Mathematisches Forschungsinstitut Oberwolfach

Report No. 47/2001

Theoretische und Mathematische Biologie

October 21st – October 27th, 2001

The present conference was organized by Andreas Dress (Bielefeld) and William Martin (Düsseldorf). The focus of the meeting was how to advance the analysis and theoretical understanding of evolutionary processes using formal tools from mathematics – with some emphasis on molecular evolution.

In contrast to standard meetings, the focus was on interdisciplinarity and discussion. By bringing mathematicians and biologists together, important issues that had not yet been resolved (or even defined) were identified and analysed, rather than focussing on recent advances or harping on things that are currently thought to be true.

Beside the morning and afternoon sessions, there were evening discussions dealing with particular topics like "The vices and virtues of maximum likelihood", "Quartet methods", or "What can we learn from whole genome research".

Abstracts

Some aspects of recombination

Michael Baake

Recombination of sequences is one possible (and important) mechanism needed contributing to the genetic variability within a population. In the first part of this talk, a simple recombination model (in continuous time) is presented that admits a solution in closed form, and an extension towards including simple mutation/selection mechanisms. In the second part, first results are shown on the equilibrium states of a discrete time recombination model involving sequences of unequal length.

An agglomerative algorithm for network construction

David Bryant

In many situations, evolutionary history and data are best represented by a network, rather than a simple tree. I describe a simple agglomerative method for constructing networks from distance data – the method generalizes the popular neighbour-joining algorithm. A system of cyclic splits is constructed with the property that a circular metric (rather than a tree metric) will correspond to the resulting system of cyclic splits.

A space of trees

David B. A. Epstein
(joint work with Jonathan Ingram)

The space of trees defined by Billera, Holmes, and Vogtmann was discussed. Algorithms were discussed for finding (a) the distance between two trees and (b) the average of a finite number of trees. Possible applications were discussed:
(i) For a given collection of taxa, find genes common to each. Build a tree for each gene. Investigate the distribution of trees to determine which genes are recombinant.
(ii) Determine how sharply a maximum likelihood tree is defined by looking at suboptimal trees.
(iii) Any method using consensus can be replaced by averaging in the space of trees.

Another space of trees

Andreas Dress

(Phylogenetic) X-trees are uniquely determined by (i) the metric, (ii) the split system, or (iii) the quartet trees they induce on X. Analysing these three data structures and their relationships, it can be shown that various optimization problems related to phylogenetic tree construction can be interpreted as geometric problems relating to the convex polytope spanned by the set of points representing X-trees in terms of these data structures (considered as canonical embeddings of the set of X-trees into some large linear space).

Heterogeneous likelihood models in phylogenetics

Peter Foster

The process of evolution can sometimes differ over the tree, and it makes sense to accommodate that with heterogeneous models allowing the characteristics of the model to differ over the tree as well. An example data set of bacterial 16S RNA genes which show convergent compositional bias and attraction is used to show that a good fit of the model to the data can be obtained with few extra parameters.

A method for interpreting alignment data

Alex Grossmann

If the two sequences in a pairwise alignment are "sufficiently close" (in a sense which can be made precise), then one can consider two "rate matrices" associated to that alignment, and defined as the logarithm of the observed Markov matrices. A multiple alignment then gives rise to a family of matrices, i. e. (in case of a 20 letter alphabet) to a cloud of points in 400-dimensional space. The reasons for considering this cloud are:
(i) For sequences obtained from a reversible continuous Markov evolution, the points of the cloud are on a straight line through the origin.
(ii) The log-det distance between two sequences is the trace of the matrix in the cloud, and can be decomposed into contributions corresponding to mutations of individual amino acids. This is an extension of the natural "counting of mutations".

This method is being applied to the study of mitochondrial genomes of multicellular animals, in a work by C. Devauchelle, M. Monnerot, and Alex Grossman.

