Final Report on Seminar "Group Testing in the Life Sciences"

Dr. Alexander Schliep
Group Leader Algorithmics
Max Planck Institute for Molecular Genetics
Berlin, Germany

Prof. Amin Shokrollahi
Ecole Polytechnique Fédérale de Lausanne
Switzerland

Dr. Nicolas Thierry-Mieg
Laboratoire TIMC-IMAG / TIMB
Centre National de la Recherche Scientifique (CNRS)
Grenoble, France

Group testing AKA smart-pooling is a general strategy for minimizing the number of tests necessary for identifying positives among a large collection of items. It has the potential to efficiently identify and correct for experimental errors (false–positives and false–negatives). It can be used whenever tests can detect the presence of a positive in a group (or pool) of items, provided that positives are rare. Group testing has numerous applications in the life sciences, such as physical mapping, interactome mapping, drug–resistance screening, or designing DNA-microarrays, and many connections to computer science, mathematics and communications, from error-correcting codes to combinatorial design theory and to statistics. The seminar brought together researchers representing the different communities working on group testing and experimentalists from the life sciences. The desired outcome of the seminar was a better understanding of the requirements for and the possibilities of group testing in the life sciences.

Due to the strongly interdisciplinary nature of the seminar, introductory talks were first given on basic biological concepts and methods with a focus on possible group testing applications (T. Beißbarth), as well as on recent developments in coding theory (S. Litsyn). More specialized talks, often reporting unpublished results, then addressed the seminar topic from the angles of the various fields represented among the participants. This ranged from presentations of new group testing algorithms on graphs to find defective edges (E. Triesch) or learn a hidden graph (H.-L. Fu), with a focus on protein-protein interaction networks (P. Damaschke), to experiment–driven talks describing applications of group testing in the agricultural sciences (F. Schaarschmidt), in the mapping of interactomes (N. Thierry-Mieg), or in microbe classification using

Dagstuhl Seminar Proceedings 08301
Group Testing in the Life Sciences
http://drops.dagstuhl.de/opus/volltexte/2008/1780

compressed sensing DNA microarrays (O. Milenkovic). New efficient algorithms for decoding the outcome of a group testing experiment were introduced (M. Jimbo), and solutions to the group testing with inhibitors problem were proposed (A. De Bonis). Finally, two talks analysed biological sequence and subsequence problems using fault tolerant interval group testing (F. Cicalese) or constrained Minkowski sums (F. Eisenbrand).

Though the number of participants was smaller than we expected, the meeting was successful in that it brought together researchers from various fields and managed to plant the seeds for cross-fertilization of ideas.

The following is a list of abstracts of the talks given in this seminar.

Biological Background and Applications of Group Testing in Biology

Tim Beißbarth
DKFZ - Heidelberg

In this talk basic biological concepts and methods like high-throughput screening with microarrays and RNAi were introduced. Beside the biological background, these techniques were evaluated in light of possible applications for group testing problems.

Learning Biological Networks via Nested Effects Models

Tim Beißbarth
DKFZ - Heidelberg

Nested Effects Models are a statistical modeling framework similar to bayesian networks. Nested Effects Models were designed to model nested effects structures in data in order to reconstruct networks - different from bayesian networks Nested Effects Models distinguish between a Signal Network and an Effects Graph. These models have been used to reconstruct biological signaling networks based on a combination of gene-expression and knock-down intervention data. The basic concepts and some biological examples were introduced in this talk.

2-Stage Fault Tolerant Interval Group Testing

Ferdinando Cicalese
University of Salerno

We study the following fault tolerant variant of the interval group testing model: Given four positive integers n, p, s, e, determine the minimum number of questions

needed to identify a (possibly empty) set $P \subseteq \{1, 2, \ldots, n\}$ ($|P| \leq p$), under the following constraints. Questions have the form "Is $I \cap P \neq \emptyset$?", where I can be any interval in $\{1, 2 \ldots, n\}$. Questions are to be organized in s batches of non-adaptive questions (stages), i.e, questions in a given batch can be formulated relying only on the information gathered with the answers to the questions in the previous batches. Up to e of the answers can be erroneous or lies.

