib.mathematik.tu-darmstadt.de/Math-Net/Preprints/Listen/pp01.html

Yevtushenko, S., et al, Concept Explorer, Open source java software available at ttp://sourceforge.net/projects/conexp, Release 1.2, 2003.

Zaki, M., and M. Ogihara, M., Theoretical Foundations of Associations Rules, in Proceedings of 3rd SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'98), Seattle, Washington, USA, June 1998.

5. For large sets of attributes in a data set, how to select variables for statistical models

Variable selection in statistical model construction is an important problem in epidemiology. Thus, in a regression, the problem is to find those variables and

combinations of variables to enter into the regression, taking into account their relevance to the outcome, the problem of multicollinearity and the problem of interpreting complex and higher order interactions. Papers by Ruczinski, Ozonoff/Pogel and Ridgeway discussed various approaches to considering higher order interactions in microarray (Imielinski), genetic (SNP) data (Ruczinski, Ozonoff/Pogel) and collinearity (Ozonoff/Pogel). Association rules, logical analysis techniques and lattice techniques were among the approaches.

Modeling strategies that put absolute penalties on the absolute magnitude of regression coefficients have the effect of setting some or many of the coefficients equal to 0. For example, the Lasso estimate of the regression coefficients for a fixed penalty, λ,

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta' \mathbf{x})^2 + \lambda \sum_{j=1}^{J} | \beta_j |,$$

has some $\beta_j = 0$ for some values of λ small enough. Tibshirani (1996) and Poggio and Girosi (1998) recognized early on that the absolute penalty had this variable selection property, which eventually led to similar properties for support vector machines. Recently, Efron et al (2004) have extended the ideas to a simple algorithm with computation of the same order as an ordinary least squares fit. Further research should investigate this method as to what extent this solves or contributes to the solution of the variable selection problem.

R. Tibshirani (1996). "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58(1):267-288.

T. Poggio and F. Girosi (1998). "A Sparse Representation for Function Approximation," *Neural Computation*, 10(6).

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani (2004). "Least Angle Regression," *Annals of Statistics*, 32(2).

Working Subgroup on Surveillance and Epidemic Detection

Michael Cook (Merck Laboratories)
A. Lawrence Gould (Merck Laboratories)
David Madigan (Rutgers University)

The Adverse Event Reporting System (AERS) is a database designed to support the FDA's post-marketing safety surveillance program for all approved drug and therapeutic biologic products. The FDA receives adverse drug reaction reports from manufacturers, physicians, and consumers. Pharmaceutical companies maintain similar systems. The primary purpose is to identify potential toxicities associated marketed drugs and vaccines. These databases can be large - AERS, for example, contains over five million reports.

A number of procedures exist for identifying potential associations but they have various limitations. A critical requirement of any new procedure is that its output can be readily understood and utilized by clinical epidemiologists and physicians.

Q1 With large-scale spontaneous reporting systems, are there better methods for identifying potential drug-adverse event relationships than currently exist (e.g., GPS/MGPS)?

Current techniques consider marginal distributions of small numbers of drugs and adverse events, most commonly one drug and one adverse event. This approach is susceptible to Simpson's Paradox as well as issues associated with multiple testing. Two lines of research might prove fruitful - building on the marginal approach, are there methods for reducing confounding due to concomitant drugs and other factors - multivariate models that predict the adverse risks as a function of all the drugs and covariates.

Q2 Can we use linguistic or other techniques to automate the coding of adverse events

Q3 Can we develop algorithms to clean verbatim spontaneous adverse event reports (e.g. Appendix 1 lists verbatim descriptions of just two drugs).

Q4 Hierarchical organization of both drugs and adverse events via data mining

Grouping drugs by class and/or pharmacologic activity as well as grouping related adverse events has the potential to improve the sensitivity for detected potential drug/adverse event associations.