A neighbour joining algorithm for quartet data

Stefan Grünewald

For a finite set X of species, let m be a function that maps every quartet tree Q to a weight $m(Q)$, where a quartet tree is a binary tree with exactly four leaves drawn from the set X. A quartet tree Q is defined to be displayed by a binary X-tree T, if T contains a subdivision of Q. The problem to find a binary X-tree for which the sum of the weights of all displayed quartet trees is maximized is called the quartet puzzling problem, and it is known to be NP-hard, thus all known exact algorithms solving this problem are too slow in the case of large species numbers n. A fast ($O(n^4)$) algorithm is presented that produces a "good" but not necessary a best binary tree. The algorithm starts with the star graph $T = K_{1,n}$ with the central vertex z. In every step, two neighbours u, v of z are joined (which means the edges uz and vz are removed, and a vertex w and edges uw, vw, wz are inserted) such that the average weight of the quartet trees displayed in the obtained tree but not in the tree considered so far is maximized.

Assessing the variability of sequence alignment

Arndt von Häseler
(joint work with Roland Fleißner, Dirk Metzler, Anton Wakolbinger)

We present an MCMC method to sample alignments and mutation parameters of pairs of DNA or amino acid sequences simultaneously. This approach allows to estimate the variability of alignments and mutations parameters. Although our model is an oversimplification of the true evolutionary process, it provides some insights into regions of the sequences that can be aligned without any ambiguity and regions that can not be aligned with certainty. We give one example where our approach actually helps to improve an alignment.

Pitfalls of molecular phylogeny which we encountered in the study of eutherian evolution

Masami Hasegawa

Molecular phylogenetics is crucial in understanding the organismal evolution, but the methods currently used for estimating trees are still immature. In this talk, I presented several examples of pitfalls of molecular phylogeny which we have encountered in the study of eutherian (placental mammal) evolution:
(i) long branch attraction,
(ii) inappropriate model of substitution,
(iii) insufficient species sampling,
(iv) insufficient topology search.

Delta-plots: A visual aid for determining how well your data is represented by a tree

Barbara Holland

In many data sets, the historical (tree-like) signal is obscured by other processes such as selection, parallel mutations, recombination, or other biases in the underlying substitution process. Tree estimation methods will output trees regardless of how poorly a tree model really explains the data. We present a tool based on statistical geometry called the Delta-plot that can be used prior to tree estimation in order to determine how much tree-like signal exists in the data. The method is extended to rank taxa in the data in order to assess how much they confound the tree-like signal. Removing taxa based on this ordering improves the accuracy of tree construction on the remaining set of taxa, and also improves other measures of tree-likeness such as the L_2 measure. Simulations show that the method has potential to identify recombinant taxa.

From sphere to torus: cosa il topo ha da dire al toro

Harald Jockusch

Complex body plans of multicellular animals (metazoa) are based on boundaries formed by cell sheets, the epithelia. From the viewpoint of mathematical topology, epithelia (like cellular membrane systems) can be described as orientable 2D surfaces embedded in 3D space, i. e. there are no edges and self-intersection does not occur. Thus, an animal with the topology of a Klein bottle will never be discovered. Early developmental stages

of metazoa like the blastulae and gastrulae as well as the evolutionary ancient Cnidarian polyps and medusae (jellyfish) are topological spheres and the same holds for the most primitive bilateria, the flatworms, with only one opening of the gut. In colonial polyps and in higher bilateria (i. e. the rest of the metazoa, including ourselves), the formation of secondary body openings , and thus the transition from sphere (genus $g = 0$, Euler characteristic $\chi = 2$) to the torus ($g = 1$, $\chi = 0$) or higher toroids ($g > 1$, $\chi < 0$) has evolved. The cellular mechanisms and the genes responsible for this "breakthrough" are presently unknown. Their detection would add to our understanding of the emergence of complex body plans.

Environmental factors restrict horizontal gene transfer

Jim Lake

Recent genome analyses have suggested that some gene types (operational genes) are readily horizontally transferred, whereas others are not. Thus, in some sense, operational genes form a world wide organism. In this work, we studied the effects of environmental factors (growth temperature, pH, etc.) on horizontal gene transfer (HGT), and found that HGT has positive associativity with temperature, i.e. genes from high temperature organisms are preferably exchanged with other high temperature organisms, etc.

