Mathematisches Forschungsinstitut Oberwolfach

T a g u n g s b e r i c h t 12/2000

Mathematische Stochastik

12. bis 18. März 2000

Organisers of the meeting were Peter Bickel (Berkeley), Simon Tavaré (Los Angeles) and Anton Wakolbinger (Frankfurt).

The meeting focussed on "Challenges for Probability and Statistics from Molecular Biology and Population Genetics". Each day started with talks by scientists with especially strong biological credentials. Among the topics addressed were

- pattern recognition in DNA and protein sequences
- the analysis of gene expression arrays
- models of sequence analysis and tree reconstruction
- genome rearrangement analysis
- evolution in spatial populations.

In an evening session the role of stochastics in bioinformatics programs was discussed.

In the tradition of universality of the Oberwolfach spring meeting, there were also sessions less closely connected with the biological theme. Some of the topics treated in these sessions were

- models from statistical physics
- stochastic processes in random environment
- branching particle systems and superprocesses
- Poisson approximation and Stein's method
- improving unstable predictors and classifiers by bagging and boosting.

The enterprise of bringing together a quite inhomogeneous body of experts worked out well. Most of the lecturers made a big effort to explain core ideas. The result was that even non-specialists could take part in the lively discussions that followed.

Abstracts

Nick Barton
Evolution in spatially continuous populations

Many questions in evolution and ecology involve stochastic reproduction in spatially continuous populations. For example, one can ask under what circumstances a stable population can be sustained; whether rates of movement and population density can be inferred from spatial patterns in gene frequency; and whether a new combination of genes can be established by random genetic drift. The simplest approach is to approximate gene flow by diffusion, and random drift by an uncorrelated "white noise"; this gives straightforward results which fit closely with simulations. However, there are fundamental difficulties in justifying such an approximation in a two-dimensional spatial continuum in which nearby individuals necessarily interact.

Peter Bickel
Recognition of highly conserved patterns in protein sequences

We exhibit a method for finding all subfamilies of a family of aligned protein sequences which have what we define as a strong motif of sites. We propose a rule for declaring validity of our findings and using an elementary strawman model calculate a bound on the rate of false decisions. We apply this method to the family of phycobiliproteins which play an important role in photosynthesis for cyanobacteria, red algae and cryptomonads. Of the 24 subfamilies found and declared to have a strong motif, 23 were found to consist of members having important functional and evolutionary commonalities.

(Joint with A. Glazer (MCB), P. Spector(Stat.), G. Wedemayer(Chem.), UCBerkeley)

Peter Bühlmann
Explaining Bagging and a new combination with Boosting

Bagging [Breiman, 1996] is one of the most effective computational intensive procedures to improve on unstable predictors or classifiers, useful especially for high dimensional data set problems. We derive theoretical results to explain the variance reduction effect of bagging in hard-decision problems, including stumps (decision trees with 2 terminal nodes) for continuous regression functions and classifiers. Hard-decisions create instability, and bagging is shown how it smoothes hard thresholds yielding smaller variance. For example, we obtain an asymptotic limiting distribution at the cube-root rate for the split point when fitting piecewise constant functions. Denoting by n the sample size, it follows that in a cylindric neighborhood of diameter $O(n^{-1/3})$ of the theoretically optimal split point, the variance reduction can be characterized analytically. Because of the slow rate and the fact that many split points are carried out in decision trees such as CART, the covariate space is filled with such neighborhoods and therefore our reasoning provides an explanation on the global scale. We illustrate bagging, and computationally cheaper versions, on some numerical examples. And we finally argue why bagging combined with boosting [Freund and Schapire, 1997] can become a powerful variance AND bias reduction technique.

(Joint work with Bin Yu, Bell Labs (Lucent Technologies) and University of California at Berkeley)

Peter Eichelsbacher
Moderate deviations for mean field Gibbs measures

We present a *moderate deviations principle* around non-degenerate attractors of the empirical measure of random variables distributed according to a *mean field Gibbs measure*. The result is applied for the Curie-Weiss model, which models a ferromagnet. This model is well known. It is ideal for doing exact calculations. Moreover the result is applied for diffusion processes with mean field interaction leading to a McKean-Vlasov limit.

The proof is based on a *rank-dependent* moderate deviations principle for a *collection* of U-empirical measures and U-statistics. We point out that *decoupling* and *randomization* offer an efficient way proving relatively sharp exponential estimates. Theses estimates are one of the main ingredients in our proofs.

