

ANNOUNCEMENT

CIMPA Summer School

Mathematical and Computational Methods in Biology

January 5-16, 2004, Valdivia, Chile

Center for Scientific Studies (CECS)

Organizing Committee:

Alejandro Maass (CMM-DIM, Universidad de Chile)

Salomé Martínez (CMM-DIM, Universidad de Chile)

Elisabeth Pécou (U. Bourgogne)

Advisory Committee:

Ricardo Badilla (BioSigma S.A.)

Ramón Latorre (CECS)

Servet Martínez (CMM-DIM, Universidad de Chile)

Bernard Prum (Genopole, Evry, France)

Courses:

Hidde de Jong:

Qualitative Modeling and Simulation of Genetic Regulatory Networks

First Week: January 5-9, 2004.

1) Yuan Lou:

A Semilinear Parabolic System for Migration and Selection in Population Genetics

Abstract:

We will discuss the semilinear parabolic system that describes the evolution of the gene frequencies in the diffusion approximation for migration and selection at a multiallelic locus. The population occupies a finite habitat of arbitrary dimensionality and shape. The selection coefficients depend on position; the drift and diffusion coefficients may do so.

In Lecture 1, we will present the model and study two alleles case (the scalar case). We will concentrate on the work of D. Henry, and indicate how the global analysis of D. Henry can

be extended from homogeneous, isotropic migration (corresponding to the Laplacian) to arbitrary migration (corresponding to an arbitrary elliptic operator). Some open problems will be discussed.

In the rest of lectures, we will discuss the multiple alleles case:

(a) Sufficient conditions are given for the global loss of an allele that is nowhere the fittest. In the case of no dominance with at least one change in the direction of selection, sufficient conditions are established for global convergence to a stable equilibrium with all the intermediate allele absent and one or two extreme alleles present.

(b) Sufficient conditions are given for the global loss of an allele and for its protection from loss.

(c) Sufficient condition for the existence of at least one internal equilibrium, and the profile of any internal equilibrium in the zero-migration limit is obtained.

(d) Further discussions and open problems.

Note: Some background in partial differential equations will be helpful.

References:

i) D. Henry, "Geometric Theory of Semilinear Parabolic Equations," Lecture Notes in Mathematics, Vol. 840, Springer-Verlag, Berlin, 1981

ii) T. Nagylaki, Conditions for the existence of clines, Genetics, Vol 80, 1975, 595-615.

iii) T. Nagylaki, The diffusion model for migration and selection, "Some Mathematical Questions in Biology" (A. Hastings, Ed.), Lectures on Mathematics in the Life Sciences, Vol. 20, pp. 55-75, American Mathematical Society, Providence, R.I., 1989

iv) Y. Lou and T. Nagylaki, A semilinear parabolic system for migration and selection in population genetics, J. Differential Equations, Vol 181 (2002), 388-418

2) Michal Kowalczyk:

Transport phenomena in eukarya: continuous and discrete models for molecular motors

Abstract:

Biological systems provide an important motivation to study active processes which on a molecular scale are able to transduce chemical energy into mechanical work and motion. Important examples are linear or rotary motor enzymes (molecular motors) which move actively along DNA (track).

Models that I am going to present describe the motion of an assembly of molecular motors in terms of their distribution function. In this description unidirectional motion is achieved thanks to diffusion which tends to spread and dissipate density and transport which concentrates density at specific sites determined by the energy landscape. The result of this collaboration is unidirectional transport of mass. This is in spite of the fact that each of the components, i.e. diffusion and transport, is spatially unbiased and acting separately there is no mean net transport of mass.

In my talk I will discuss some PDE models for the motion of the molecular motors. As it turns out proving that those models indeed exhibit unidirectional transport is rather delicate and their rigorous analysis involves techniques that have been developed in recent years in the context of the Monge-Kantorovich mass transfer problem.

4) Stphane Robin:

Statistical analysis of microarray data

Abstract:

The aim of this course is to present the main issues in statistical analysis of microarray. Most popular and standard tools will be presented and discussed and some more recent approaches will be introduced. This program gives a sketch of the content of each of the four lectures.

1.- Introduction and preliminary studies: the first lecture will be devoted to a general introduction to microarray technology, to underlying biological issues and their translation into statistical terms. Planning of experiments and normalization techniques will be presented as necessary preliminary steps before further statistical analysis.

- Biological context: functional genomics.