A chloroplast story

Peter Lockhart

As sequences diverge, information is lost and the biochemical and cellular constraints can give rise to apparent phylogenetic signals that may not reflect the true underlying phylogeny. This potential problem was discussed in the context of our understanding of chloroplast origins.

How many genes in Arabidopsis come from cyanobacteria

William Martin

Chloroplasts were once free-living cyanobacteria. During the process of organelle genome reduction, they donated many genes to the nuclear chromosomes through a process called endosymbiotic gene transfer – a special case of horizontal gene transfer. Using BLAST data base searches and maximum likelihood protein phylogenies, we have compared all 24990 proteins encoded in the Arabidopsis genome to the proteins encoded in three cyanobacterial genomes, 16 other reference prokaryotic genomes, and yeast as an internal control. We can show that about 4500 protein-coding genes in the Arabidopsis genome (18% of the total) are more similar to their cyanobacterial homologues then they are to the homologues from any other genome sampled, indicating that those genes were acquired from the cyanobacterial antecedent of plastids. The Arabidopsis proteins encoded by these laterally acquired genes encompass all functional classes and the majority of them are targeted to all compartments other than chloroplast. Overall, these quantitative estimates indicate that the genetic contribution of the ancestor of plastids to the complement of plant nuclear genes substantially exceeds previous expectations and that the impact of this transfer extends far beyond the organelle from which these genes were laterally acquired, uncovering a vast genomic heritage of plastid origins in nuclear chromosomes.

Whole genome assembly: tactical and strategic implications

Eugene Myers

The assembly of a shotgun data set of end-sequence reads was considered controversial at the time Celera proposed to apply the method to the human genome in mid 1998. Critics claimed that the computations would involve an impossible amount of computer time, that the size and repetitiveness of the genome would confound all attempts at assembly should sufficient computer efficiency be achieved, and that even if an assembly was produced it would be of an extremely poor quality and partial nature.

In 1999, the informatics research team at Celera produced an assembly of the Drosophila genome from a whole genome shotgun data set consisting of 3.2 million reads, 72% of which were paired-end reads of 2Kbp and 10Kbp inserts in a 1 to 1.32 mix. The assembly consisted of completely ordered and oriented contigs covering an estimated 97.2% of the genome with only 1630 gaps of average size 1,415bp.

Results of high quality have now been obtained for the human and mouse genomes, and our reconstruction of the human genome has been reported upon in Science. We discuss our algorithmic approach, the strategic pros and cons of the method, and the implications for the future of bioinformatics.

Time scales and evolutionary processes

David Penny

The same basic models of evolution apply for time-scales from generations, to populations, to genera, to long-range macroevolution. However, what we observe depends on the time scale of our observation. At shorter times, ancestral character states (haplotypes) are still present in the population, and there should be non-binary trees. Between generations we observe deleterious as well as neutral mutations. For a variety of reasons (including insufficient knowledge), there may be cycles in the graph, leading to non-tree models. Again, over the shorter term, sites in a sequence are expected to remain in the same rate class, and a gamma model for the distribution of rates may be appropriate. A small number of hypervariable sites may dominate the observed rate. Over longer times, almost all these features change. We move toward a tip-labeled binary tree, and expect deleterious mutations to be eliminated. There are changes in the 3-D structure of macromolecules favouring covarion models, and in some cases non-stationary models. Similarly, the effect of a small number of fast evolving sites may not be observed because they can no longer be aligned. Thus, we measure (genuinely) different rates at different time periods, even though the underlying process is identical. However, over all time periods we tend to use the same analytical models. That is, we assume that sequences evolve on binary trees with mutations and duplications (or speciation) occurring only at the tips, and with substitutions occurring in the same manner and rate at a given site throughout the tree. We need to discuss the consequences of a single process measure at very different time-scales. Additional sequences may be more important for consistency.