(Joint work with Tim Zajic, Minnesota).

Alison Etheridge
Evolution in spatially continuous populations: some mathematical questions.

This 'partner' talk to that of Nick Barton raises a number of mathematical questions arising from the study of models of populations evolving in 2-dimensional continua. Although partially addressed by classical stepping stone/coalescent models, some of the assumptions of those models are violated in our context and there is concern that their predictions might therefore be misleading. Perhaps the most significant challenge is to produce a class of spatially continuous models in which local rules regulate global population size in such a way as to prevent 'clumping' and extinction.

Klaus Fleischmann
Catalytic, mutually catalytic, and cyclically catalytic branching

Measure-valued processes on R^d are presented which model the phenomenon of moving, branching, and locally interacting populations in three different types of settings: one of two populations acts autonomously, two populations catalyze each other, and $K \geq 3$ populations interact cyclically.

The survey is based on joint work with Don Dawson, Jean-François Delmas, Alison Etheridge, Achim Klenke, Leonid Mytnik, Ed Perkins, and Jie Xiong.

Nina Gantert
Random walks on Galton-Watson trees

We first choose a Galton-Watson tree according to a (supercritical) branching process. This tree defines a simple random walk which moves from a vertex with equal probabilities to one of its neighbours. In analogy to the "random walk in random environment" model, the random tree here takes the role of a random environment. We also consider "λ-biased" random walks which do not move with equal probabilities to all neighbours but have a bias towards the root. Let $|X_n|$ denote the distance of the walker at time n to the root of the tree. It is known that $|X_n|/n$ converges almost surely to a deterministic speed $v_\lambda > 0$. Our main result is a large deviation principle for the distributions of $|X_n|/n$. In particular, we show that the exponential decay of speedup and slowdown probabilities for a "typical" tree is the same as if we average over all trees. This is in sharp contrast with the results for random walk in random environment which partly motivated this work.

The talk is based on joint work with Amir Dembo, Yuval Peres and Ofer Zeitouni.

Jochen Geiger
Branching processes in random environment

A branching process in random environment is a Galton-Watson process in which particles have generation dependent offspring distributions picked at random. Generalizing a result of Kozlov in the special case of offspring distributions with linear fractional generating functions we show that, as $n \to \infty$, the non-extinction probability of a branching process in general i.i.d. random environment is asymptotically $\beta n^{-\frac{1}{2}}$ for some $0 < \beta < \infty$. A key ingredient for the proof is a representation of the non-extinction probability in terms of a random walk which is derived from a spinal decomposition of the family tree produced by the branching process. The analysis of this formula shows that only extremely favorable environments in which the population has a reasonable chance to survive contribute to the unlikely event of non-extinction.

(Joint work with G. Kersting, Frankfurt)

Hans-Otto Georgii
Phase transition and percolation in the two-dimensional Ising model

The occurrence of phase transitions in lattice models of Statistical Mechanics is often intimately related to the the presence of infinite clusters in certain associated random graphs. This holds in particular for the the Ising model on the square lattice $\mathbb{Z}^2$: in the case of a unique Gibbs measure all clusters of +spins are finite almost surely, but if several extremal Gibbs measures (phases) exist, the +phase almost surely shows an infinite +cluster. This relationship between phase transition and percolation can be pushed further to obtain information on the number of distinct phases: there exist no more than two different phases if and only if the coexistence of infinite + and −clusters is impossible almost surely for each Gibbs measure. A result of Russo–Aizenman–Higuchi (1980) states that the latter is indeed the case. We present a new streamlined proof of this result.

(Joint work with Y. Higuchi, Kobe University.)

Arndt von Haeseler
Models of sequence evolution and tree reconstruction

We give a brief overview about currently used methods to reconstruct phylogenetic trees. Prominent methods are maximum parsimony, distance based methods like neighbor-joining and maximum likelihood approaches. Special emphasis is put on the latter approach, because this provides a statistical framework to actually test our assumptions about the evolutionary process. We show, using mitochondrial DNA sequences as an example, that more complex models, e.g., models that include a transition-transversion parameter or models that include dependencies among sites, usually describe the evolutionary process more appropriate than simple models.

However, comparing log-likelihoods to the so-called unconstrained log-likelihood reveals that we are still some distance away from a satisfactory description of the process that created the diversity in DNA sequences.

Part of this work is joint with Michael Schöniger and Korbinian Strimmer.