The study of interval group testing is motivated by several applications. remarkably, by the problem of identifying splice sites in a genome. In particular, such application motivates to focus algorithms that are fault tolerant to some degree and organize the questions in few stages, i.e., on the cases when s is small, typically not larger than 2. To the best of our knowledge, we are the first to consider fault tolerant strategies for interval group testing.

We completely characterize the fully non-adaptive situation and provide tight bounds for the case of two batch strategies. Our upper and lower bound only differ by a factor of $\sqrt{11/10}$ for the case $p = 1$, and by a factor of $\sqrt{4/3}$ for the general case $p \geq 2$.

Learning Hidden Hubs and Vertex Covers by Neighborhood Queries and Group Tests

Peter Damaschke
Chalmers University

Determining and analyzing protein-protein interaction networks is a major step in the understanding of complex mechanisms in the living cell. We address the question of finding a big fraction of the pairwise interactions with minimal experimental effort, and in a small number of stages where experiments are done in parallel, that is, non-adaptively. Experiments with "bait" proteins return all "prey" proteins interacting with the bait.

Expressed in graph-theoretic language, we are given a graph with initially unknown edge set and may be able to ask the following types of queries: neighborhood queries (return all neighbors of a given vertex), neighborhood group tests (is there some edge between a given vertex and a given set of vertices?), and edge group tests (is there some edge in a given set of vertices?). Due to well-known results for group testing, $O(r \log n)$ adaptive or $O(r^2 \log n)$ nonadaptive neighborhood group tests can simulate a neighborhood query to a vertex of degree at most r. On the other hand, in protein-protein interaction networks most vertices have small degrees. This motivates the study of strategies for detecting most edges in an unknown graph by neighborhood queries, which may actually be realized by group tests.

The basic idea is to choose random bait vertices in a first stage. Then, vertices appearing several times as preys are likely to have high degrees and should be queried in further stages. We analyze 2- and 3-stage strategies and estimate the number of queries needed to detect, with high probability, all high-degree vertices, or a small ver-

tex cover or a small dominating set of vertices, both for general graphs and for graphs with scale-free degree distributions. A set of k vertices that dominates a given fraction of all vertices in the graph can be found using $O(k)$ neighborhood queries.

We also study the problem of learning hidden vertex covers of size k (that is, vertex covers of size k in a previously unknown graph of size n) by edge group tests. Using $(2,k)$-disjunct matrices we can learn them by $O(k^3 \log n)$ nonadaptive queries. We can decode the answers within the time bound of a parameterized algorithm for vertex cover enumeration, which is $O(b^k poly(n))$ for some constant base $b < 2$ and a polynomial $poly$.

Back to neighborhood group tests, since vertex degrees r are very different and strategies with few stages are preferable for unraveling networks, the problem of group testing with unknown r and minimum adaptivity arises. This is also of independent interest. By recent results of De Bonis, Gasieniec, Vaccaro and Eppstein, Goodrich, Hirschberg, $O(r \log n)$ queries in two stages are sufficient, but r must be known in advance. For the case of previously unknown r we show that any deterministic strategy with this optimal query number must use $\Theta(\log r / \log \log r)$ stages in the worst case. On the other hand, some randomized strategy solves the same problem in only three stages, with high probabality, where randomization is used only for estimating r.

(The talk provided yet unpublished, preliminary results from an ongoing project supported by VR, the Swedish Research Council, grant no. 2007-6437.)

Group testing with inhibitors

Annalisa De Bonis
University of Salerno

Group testing with inhibitors (GTI) is a variant of classical group testing where in addition to positive items and negative items, there is a third class of items called *inhibitors*. In this model the response to a test is YES if and only if the tested group of items contains at least one positive item and no inhibitor. This model of group testing has been introduced by Farach *et al.* for applications in the field of molecular biology. We have investigated the GTI problem both in the case when the exact number of positive items is given, and in the case when the number of positives is not given but we are provided with an upper bound on it. For the latter case, we present a lower bound on the number of tests required to determine the positive items in a completely nonadaptive fashion. Also under the same hypothesis, we show an improved lower bound on the number of tests required by *any* algorithm (using any number of stages) for the GTI problem.