Covarions, rates, recombination and lateral gene transfer; problems in recovering a molecular phylogenetic tree of life

Andrew Roger

Deriving a molecular "tree of life" is hindered by problems of two sorts:
(a) methodological artefacts leading to conflicts between trees,
(b) biological effects such as horizontal (lateral) gene transfer between distantly related genes or within-gene recombination events.

One major methodological effect concerns the stationarity of the evolutionary rates at sites across deep splits in the tree of life. We have detected significant changes in the rates at sites between paralogs in the elongation factor GTPase superfamilies. Such changes appear to relate to shifts in the functional properties of molecules such as binding of tRNA and proteins. Such changes violate widely used Markov model for sequence change that assumes rates are constant through time and could introduce serious bias into our estimate of the tree of life. Recombination also may be prevalent between distantly related, but homologous sequences – thus a single bifurcating tree does not adequately describe the history of many genes and genomes. Maximum likelihood methods are very powerful in detecting recombination if used in a sliding window method. Significance statistics for likelihood ratios between trees must rely on parametric bootstrapping techniques.

Information theoretic limits to phylogeny reconstruction

Mike Steel

Information theory provides a convenient vehicle for expressing and deriving limits on the extent to which deep species divergences can be recovered from sequence data. In this talk, I apply some recent results of Evans et al. (Adv. Appl. Prob. 2000) to derive explicit and readily applicable bounds when sequences evolve under a 2-state Markov model. I also describe some of the issues to be considered in extending these bounds to multi-state data.

Surprising results of molecular systematics and the search for an objective a-priori estimation of data quality

Johann Wolfgang Wägele

Many scientists that reconstruct relationships and phylogeny of living organisms tend to believe that a tree with good "statistical support" (e.g. high bootstrap-support values) derived from an alignment of DNA sequences represents the real phylogenetic history. However, the tree may not be correct when the data are not informative, and in some cases it can be shown that such a tree is not plausible when additional data not used for tree construction do not fit to the scenario derived from the estimated optimal topology. "Not informative" means that signals are not better than background noise, even if they are mutually compatible and fit to a tree topology. We show that spectra of split-supporting positions are a promising tool for exploratory data analysis. This tool is independent of any assumption about tree topologies and can correct for multiple substitutions. For this purpose the occurrence of noise (characters different from e.g. ingroup consensus character states) is allowed in supporting positions as long as patterns of noise seem to be random.

Using spectra one can show that adding informative sequence positions to an alignment signal accumulates in few splits but noise scatters over a large number of "nonsense splits".

Therefore, with increasing sequence length the relation signal:noise also increases, as expected. Spectra also allow direct comparison of the information content of alignments without the necessity to construct trees.

The DNA-alignment used for the original substantiation of the Ecdysozoa hypothesis (the clade Arthropoda + Cycloneuralia) is discussed. The signal in favour of the Ecdysozoa present in an 18SrDNA alignment is not better than background noise detected in spectra of supporting positions. Furthermore, additional morphological characters supporting the competing Articulata hypothesis (Arthropoda + Annelida) are of higher complexity than characters supporting the Ecdysozoa and they fit to an evolutionary scenario that explains evolution of locomotion (derived from a tree with the Articulata hypothesis on which behaviour and characters important for locomotion can be plotted).

Solving the quartet puzzling problem as integer linear optimization problem

Jan Weyer-Menkhoff

The quartet puzzle method is one of many methods to reconstruct phylogenetic trees. For using this method, an algorithm for solving a certain optimization problem has to be developed. This can be done by translating the original problem into the language of integer linear programming.

In this talk, I will explain how this can be done, how the linear optimization problem can be solved fast enough, and I will also present some first results, and report some first experiences of using this approach.