Susan Holmes
Confidence statements and averages for trees

This work presents two different ways of coding semi-labeled binary trees such as phylogenetic trees, hierarchical clustering trees or classification and decision trees. The aim is to provide answers to questions in biology arising from comparisons of gene trees for the same set of species that seem to conflict and from confidence regions for trees. The first encoding is the matching representation which constructs a bijection between the set of all binary trees with n labeled leaves and the set of matchings of $2(n-1)$ labels into pairs. Although this provides a useful coding for computational purposes, it implies a distance that from dissymmetry between changes including ancestors and those only including leaves. In joint work with Billera and Vogtmann, a more satisfactory approach has been provided by a geometric construction as the cube complex of cubes of dimension $n-2$, glued together by faces corresponding to unresolved trees. This space is seen to be CAT(0) using the simplicial link and thus the existence of geodesics can be proved. Centroids are constructed inductively from the half points on geodesics. This provides a good basis for constructing a Mallow's model on the space of trees.

(Joint work with Louis Billera, Persi Diaconis and Karen Vogtmann.)

Achim Klenke
Biodiversity of catalytic super-Brownian motion

We investigate the structure of the equilibrium state of three–dimensional catalytic super-Brownian motion where the catalyst is itself a classical super-Brownian motion. We show that the reactant has an infinite local biodiversity or genetic abundance. This contrasts the finite local biodiversity of the equilibrium of classical super-Brownian motion.

As a key tool for the problem we show that in $d = 3$ the reactant matter propagates everywhere in space immediately.

(Joint work with Klaus Fleischmann)

Timo Koski
A model for predictive mixtures and for classification

Detection of protein sequence homologies can be done by using mixtures of Dirichlet distributions. These are statistical models for motifs in multiple alignments of protein sequences. We derive this mixture using an assumption of infinite exchangeability and predictive sufficiency. By this argument it is immediate that we are dealing with predictive classification of protein sequences in the sense of predicting a portion of a sequence based on a motif. Finally a result about the distribution of the score based on an exchangeable representation is outlined.

Eva Löcherbach
Asymptotic statistics for Markovian particle systems with killing and jumps

We consider finite systems of particles built from pasting together pieces of strong Markov processes which are killed at position dependent rates and connected via transition kernels. Such systems can e.g. be models for finite systems of branching diffusions. We derive an explicit formula for the likelihood ratio process of such particle systems on a suitable path space in situations where the motion of particles, the killing rate and the jump

kernel depend on some unknown finite dimensional parameter ϑ. For ergodic submodels, under smoothness assumptions on the parametrization, we obtain a limit theorem for the likelihoods when comparing parameters $\vartheta + \frac{1}{\sqrt{n}}h$ with a fixed parameter ϑ, as the observation time n tends to infinity. In terms of the theory of asymptotic statistics due to Le Cam, this means that local asymptotic normality (LAN) holds at ϑ. As a consequence, we can characterize asymptotically efficient estimators for the unknown parameter. In null-recurrent models, similar results are obtained.

Matthias Löwe
Reconstruction of random scenery

A d-dimensional (random) m-color scenery ξ is a (random) coloring of the d-dimensional integer lattice $\mathbb{Z}^d$ by $m(\geq 2)$ colors:

$$\xi : \mathbb{Z}^d \longrightarrow \{0, \ldots, m-1\}$$

Two sceneries are equivalent, if they just differ by shifts and rotations of $\mathbb{Z}^d$.
We assume that ξ is not directly observable. Instead, let $(S_n)_{n\geq 0}$ denote simple random walk on $\mathbb{Z}^d$ starting in the origin.
The scenery problem in general asks for non-trivial conclusions we can deduce from the color record

$$\chi := (\xi(S_n))_{n\geq 0}$$

of the random walk to the scenery it has been produced on. Of course, the most ambituous question is whether ξ can be reproduced (up to the equivalence '$\sim$') from χ.
In this talk we review results on the scenery problem in the past decade. We mainly focus on the following theorem, which gives a first result that also 2-dimensional sceneries may be reconstructed from the color record of simple random walk.

Theorem Let ξ be a 2-dim. i.i.d. scenery, i.e. $\xi(i), i \in \mathbb{Z}^2$ are i.i.d. random variables with

$$\mathbb{P}(\xi(i) = k) = \frac{1}{m}$$

$\forall i \in \mathbb{Z}^2, \forall k \in \{0, \ldots, m-1\}$.