As far as it concerns the case when the exact number of positives is known, we give an efficient two-stage algorithm. Instrumental to our results are new combinatorial structures introduced in this paper. In particular we introduce generalized versions of the well known superimposed codes and selectors that we believe to be also of independent interest.

Constrained Minkowski Sums

Friedrich Eisenbrand
EPFL

In this talk, we introduce the notion of a constrained Minkowski sum which for two finite point-sets $P, Q \subseteq \mathbb{R}^2$ and a set of k inequalities $Ax \geq b$ is defined as the point-set $(P \oplus Q)_{Ax \geq b} = \{x = p + q \mid p \in P,\, q \in Q,\, Ax \geq b\}$. We show that typical subsequence problems from computational biology can be solved by computing a set containing the vertices of the convex hull of an appropriately constrained Minkowski sum. We provide an algorithm for computing such a set with running time $O(N \log N)$, where $N = |P| + |Q|$ if k is fixed. For the special case $(P \oplus Q)_{x_1 \geq \beta}$, where P and Q consist of points with integer x_1-coordinates whose absolute values are bounded by $O(N)$, we even achieve a linear running time $O(N)$. We thereby obtain a linear running time for many subsequence problems from the literature and improve upon the best known running times for some of them. The main advantage of the presented approach is that it provides a general framework within which a broad variety of subsequence problems can be modeled and solved.

We also consider arbitrary subsets S of the Minkowski sum of P and Q and show that the convex hull of S has at most $O(|P|^{2/3}|Q|^{2/3} + |P| + |Q|)$ extreme points.

The talk is based on joint work with Thorsten Bernholt, Thomas Hofmeister, Janos Pach, Thomas Rothvoß and Nir Sopher.

Group Testing Algorithms for DNA Library Screening based on BP and CCCP

Masakazu Jimbo
Nagoya University

The study of gene functions requires high-quality DNA libraries. However, a large number of tests and screenings are necessary for compiling such librarires. We describe an algoritm for extracting as much information as possible from pooling experiments for library screening. Collections of clones are called pools, and a pooling experiment is a group test for detecting all positive clones. The probability of positiveness for each clone is estimated according tot he outcomes of the pooling experiments. Clones with high chance of positiveness are subjected to confirmatory testing. Here we describe a new positive clone detecting algorithm based on CCCP (Concave-Convex Procedure) together with BNPD (Bayesian Network Positive Decoder) which was introduced by Uehara and Jimbo (2008). The performance of CCCP adn BP are compared, but simulation, with that of the Markov chain pool result decoder (MCPD) proposed by Knill et al.

(1996). Moreover, the combinatorial properties of pooling designs suitable for the proposed algorithm are discuss in conjunction with combinatorial designs and d-disjunct matrices.

Learning a Hidden Graph with an Adaptive Algorithm

Hung-Lin Fu
Nat. Chiao-Tung University

We consider the problem of learning a *hidden graph* using *edge-detecting queries* in a model where the only allowed operation is to query whether a set of vertices induces an edge of the hidden graph or not. In this paper we present an *adaptive algorithm* that learns a general graph with n vertices and m edges using at most $(2\log n + 9)m$ queries.

A short survey of coding theory

Simon Litsyn
Tel Aviv University

In the talk we explain recent developments in coding theory. We mainly concentrate on the historical perspective, with an emphasis on comparison of algebraic methods to iterative schemes.

We explain as well different applications of coding theory, mainly in communications, storage, mathematics (lattices and groups) and biology.

Connections between codes and such combinatorial objects as schemes and designs are considered. It is shown how codes can be used to construct many known designs. Since these designs are related to codes, efficient decoding algorithms can be employed in settings employing the corresponding designs.