- Microarray technology: hybridization reaction, different technologies.

- Statistical issues: the high variability of the data requires careful statistical analysis. Typical problems are class discovery, class comparison and class prediction.

- Experimental designs: how to organize the experiments in view of analyzing a specific contrast. A special attention will be paid to the glass slide technology (loop designs, star designs, swaps).

- Normalization: due to numerous technical biases, data have to be normalized. Which effects should be normalized and by which mean?

- Background: Elementary biology, basic statistics and analysis of variance.

- Bibliography: Brown and Botstein (1999), Churchill (2002), Kerr et al. (2002) 1

2.- Class discovery: the second lecture will deal with the discovery of groups of genes having similar expression profile in a given set of conditions. Clustering techniques have become most popular tools in microarray data analysis. However their relevance is not always obvious and underlying hypotheses are generally not clearly stated. Similarity and dissimilarity: the definition of groups of genes is always based on some (arbitrary) measurement of the distance between genes. Hierarchical clustering: this technique, which provides the famous Eisen plots, will be described in details and discussed. K means is another frequently used algorithm. Its properties will be presented, partly as an introduction to mixture models. Mixture models allow to take into account the variability of the data in a clustering procedure. The expectation-maximization(EM) algorithm will be described. How many groups? Some standard criterion to choose the number of groups will be presented.

- Background: Statistics and probability (likelihood, conditional probability and Bayes rule).

- Bibliography: Eisen et al. (1998), Golub et al. (1999).

3.- Class comparison: A typical problem is the detection of genes associated with some disease. Such genes will be revealed by their differential expression between, says, normal tissues and tumors. This is often called differential analysis. Hypothesis testing: the basic principles of statistical testing will be briefly reminded (test statistic, risk, power, etc). Variance modeling: the estimation of the residual variability has strong consequences on the power of the test. Gene by gene estimates lead to un-powerful tests while unique

estimate is often unrealistic. Intermediate solution must be found. Multiple testing: performing thousands of tests at the same time may lead to numerous false positives. Some multiple testing procedure will be discussed.

- Background: Hypothesis testing, risks, t-test, p-value.

- Bibliography: Dudoit et al. (2002c), Rudemo et al. (2002), Dudoit et al. (2002b), Pan (2002).

4.- Class prediction: A last typical problem is the prediction of the type of tissue (normal or tumor) according to their expression profiles. Classifiers are fitted to tissues of known types and then applied to unknown tissues. Discriminant analysis: linear (LDA) and quadratic (QDA) methods are based on a gaussian modeling of the data. Algorithmic approaches: support vector machines (SVM), classification trees (CART) or other methods that do not make distribution assumptions have often better performances. Generalization risk: classifiers always present good performances on the training data set. The problem is to evaluate error risk on new data. Gene selection: how to select a reduced subset of genes to predict the type of the tissue with a reasonably low error rate.

- Background: Conditional probability, Bayes rule and multivariate statistics. Other notions will be introduced during the lecture.

- Bibliography: Guyon et al. (2002), Dudoit et al. (2002a), Brown et al. (2000) References Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA.* 97 262267. Brown, P. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics. Supplement* 21 3337. Churchill, G. (2002). Fundamentals of experimental designs for cDNA microarray. *Nature Genetics.* 32 490495. Dudoit, S., Fridlyand, J. and Speed, T. P. (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* 97 (457) 7787. Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2002b), Multiple hypothesis testing in microarray experiments. Technical report, U.C. Berkeley, Division of Biostatistics, www.bepress.com/ucbbiostat/paper110. submitted. Dudoit, S., Yang, Y., Callow, M. J. and Speed, T. P. (2002c). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica.* 12 111139. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA.* 1486314868. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression. *Science.* 286 531537. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning.* 46 389 422. Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J. and Churchill, G. A. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica.* 12 203218. Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics.* 18 (4) 546554. and Lindahl, P. <http://www.math.chalmers.se/~rudemo/>.

Other sources. Interesting papers appear also in: Bioinformatics or J. Comp. Biology. Vol. 12 of Statistica Sinica and 32 of Nature Genetics in 2002 were specially devoted to the analysis of microarray data.

Recent and up-to-date methods are generally available on the web, several months (or years!) before their publication in regular journals. Some groups working on statistical analysis of microarray data give access to their manuscripts or slides. See for example:

- www.bioconductor.org,
- www.stat.berkeley.edu/users/terry/zarray/Html,
- www.jax.org/staff/churchill/labsite/research/expression/.