Then there exists a number M_0 , such that whenever $m \geq M_0$, there exists a measurable mapping

$$\mathcal{A} : \{0, \ldots, m-1\}^{\mathbb{N}} \mapsto \{0, \ldots, m-1\}^{\mathbb{Z}^2}$$

for the set of observations to the set of 2-dim sceneries with

$$\mathbb{P}(\mathcal{A}(\chi) \sim \xi) = 1$$

(i.e. almost every scenery can be reconstructed almost surely up to shifts and rotation.)

This answers a question asked by Harry Kesten.

(Based on joint work with Heinrich Matzinger)

Dirk Metzler
Poisson approximations for genetic fingerprints

DNA fingerprinting techniques are used in various fields of biology to estimate the phylogeny of a number of individuals. Some of these techniques, for example RAPD-PCR, are based on the occurrence of certain patterns on the DNA sequence. Many difficulties in the analysis of DNA fingerprinting data arise from the fact that pattern occurrences on neighboured positions of DNA sequences are stochastically dependent due to possible pattern overlapping effects. The problem of quantifying these dependencies can be formulated in terms of a Poisson process approximation for a Poisson cluster process. I show how a variation of the Chen-Stein method can be used to find an upper bound for the total variation distance of the joint distribution of the fingerprint configuration of some individuals as it is induced by the Jukes Cantor model for DNA sequence evolution, and the distribution of the fingerprint configuration that is induced by a model in which pattern overlapping effects are neglected. The relationship of the individuals, which is assumed to be tree-like, is taken into account in the upper bound for the total variation distance.

Hans-Werner Mewes
Bioinformatics of genomes: a practical approach

The analysis of genomic DNA is the essence of modern systematic molecular biology. The analysis and interpretation of large amounts of data have to be supported by mathematical methods. Bioinformatics has to link factual data (e.g. sequences, structures) to attributes provided by experimental biology, e.g. protein function. Here, attributes are frequently associated to sequence information, but may also be entirely context dependent (e.g. protein A binds protein B only in presence of C.) Statistical methods allow for the the estimation of confindence for the prediction of attributes, mostly based on sequence/structure similarity of proteins. After the completion of the sequence of about 25 prekaryotic and 2 eukaryotic genomes, research in molecular biology focuses on function of the genetic elements (functional genomics). At present, most work tries to elucidate function of the individual genes, in the future cellular networks like metabolic and regulatory pathways will challenge prediction methods.

Martin Möhle
A classification of ancestral trees (coalescent processes) for exchangeable population models

A class of population models with non-overlapping generations and fixed population size N is considered. It is assumed that the family sizes within a generation are exchangeable random variables. A weak convergence criterion is established for a properly scaled ancestral process as N tends to infinity. It results in a full classification of the coalescent generators for the case of exchangeable reproduction. In general the coalescent process allows for simultaneous multiple mergers of ancestral lines. Kingman's coalescent appears if and only if triple mergers of ancestral lines are asymptotically negligible in comparison with binary mergers.

Peter Mörters
How fast are the particles of super-Brownian motion?

We study super-Brownian motion in the historical setting. In this setting each particle of super-Brownian motion alive at time t is represented by a path $w : [0,t] \to \mathbb{R}^d$ and the state of historical super-Brownian motion is a measure on the set of paths. Typical paths are Brownian motion paths, however in the uncountable collection of paths in the range of a super-Brownian motion there are some which at exceptional times move faster than Brownian motion. We show that the maximal speed of all paths during a given time period E is given by

$$\sup_{s\in E}\sup_{w\in\text{range}} \limsup_{h\downarrow 0} \frac{|w(s)-w(s-h)|}{\sqrt{h\log(1/h)}} = \sqrt{2+2\dim_P(E)}\,,$$

where $\dim_P$ is the packing dimension of E. This complements earlier results of Dawson/Perkins and Verzani. We also show that the Hausdorff dimension spectrum of fast paths in the range of historical super-Brownian motion of dimension at least 2 is given by

$$\dim\left\{w\in\text{range} : \limsup_{h\downarrow 0} \frac{|w(\zeta)-w(\zeta-h)|}{\sqrt{h\log(1/h)}} \geq a\right\} = 4-a^2\,,$$

where ζ denotes the lifetime of the stopped path w. The main tools of the proof are Le Gall's Brownian snake, the limsup random fractals of Khoshnevisan, Peres and Xiao and an extension of a uniform dimension formula of Serlet.
Parts of this research were carried out while the author was post-doctoral research fellow at the DFG-Graduiertenkolleg "Stochastische Prozesse und probabilistische Analysis" in Berlin.