Compressive Sensing DNA Microarrays

Olgica Milenkovic
Univ. of Illinois - Urbana Champaign

Accurate identification of large numbers of biological agents in an environment is an important and challenging research problem. DNA microarrays are frequently applied solutions for microbe detection, classification and other biosensing applications. A DNA microarray is an array of genetic sensors containing DNA sequences termed *probes*. Each spot has a large number of copies of a single probe sequence.

From the perspective of a microarray, each DNA sequence can fundamentally be viewed as a sequence of four DNA bases $\{A, T, G, C\}$ that tend to bind with one another in complementary base pairs: A with T and G with C. A DNA strand in a target organism's genetic sample will tend to bind or "hybridize" with its complementary probe on a microarray to form a stable structure. This is the underlying principle behind all DNA microarrays.

A Compressive Sensing Microarray (CSM) is a new device for DNA-based identification of target organisms that leverages the nascent theory of Compressive Sensing (CS). In contrast to a conventional DNA microarray, in which each genetic sensing probe is designed to respond to a single target organism, in a CSM each probe responds to a group of targets. As a result, significantly fewer total sensor spots are required. We study how to design group identifier probes that simultaneously account for both the constraints from the CS theory and the biochemistry of probe-target DNA hybridization. We employ Belief Propagation as a CS recovery method to estimate target concentrations from the microarray intensities.

(This is joint work with with Wei Dai, Mona A. Sheikh, Richard G. Baraniuk.)

Group Testing in the Agricultural Sciences

Frank Schaarschmidt
Universität Hannover

Rare proportions such as the proportion of infected individuals in a population, the proportion of genetically modified seeds in conventional seed lots or the probability that a disease is transmitted by individuals of insect species are of interest in ecological or agricultural sciences. Group testing can be used to estimate the probability π of this rare event in a population or to decide upon hypotheses concerning π. The talk reviews statistical procedures to estimate confidence intervals for π based on group testing including their assumptions. The deficiencies of simple methods are illustrated. Further, different situations of experimental design are considered, with focus on choosing the size of groups (pools) when interest is in rejecting the null-hypotheses of statistical tests concerning π. Approaches to compare two or several treatments with respect to π are discussed briefly.

Smart-pooling: increasing accuracy, coverage and efficiency in high-throughput screening

Nicolas Thierry-Mieg
TIMC-IMAG, CNRS - Grenoble

Smart-pooling is an experimental methodology susceptible of increasing efficiency, accuracy and coverage in high-throughput screening projects. It consists in assaying

well-chosen pools of probes, such that each probe is present in several pools, hence tested several times. The goal is to construct the pools so that the positive probes can usually be directly identified from the pattern of positive pools, despite the occurrence of false positives and false negatives. While striving for this goal, two interesting mathematical or computational problems emerge: the pooling problem (how should the pools be designed?), and the decoding problem (how to interpret the outcomes?). In this talk I will discuss these questions and the solutions we have proposed. I will then present the results of validation experiments that we have performed in the context of Y2H interactome mapping. First, we rescreened a 100x940 subspace of the human CCSB-HI1 interactome, demonstrating high specificity and sensitivity compared to the standard CCSB protocol. Second, we performed extensive screening in high-density formats (384 and 1536) of the complete worm ORFeome (13000 preys) using 12 baits: 5 methods were applied, including full duplicated individual-prey arrays and 3 different smart-pooling schemes.

An algorithm for finding defective edges in a graph

Eberhard Triesch
RWTH Aachen

The main goal of the talk is to present a group testing algorithm for finding d defective edges in a graph $G = (V, E)$ when the allowed tests are as follows: For a subset $W \subset V$, we may test whether one of the defective edges has both endpoints in W or not. The algorithm is easy to implement and finds the edges by $d(\log(n/d) + 6)$ tests if d is known and by $d(\log(n/d) + 9)$ tests if d is unknown.