Q5 Use data mining techniques to detect duplicate adverse event reports

APPENDIX 1

ADVAIR
ADVAIR (ADVAIR)
ADVAIR (FLUTICASONE PROPIONATE/SALMETEROL XINAFOATE)
ADVAIR (SALMETEROL XINAFOATE; FLUTICASONE PROPIONATE)
ADVAIR (SALMETEROL/FLTICASONE)
ADVAIR (SALMETEROL/FLUTICASONE)
ADVAIR DISC (SERETIDE MITE)
ADVAIR DISKUS
ADVAIR DISKUS 100/50
ADVAIR DISKUS 250/50
ADVAIR DISKUS 500/50
ADVAIR HFA
ADVAIR MULTI DOSE POWDER INHALER (FLUTICASONE + SALMETEROL)
ADVAIR MULTI DOSE POWDER INHALER (FLUTICASONE+SALMETEROL)
ADVAIR(SALMETEROL XINAFOATE / FLUTICASONE PROPIONATE)
FLUTICASEON EPROPIONATE (FLOVENT)
FLUTICASONE
FLUTICASONE (FLUTICASONE PROPIONATE)
FLUTICASONE (FLUTICASONE)
FLUTICASONE + SALMETEROL
FLUTICASONE +FLONASE+ NASAL SP
FLUTICASONE INHALER
FLUTICASONE MDI
FLUTICASONE NASAL SPRAY
FLUTICASONE ORAL INHALER
FLUTICASONE PROP
FLUTICASONE PROP INH
FLUTICASONE PROPIONATE
FLUTICASONE PROPIONATE (+) SALMETEROL XI
FLUTICASONE PROPIONATE (FLIXONASE)
FLUTICASONE PROPIONATE (FLONASE)
FLUTICASONE PROPIONATE (FLOVENT)
FLUTICASONE PROPIONATE (FLUTICASONE PROPIONATE)
FLUTICASONE PROPIONATE AND SALMETEROL XINAFOATE
FLUTICASONE PROPIONATE(+) SALMETEROL XI
FLUTICASONE PROPRIONATE . SALMETEROL
FLUTICASONE(FLUTICASONE)
FLUTICASONE+SALMETEROL
FLUTICASONE/SALMETEROL
FLUTICASONE/SALMETEROL 100/50M
FLUTICASONE/SALMETEROL 500/50M
FLUTICASONER + SALMETEROL
SALMETAROL XINAFOATE
SALMETERL ORAL INHALER
SALMETEROL
SALMETEROL (SALMETEROL)
SALMETEROL HYDROXYNAPHTHOATE
SALMETEROL INH
SALMETEROL INHAL AEROSOL
SALMETEROL INHALER
SALMETEROL MDI
SALMETEROL ORAL INHALER
SALMETEROL XINAFOATE
SALMETEROL XINAFOATE (SEREVENT INHALER)
SALMETEROL XINAFOATE (SEREVENT)
SALMETEROL XINAFOATE(SEREVENT)
SALMETEROL XINOFOATE
SALMETEROL/FLUTICASONE
SALMETEROL/FLUTICASONE PROPIONATE
SALMETEROL/FLUTICASONE PROPIONATE 50/500

Working Subgroup on Text Data Mining

Graham Cormode (DIMACS)
Lynette Hirschman (Mitre Corporation)

We identified several potential applications for text mining technology to epidemiological data. The applications fall into several classes:

1. Encoding of free text information into a rich classification scheme (e.g., multi-sentence symptom reports into symptom encodings).
2. Mapping of vocabulary into standardized canonical form, including handling of misspellings, alternate terms, abbreviations (spelling correction, terminology and abbreviation handling)
3. Capture of data relevant for detecting disease outbreaks globally (text classification, information extraction).

Underlying applications 1 and 3 is the need to develop techniques (lightly supervised learning?) that would allow the rapid creation of tailored information extraction and classification systems for specific applications.