Tobias Müller
A new method for modeling protein evolution

The estimation of amino acid replacement frequencies during molecular evolution is crucial for many applications in sequence analysis. Score matrices for database search programs or phylogenetic analysis rely on such models of protein evolution. Pioneering work was done by M. Dayhoff *et al.* (Atlas of Protein Sequences and Structure, 1978, 5, 345 – 352), who formulated a Markov model of evolution and derived the famous *PAM* score matrices. Her estimation procedure for amino acid exchange frequencies is restricted to pairs of proteins that have a constant and small degree of divergence. Here we present an improved estimator, called the resolvent method, that is not subject to these limitations. This extension of Dayhoff's approach enables us to estimate an amino acid substitution model from alignments of varying degree of divergence. Extensive simulations show the capability of the new estimator to recover accurately the exchange frequencies among amino acids. Based on the SYSTERS database of aligned protein families (Krause & Vingron, Bioinformatics, 1998, 14(5), 430 – 438) we recompute a series of score matrices.

(Joint with Martin Vingron)

Olle Nerman
Everything at once: Some reflections and experiences from an effort to build bioinformatics at all levels, jointly with biologists

Chalmers University of Technology and Gothenburg University cooperate in a recently started joint bioinformatics programme which I and Anders Blomberg, a yeast proteomics microbiologist at Gothenburg University, are jointly responsible for. The ambition is to build a broad bioinformatics scientific programme as fast as possible, on the undergraduate, the graduate and senior research levels. My talk was a short decription of what we have done, what we plan to do, and our experiences so far.

Anne-Mette Krabbe Pedersen
Probabilistic models of DNA sequence evolution with context dependent rates of substitution

We consider Markov processes of DNA sequence evolution in which the instantaneous rates of substitution at a site are allowed to depend upon the states at the sites in a neighbourhood of the site at the instant of the substitution. We characterize the class of Markov process models of DNA sequence evolution for which the stationary distribution is a Gibbs measure, and give a procedure for calculating the normalizing constant of the measure. We develop an MCMC method for estimating the transition probability between sequences under models of this type. Finally, we analyze an alignment of two HIV–1 gene sequences using the developed theory and methodology.

(Joint work with Jens Ledet Jensen)

Gesine Reinert
Stein's method and application to sequence analysis

Often biological sequences are modelled as a stationary m-order Markov chain on a finite alphabet. Typical questions include identifying significantly rare or significantly frequent short substrings (*words*) in the sequence. Typical heuristics concern normal, Poisson, and compound Poisson approximations. A common feature of these problems is that the observed sequence is often too long for exact results, but too short to be in the asymptotic regime. An additional complication is the dependence between different occurrences of words, due to self-overlap and overlap between words. Stein's method not only enables us to disentangle this dependence, but also to derive bounds on the approximation to the target distribution (normal, Poisson, or compound Poisson in the examples considered). These bounds can be used to give conservative confidence intervals.

(Joint work with S. Schbath and M. S. Waterman)

Hugh Salamon
On the identification of deletion polymorphism in non-repetitive DNA from genome-wide probe hybridization data

Genomic diversity within and between populations is due to single nucleotide mutations, changes in repetitive DNA systems, recombination mechanisms, and also due to insertion and deletion events. The contribution of these sources to diversity, whether purely genetic or of phenotypic consequence, can only be investigated if we have the means to

quantitate and characterize diversity in many samples. With the advent of complete sequence characterization of representative genomes of different species, the possibility of developing protocols to screen for genetic polymorphism across entire genomes is actively being pursued. The large numbers of measurements such approaches yield demand that we pay careful attention to the numerical analysis of data. Because individual probe-pair hybridization intensities exhibit limited sensitivity/specificity characteristics to detect deletions, data-analytical methodology to exploit measurements from multiple probes in tandem locations across the genome was developed. The Tandem Set Terminal Extreme Probability (TSTEP) algorithm designed specifically to analyze the tandem measurements data was applied to data from an Affymetrix GeneChip (TM) targeted at the fully sequenced genome of Mycobacterium tuberculosis. The algorithm was shown to discover relative deletions with high sensitivity in two related organism genomes. The TSTEP algorithm provides a foundation for similar efforts to characterize deletions in large numbers of hybridization measures in similar-sized and larger genomes.