Edited by Stefan Grünewald

Participants

Prof. Dr. Michael Baake
mbaake@uni-greifswald.de
Fachrichtung Mathematik/Informatik
Universität Greifswald
Friedrich-Ludwig-Jahn-Str. 15a
17489 Greifswald

Prof. Dr. David Bryant
bryant@math.mcgill.ca
Dept. of Mathematics and Statistics
McGill University
Burnside Hall
805 Sherbrooke Street West
Montreal QC, H3A 2K6
CANADA

Prof. Dr. Andreas Dress
dress@mathematik.uni-bielefeld.de
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld

Daniel Eberhard
daniel.eberhard@biologie.Uni-Bielefeld.DE
Developmental Biology & Molecular Pathology
University Bielefeld
33501 Bielefeld

Prof. Dr. David B.A. Epstein
dbae@maths.warwick.ac.uk
Mathematics Institute
University of Warwick
Gibbet Hill Road
GB-Coventry, CV4 7AL

Peter Foster
p.foster@nhm.ac.uk
Department of Zoology
The Natural History Museum
Cromwell Road
GB-London, SW7 5BD

Prof. Dr. Alex Grossmann
grossman@genopole.cnrs.fr
Laboratoire Genome et Informatique
Tour Evry 2
523 Place des Terraces de l'Agora
F-91034 Evry Cedex

Dr. Stefan Grünewald
grunew@Mathematik.Uni-Bielefeld.DE
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld

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

Prof. Dr. Masami Hasegawa
hasegawa@ism.ac.jp
The Institute of Stat. Mathematics
4-6-7 Minami Azabu, Minato-ku
Tokyo 106-8569
JAPAN

Barbara R. Holland
B.R.Holland@massey.ac.nz
Lehrstuhl für Spezielle Zoologie
Ruhr-Universität Bochum
Universitätsstr. 150
44780 Bochum

Jonathan Ingram
jingram@maths.warwick.ac.uk
Mathematics Institute
University of Warwick
Gibbet Hill Road
GB-Coventry, CV4 7AL

Dr. Harald Jockusch
h.jockusch@biologie.uni-bielefeld.de
Developmental Biology & Molecular Pathology
University Bielefeld
33501 Bielefeld

Prof. Dr. Jim Lake
lake@ewald.mbi.ucla.edu
232 Molecular Biology Institute
UCLA
611 Charles Young Drive
Los Angeles, CA 90095
USA

Dr. Peter J. Lockhart
p.j.lockhart@massey.ac.nz
Institute of Molecular BioSciences
Massey University
Palmerston North
NEW ZEALAND

Prof. Dr. William Martin
w.martin@uni-duesseldorf.de
Botanisches Institut II
Universität Düsseldorf
Universitätsstr. 1
40225 Düsseldorf

Prof. Dr. Eugene Myers
gene.myers@celera.com
Informatics Research
Celera Genomics
45 W Gude Drive
Rockville MD 20890
USA

Prof. Dr. David Penny
d.penny@massey.ac.nz
Inst. for Molecular BioSciences
Massey University
Private Bag 11 - 222
Palmerston-North 5300
NEW ZEALAND

Achim Radtke
radtke@eva.mpg.de
Max-Planck-Institut für evolutionäre Anthropologie
Inselstr. 22
04103 Leipzig

Dr. Andrew J. Roger
aroger@is.dal.ca
CIAR Program i.Evolutionary Biology
Dep.of Biochemisty a.Molecular Bio.
Dalhousie University
Halifax, N.S. B3H 4H7
CANADA

Prof. Dr. Mike Steel
m.steel@math.canterbury.ac.nz
Biomathematics Research Centre
University of Canterbury
Private Bag 4800
Christchurch
NEW ZEALAND

Prof. Dr. Johann Wolfgang Wägele
johann.w.waegele@rz.ruhr-uni-bochum.de
Lehrstuhl für Spezielle Zoologie
Ruhr-Universität Bochum
Universitätsstr. 150
44780 Bochum

Jan Weyer-Menkhoff
jweyer@mathematik.uni-bielefeld.de
Graduiertenkolleg Struktur-bildungsprozesse
Fakultät für Mathematik
Postfach 100131
33501 Bielefeld