1. Mapping of free text information to "encoding"
    a. AERS and VAERS databases have patient symptom data in free text; this is currently manually encoded into MEDRA terms, but this is a large vocabulary, not optimally organized for data clustering (e.g., there is hierarchical organization, but hypothyroidism and hyperthyroidism are aggregated into one category at the next level of the hierarchy.
    Task a1 could be to automate the encoding
    Task a2 could be to use data mining and lightly supervised learning to develop a new improved hierarchy for improved automated encoding.
    This resource would be of interest to the FDA and the pharmas; also to researchers involved in medical text mining and information extraction (Russ Altman at Stanford, Hongfang Liu at U Maryland, James Pustejovsky at Brandeis

    b. Also involved with these two databases is the need for a mapping to a standardized set of drug names; this would require a mix of good spelling correction algorithms, and a resource with "canonical" drug names and brand names and variants; also a good "abbreviation expansion" program.

    c. Capture of free text information for the National Plant Diagnostic Network; this would allow capture of retrospective records into the new standardized format; this is "ground truth" data available now, and it would be possible to use these resources perhaps to pose a "challenge evaluation" problem. To do this would require some funding to put together the data set, including training and test data, plus evaluation

software. The training set would include hand coded data as "ground truth". Test set data would be kept blind, and then released at some point to interested participants, who would build an automated system to encode the data. Ideally, there should be a long-term funding plan; for example, some govt agency might be willing to fund an evaluation, and then, as a follow on, fund several of the "high performing" groups to build and evaluate a running system.

2. Surveillance for disease outbreaks
    a. There is interest in detection of early warnings and indicators of disease through collection of global newswire and web-based information (sources like ProMED, EID, and local and national on-line news sources). Current systems (e.g., MITAP, http://mitap.sdsu.edu) collect online sources (mostly in English, soon in Chinese) and "bin" the news, for example, into reports related to diseases (or reports by area or source). It would be possible to use the data already collected as the basis for experiments on early detection of international outbreaks (could we have picked up the SARS outbreak if we had been able to track Chinese local on-line news). This would involve elements of "topic detection and tracking", new topic detection research (followed in TREC and the TDT evaluations) as well as cross language information retrieval and document routing.

Working Subroup on Biostatistics, Molecular and Genetic Epidemiology

Ana Dutra (New Jersey Institute of NeuroScience)
Tomasz Imielinski (Rutgers University)
Mark Nicolich (Exxon Mobil Biomedical Sciences, Inc.)
Ingo Ruczinski (Johns Hopkins University)
Bill Shannon (Washington University)
Frank Wong (DIMACS)

This group focused on the problem of associating single nucleotide polymorphisms (SNPs) with observable traits, i.e. phenotypes (many of the topics discussed also apply to micro array data). We considered four specific questions within this area:

1) How can we identify SNPs and gene-gene interactions that are associated with phenotypes?

2) Can we improve association studies by stratification methods?

3) How can we help bridging the various disciplines concerned with association studies?

1) During the workshop, methods of how to assess variable importance and gene-gene interactions (Epistasis) were discussed by Bill Shannon (using combinatorial partitioning), Ingo Ruczinski (using Logic regression), and Tomasz Imielinski (using association rules). Many others have been used in the past, such as linear and logistic regression, CART and MARS. A collection of tools are listed for example in the annual GAW workshop proceedings (for references, see http://www.gaworkshop.org/pubs.htm). The problematic common to all approaches is the vast search space for the models under consideration. Often, the number of predictors even exceeds the number of observations.

2) Stratification looks for the association of SNP differences with phenotype differences controlling for other (genetic and possibly non-genetic) differences in the individuals. Given the dimensionality of the data this poses some non-trivial issues, such as correctly matching subjects. Particularly worthwhile are partitioning (discussed by Bill Shannon) and possibly propensity score adjustments. Other possible approaches include tree models (proposed by Chipman, George and McCulloch) or various methods introduced by Ross Quinlan (http://www.rulequest.com/) and some methods discussed by Ruczinski, Kooperberg, Leblanc (Logic regression, JCGS, 2003).