David Sankoff
Duplication, rearrangement and reconciliation

Given N genomes in the form of gene orders, allowing multiple (approximate) copies of each gene within a genome, as well as a fixed species tree and one gene tree for each gene family, we want to infer gene orders of the ancestral genomes, including possibly multiple copies of each gene. To solve this, we make use of

1. the notion of exemplar genomes, consisting of one member of each gene family per genome, to generalize notions of (pairwise) genome distances

2. a reduction of the breakpoint median problem to the Traveling Salesman Problem in order to implement a steinerization strategy to find ancestral gene orders

3. reconciliation techniques to project a gene tree onto a species tree which provides information as to the gene content of each ancestral node, as well as the ancestry of each gene copy.

The solution starts with the reconcilation step followed by iterative application of exemplar and median algorithms across the species tree.

Gary Stormo
Interesting statistical problems in molecular biology

I gave two talks at the meeting. The first was an overview of basic molecular biology that described the properties of DNA, RNA and proteins, and the information flow between them. I described briefly the current worldwide efforts to sequence the complete genomes of many different species, including humans which will be completed later this year. I then described some challenging statistical problems in the analysis of biological sequences. These include: identifying the protein-coding regions from the raw genomic DNA; predicting the functions of the proteins based on homology to known proteins; predicting the structure of proteins; predicting the structure of RNA using comparative information; predicting regulatory networks, i.e. the collection of genes controlled by each regulatory protein; inferring the phylogenic relationships between genes and species. Other challenging statistical problems on biological sequences were described by other speakers.

The second talk was about our own research at discovering binding sites on DNA sequences for proteins involved in gene regulation. This is a pattern recognition or pattern discovery problem. One starts with a collection of genes that are known to be regulated by a particular protein, but one doesn't know where the binding sites for the protein are, only that they should occur somewhere (perhaps within 1000 bases) "upstream" of the regulated genes. One uses a model of how proteins bind to DNA, i.e. how their specificity is deteremined. In our cases we usually start with a simplified assumption of additive interactions between the amino acids of the protein and the bases of the binding sites, but more complicated models may be employed if necessary. Then we model the entire pattern discovery problem as one of maximizing the probability of each gene having a high affinity binding site for the protein, using the Boltzmann distribution for the probability function. We show that under some simplifying assumptions this is equivalent to identifying the set of binding sites with maximum relative entropy compared to the background genomic DNA. I described a couple of algorithms to search for that maximum, and how one could relax some of the assumptions normally used. In particular, one can calculate the actual partition function if the whole genome sequence is known, and eliminate the approximation of a random background. I showed some examples of the use of this method to discover binding sites for regulatory proteins in eukaryotic cells.

Simon Tavaré
Can we seen the forest for the trees?

Trees arise in many examples, and on a variety of time scales, in molecular biology. The classic example involves 'the' phylogenetic tree of a series of species inferred from molecular data. In this setting the tree is thought of as a parameter, to be estimated from the sequence data. In the population genetics setting, trees arise in a description of the ancestry of a sample of molecules. In this case the tree is random; a different sample produces a different tree. A tumor may be represented as the result of a cell duplication process, the cells sampled in the final clonal expansion of the tumor being related by the tree-like ancestry of a cell lineage. In all these cases, molecular variation observed in a sample is modeled by superimposing a mutation process on the tree. I will describe several such examples, focusing on the common ingredient of computational inference for models with dependence generated by a tree (or graph). A wide variety of methods, including importance sampling and Markov chain Monte Carlo, have been used to study such issues. I will conclude by describing in some detail a computational approach to resolving the difference between fossil record and molecular evolution estimates of species divergence times. This is a novel inference problem for branching processes.

Martin Vingron
Analysis of large scale gene expression data

During the last few years it has become possible to measure the amount of DNA for many genes in a cell simultaneously. Mostly, this is done using arrayed DNA spots, so-called DNA-chips. The result of an experiment is a vector of intensities for several thousand genes. The talk is discussing the various computational and mathematical problems arising in the design of the arrays, in the comparison of different experiments, the analysis of scanned images, and the final interpretation of the data. Correspondence analysis is suggested as one way of visualizing sets of experiments.

Gunter Weiss
On the number of mitochondria inherited in dogs

A typical eukaryotic cell contains about 1000 mitochondria which carry their own genome (mtDNA) in about ten copies per mitochondrion. These cell organelles replicate independently of the cell cycle. When a cell divides the mtDNA population of the cell is randomly distributed between the two daughter cells. At this stage genetic drift (or sampling error) can act as an evolutionary force. If the mtDNA population consists of more then one type, the daughter cells may differ in their frequency spectra of mtDNA types. This effect seems to be enhanced in the germ line, such that differences between different offspring become visible in studies of mtDNA. This enhancement may be caused by incomplete sampling of the mtDNA population during oogenesis, i.e. the hypothesis of a bottleneck in mitochondrial inheritance.

In order to test this hypothesis we estimated the number of inherited segregational units of mtDNA by using tandem repeat data from a study of domesticated dogs. We modelled the sampling error as a function of the number of inherited units and added the effect of possible measurement error in the data. Under the assumption that the measurement error is comparatively small (which has yet to be confirmed by experiment) our estimation procedure yield a bottleneck significantly smaller than 100 together with a most probable value of about 25 segregational units.

If this result holds true for human biology, it will have a major impact on the understanding of diseases caused by mtDNA deficiency.

Data were kindly provided by Peter Savolainen.

Berichterstatter: Dirk Metzler

Participants

Dr. Ellen Baake
baake@zi.biologie.uni-muenchen.de
Zoologisches Institut
Universität München
Luisenstr. 14
80333 München

Prof. Dr. Nicholas Barton
n.barton@ed.ac.uk
Ashworth Laboratories
Inst. of Cell, Animal and
Population Biology
West Main Road
Edinburgh EH9 3JT
SCOTLAND

Prof. Dr. Peter J. Bickel
bickel@stat.berkeley.edu
Department of Statistics
University of California
367 Evans Hall
Berkeley , CA 94720-3860
USA

Matthias Birkner
birkner@math.uni-frankfurt.de
Fachbereich Mathematik
Universität Frankfurt
60054 Frankfurt

Dr. Tom Britton
tom.britton@math.uu.se
Department of Mathematics
University of Uppsala
P.O. Box 480
S-75106 Uppsala

Dr. Peter Bühlmann
buhlmann@stat.math.ethz.ch
Seminar f. Statistik
ETH-Zentrum
LEO D12
CH-8092 Zürich

Dr. Peter Eichelsbacher
peter@mathematik.uni-bielefeld.de
Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum

Dr. Alison M. Etheridge
etheridg@stats.ox.ac.uk
Department of Statistics
University of Oxford
1 South Parks Road
GB-Oxford OX1 3TG

Dr. Klaus Fleischmann
fleischmann@wias-berlin.de
Weierstraß-Institut für
Angewandte Analysis und Stochastik
im Forschungsverbund Berlin e.V.
Mohrenstr. 39
10117 Berlin

Dr. Nina Gantert
gantert@math.tu-berlin.de
Fachbereich Mathematik
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin

Dr. Jochen Geiger
geiger@mi.informatik.uni-frankfurt.de
Fachbereich Mathematik
Universität Frankfurt
Postfach 111932
60054 Frankfurt

Prof. Dr. Hans-Otto Georgii
georgii@rz.mathematik.uni-muenchen.de
Mathematisches Institut
Universität München
Theresienstr. 39
80333 München

Prof. Dr. Friedrich Götze
goetze@mathematik.uni-bielefeld.de
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld

Steffen Grossmann
grossman@math.uni-frankfurt.de
Fachbereich Mathematik
Universität Frankfurt
60054 Frankfurt

Dr. Arndt von Haeseler
arndt@eva.mpg.de
Max-Planck-Institut für
evolutionäre Anthropologie
Inselstr. 22
04103 Leipzig

Prof. Dr. Susan Holmes
Department of Statistics
Stanford University
Sequoia Hall
Stanford , CA 94305-4065
USA

Prof. Dr. Reinhard Höpfner
hoepfner@mathematik.uni-mainz.de
Institut für Informatik
Universität Mainz
Staudingerweg 9
55122 Mainz

Prof. Dr. Götz Kersting
kersting@math.uni-frankfurt.de
Fachbereich Mathematik
Universität Frankfurt
Postfach 111932
60054 Frankfurt

Dr. Achim Klenke
klenke@mi.uni-erlangen.de
Mathematisches Institut
Universität Erlangen
Bismarckstr. 1 1/2
91054 Erlangen

Dr. Timo Koski
timo@math.kth.se
Dept. of Mathematics
Royal Institute of Technology
Lindstedtsvägen 25
S-100 44 Stockholm