3) In these types of studies, a hypothesis about the association between variations in the genome and some phenotype is studied. A common question is how to allocate the given amount of resources ($$). Is it better to sequence many patients on a few sites in the

genome, or is it better to sequence few people on many sites ("needle in the haystack")? Is it best to focus on one or two pathways, or more? How do we sample the populations? How do we design those studies (case control, cohort, case-cohort, other?) and how do we analyze those studies (in particular, what software is available for a given study design)? To tackle those tasks, a close collaboration between all disciplines involved is crucial. But how can we help to bridge the various disciplines (such as Biology, Computer Science, Mathematics, Statistics, Chemistry) concerned with SNP association studies? How can we get new researchers in this field up to speed? Clearly lacking is a good text for the introduction into the biology behind SNPs and haplotypes, and an overview of the field such as "Structural Bioinformatics" by Bourne and Weissig for the field of Proteomics. Besides those tutorials, review papers (for example addressing the question when to use which method for haplotype reconstruction) would be extremely helpful.

Working Subgroup on Spatial and Non-Human Epidemiology

Tom Jacob (University of North Texas),
Kenton Morgan (University of Liverpool),
Bill Smith (USDA Forestry Service),
Carla Thomas (University of California, Davis),
Dan Wartenberg (Rutgers University)

The process

- Major concerns
- Detecting new diseases
- detecting changes in disease prevalence or distribution
- spatial data analysis

The Questions

- Non-denominator based disease reporting
  - How do we make meaningful inferences?
- Spatial data which involves polygons, lines, and points
  - how do we merge and mine them and estimate accuracy and uncertainty?
- Multiscalar environment
  - How do we work in with data at different time intervals and at individual animal or herd level?

Meaningful inferences from non-denominator based disease reporting?

- US and UK Organizations have these data
  - US: Historical databases 1970's
    Plant diagnostic database 2002
    Forest Health Monitoring Program 1990
    Natureserve database – Nature Conservancy
  - UK: VIDA - Veterinary Investigation Data Analysis

How do we merge and mine spatial data which involves polygons, lines, and points and estimate and represent accuracy and uncertainty?

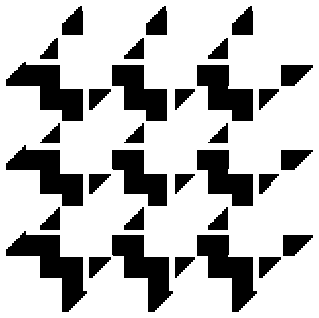
- Datasets
  - Polygonal datasets - common in government programs e.g digital elevation map
  - Linear/interpolated datasets - natural resources
  - Point data - permanent inventory plots
- How can it be merged without information loss?
- What visualization tools can be used to display uncertainty?

How do we work in a multiscalar environment i.e. with data collected at different time intervals or at different levels e.g. animal or herd level?

– data collected
- at different time intervals
- at different levels e.g animal or herd
- in different format e.g ordinal,
- How can the datasets be integrated?
- How can the outputs be displayed?

DIMACS

Center for Discrete Mathematics & Theoretical Computer Science
Founded as a National Science Foundation Science and Technology Center



DIMACS Working Group on Data Mining and Epidemiology

Second Meeting, March 18-19, 2004
DIMACS Center, CoRE Building, Rutgers University

Organizers:

James Abello, DIMACS, abello@dimacs.rutgers.edu
Graham Cormode, DIMACS, graham@dimacs.rutgers.edu
Kenton Morgan, University of Liverpool, k.l.morgan@liverpool.ac.uk
David Ozonoff, Boston University, dozonoff@bu.edu

Presented under the auspices of the Special Focus on Computational and Mathematical Epidemiology.