Prof. Dr. Christof Külske
kuelske@wias-berlin.de
Weierstraß-Institut für
Angewandte Analysis und Stochastik
im Forschungsverbund Berlin e.V.
Mohrenstr. 39
10117 Berlin

Dr. Andreas E. Kyprianou
kyprianou@math.uu.nl
Mathematisch Instituut
Rijksuniversiteit te Utrecht
P. O. Box 80.010
NL-3508 TA Utrecht

Eva Löcherbach
locherba@ccr.jussieu.fr
Laboratoire de Probabilites
Universite Paris 6
4, Place Jussieu
F-75252 Paris

Matthias Löwe
loewe@sci.kun.nl
Mathematisch Instituut
Katholieke Universiteit Nijmegen
Toernooiveld 1
NL-6525 ED Nijmegen

Dr. Dirk Metzler
dmetzler@ath.uni-frankfurt.de
Max-Planck-Institut für
evolutionäre Anthropologie
Inselstr. 22
04103 Leipzig

Prof. Dr. Hans-Werner Mewes
Max-Planck-Institut für Biochemie
Am Klopferspitz 18a
82152 Martinsried

Dr. Martin Möhle
moehle@mathematik.uni-mainz.de
Fachbereich Mathematik
Universität Mainz
55099 Mainz

Peter Mörters
peter@mathematik.uni-kl.de
Fachbereich Mathematik
Universität Kaiserslautern
Erwin-Schrödinger-Straße
67663 Kaiserslautern

Tobias Mueller
t.mueller@dkfz-heidelberg.de
Deutsches Krebsforschungszentrum
Theoret. Bioinformatik (Abt. H0300)
Im Neuenheimer Feld
69120 Heidelberg

Prof. Dr. Olle Nerman
nerman@math.chalmers.se
School of Math. and Computer
Science
Chalmers University of Technology
and Gothenberg University
S-41296 Goteborg

Prof. Dr. Anne-Mette K. Pedersen
annemet@mudpop.bio.au.dk
Department of Theoretical Statist.
Institute of Mathematics
University of Aarhus
Ny Munkegade
DK-8000 C Aarhus

Gesine Reinert
g.reinert@statslab.cam.ac.uk
Kings College
Research Center
Kings Parade
GB-Cambridge CB2 1ST

Prof. Dr. Hugh Salamon
Berlex Laboratories, Inc.
15049 San Pablo Avenue
Richmond , CA 94804-0099
USA

Prof. Dr. David Sankoff
sankoff@ere.umontreal.ca
Dept. of Mathematics and Statistics
University of Montreal
C. P. 6128, Succ. Centreville
Montreal , P. Q. H3C 3J7
CANADA

Prof. Dr. Gary Stormo
stormo@genetics.wustl.edu
Department of Genetics
Washington Univ. Medical School
660 S. Euclid, Box 8232
St. Louis , MO 63110
USA

Jan M. Swart
janswart@sci.kun.nl
Fachbereich Mathematik
TU Berlin
Straße des 17. Juni 137
10623 Berlin

Prof. Dr. Simon Tavare
Dept. of Mathematics, DRB 155
University of Southern California
1042 W 36 Place
Los Angeles , CA 90089-1113
USA

Prof. Dr. Martin Vingron
m.vingron@dkfz-heidelberg.de
Deutsches Krebsforschungszentrum
Theoret. Bioinformatik (Abt. H0300)
Im Neuenheimer Feld
69120 Heidelberg

Prof. Dr. Anton Wakolbinger
wakolbin@math.uni-frankfurt.de
Fachbereich Mathematik
Universität Frankfurt
Robert-Mayer-Str. 6-10
60325 Frankfurt

Prof. Dr. Michael Waterman
Dept. of Mathematics, DRB 155
University of Southern California
1042 W 36 Place
Los Angeles , CA 90089-1113
USA

Dr. Gunter Weiss
weiss@eva.mpg.de
Max-Planck-Institut für
evolutionäre Anthropologie
Inselstr. 22
04103 Leipzig

Prof. Dr. Jian Zhang
zhang@eurandom.tue.nl
EURANDOM
Technical University Eindhoven,R.C.
P.O. Box 513
Antwoordnummer 513
NL-5600 VB Eindhoven

Prof. Dr. Willem R. van Zwet
vanzwet@math.leidenuniv.nl
Mathematisch Instituut
Rijksuniversiteit Leiden
Postbus 9512
NL-2300 RA Leiden