Workshop Program:

Thursday, March 18, 2004

8:00 - 8:50 Breakfast and Registration - CoRE Building - 4th floor

8:50 - 8:55 Welcome and Opening Remarks
Brenda Latka, DIMACS Associate Director

8:55 - 9:05 Opening Remarks:
James Abello, and Graham Cormode, DIMACS, Rutgers University

9:05 - 9:45 Biostatistical Challenges in Molecular Epidemiology
William Shannon, Washington University

9:45 - 10:00 Coffee Break-Discussion

10:00 - 10:50 Data Mining Overview
Dave Madigan, Rutgers University

10:50 - 11:00 Coffee Break - DIMACS Lounge

11:00 - 11:25 Perspectives on Automated Methods for Pharmacovigilance Signal Detection

A. Lawrence Gould and Peter K. Honig, Merck Research Laboratories

11:25 - 11:50 Pattern Analysis and Data Mining Efforts of the National Plant Diagnostic Network for Early Detection of Infectious Crop Disease and Pest Outbreaks
Carla Thomas, University of California
Leonard Coop, and Hang-Kwang Luh, Oregon State University

11:50 - 12:15 Observational Data Sets in Veterinary Epidemiology, The Challenges for New Data Mining Techniques
Kenton Morgan, University of Liverpool, UK

12:15 - 1:45 Lunch - DIMACS Lounge

1:45 - 2:10 Retooling Propensity ScoreTechniques with Machine Learning for Evaluating Solutions to the Los Angeles Drug Abuse Epidemic
Greg Ridgeway, The Rand Corporation

2:10 - 2:35 Finding Interactions and Assessing Variable Importance in SNP Association Studies
Ingo Ruczinski, Johns Hopkins University

2:35 - 3:00 Coffee Break-Discussion - DIMACS Lounge

3:00 - 3:25 Case-Control Surveillance Methods
Michael N. Cook, Merck Research Laboratories

3:25 - 3:50 The Generalized Contingency Table, its Concept Lattice and Connections with 2x2 Tables
Dave Ozonoff, Boston University and Alex Pogel, New Mexico State University

3:50 - 4:15 The Exploration of Spatial Data Mining (and Mind Mining) to model the risk of Emerald ash borer (EAB) (Agrilus planipennis) and its likely spread from current areas of infestation
Bill Smith, USDA Forest Service

4:15 - 4:30 Coffee Break/Working Groups - DIMACS Lounge

4:30 - 4:50 Epidemiological Factors of Survival Time for Cancer Patients Discovered via SVM Learning Classification Method: An Experimental Study on SEER Data
Ilya Muchnik and Jixin Li, DIMACS, Rutgers University

4:50 - 5:10 Graph Theoretical Methods in Epidemiology
James Abello, DIMACS- Rutgers University

5:10 - 6:30 Working Groups Discussions

6:30 - 8:30 Dinner - Reception - DIMACS Lounge

8:30 Van to the Hotel

Friday, March 19, 2004

8:00 - 8:50 Breakfast and Registration - CoRE Building - 4th floor

8:50 - 9:05 Introductory Remarks
James Abello, Graham Cormode, DIMACS, Rutgers University

9:05 - 10:00 Association Rule Mining of Biological Data Sets
Tomasz Imielinski, Rutgers University

10:00 - 10:20 Capture and Use of Free Text Information for Tracking Disease Outbreaks
Lynette Hirschman, The MITRE Corporation

10:20 - 11:45 Working Groups Discussions and Report Preparation

11:45 - 12:30 Descriptive Epidemiology for Data Miners
Dona Schneider, Rutgers University

12:30 - 1:45 Lunch - DIMACS Lounge

1:45 - 2:45 Working Groups Presentations

2:45 - 3:30 Discussion and Future Plans

3:30 - 3:45 Closing Remarks

4:30 Van to the Hotel


Previous: List of Participants


Workshop Index


DIMACS Homepage **Contacting the Center**

Document last modified on March 12, 2004.