Rudemo, M., Lobovkina, T., Mostad, P., Scheidl, S., Nilsson, S. Variance models for microarray data. (2002).

4) Rinaldo Schinazi:

Using interacting particle systems to model population biology

Abstract:

We consider models for which the space is discrete (usually the lattice \mathbb{Z}^d) and the time is continuous. The state of each spatial site evolves randomly and according to the states of neighboring sites. We are interested in understanding how this local behavior affects the population as a whole. The local behavior is suggested by some biology hypothesis and analyzing the model allows us to test the biology hypothesis. Many of our models are built on the so called “contact process”.

We will start by discussing some of the properties of this fundamental process and then we will introduce several other models that will allow us to discuss some biology hypotheses like: the role of aggregation in the transmission of infectious diseases and in mass extinctions, the appearance of drug resistant diseases, the role of reinfection in infectious diseases.

Background:

- Elementary stochastic processes, in particular random walks on \mathbb{Z}^d and Galton-Watson processes at the level of book, “Classical and spatial stochastic processes” (R. Schinazi Birkhauser).
- Articles on the research of R. Schinazi can be downloaded from schinazi@cmi.univ-mrs.fr

Second Week: January 12-16, 2004.

1) Michael Waterman:

Abstract:

Lecture 1. Dynamic Programming for Sequence Alignments:

In the early 1970s various workers began to study the problem of how to arrange two biological sequences (usually protein sequences then) into an alignment. When letters from the respective sequences are located one above the other, it indicates a common evolutionary history. The deletions and insertions make the problem a combinatorial challenge, and for some time the proper solutions were unknown. However people such as David Sankoff made substantial progress as did the introduction of more sophisticated alignments which included multiple letter indels.

The reason this problem has been so central to computational biology is that the evolutionary process often conserves useful and essential sequence features, and alignment can show novel biological explanations. Often today sequence function is inferred from alignment alone.

Dynamic programming provides a flexible and easy framework for solving alignment problems. Global alignment of entire sequences, local alignment showing conservation in otherwise unrelated sequences, overlap detection for sequence assembly, and other problems are all solved in this setting. Too expensive for database searching, dynamic programming continue to be considered the most rigorous of all approaches.

Reference:

- Introduction to Computational Biology: Sequences, Maps and Genomes. (1995), M.S. Waterman. Chapman Hall, 431 pages. Chapter 9.

Lecture 2: Statistical Distributions for Sequence Alignment Scores:

Any two sequences can be aligned and random sequences can have impressive (and biologically meaningless) alignments. The goal of this lecture is to introduce the problem of assigning p-values to alignment scores. For this we need to make assumptions about the sequence letter distributions and then derive the score distributions.

Global alignments are the most obvious problem to study. The global alignment scores are those about which we know the least however. Certain tools provide some insight but a useful approximation has yet to be found.

Local alignments arise in database searching so that these distributions are the most important to biology. Surprisingly perhaps there is great progress on this statistical problem. A basic insight comes from Poisson approximation. It is this insight that gives the widely used e-values in BLAST output.

Other situations give some novel twists as arises in a word count statistic used in EST studies. The lecture may not reach this subject, however.

References:

- Introduction to Computational Biology: Sequences, Maps and Genomes. (1995), M.S. Waterman. Chapman Hall, 431 pages. Chapter 11.
- Lippert, R.A., Huang, H., and Waterman, M. (2002) Distributional regimes for the number of k -word matches between two random sequences. Proc. Natl. Acad. Sci. USA, **99** 13980-13989.

Lecture 3: Eulerian Graphs, Sequence Assembly and Multiple Alignment:

Sequencing of a large genome is often done by shotgun sequencing, where short DNA fragments are reassembled to reveal the original genomic sequence. Traditional assembly programs follow three steps: pairwise comparison, layout and multiple alignment (consensus). This approach can mis-connect fragments and contigs because of long near-perfect repeats that are often prevalent in genome sequences.

We introduce an Eulerian path approach to the DNA fragment assembly that originated with Idury and Waterman, 1995, and then advanced by Pevzner, et. al. 2001. This combinatorial approach bypasses the traditional “overlap-layout-consensus” approach and successfully resolves many of the troublesome repeats in practical assembly projects. The computational efficiency of assembly by the Eulerian path approach is significantly more efficient than traditional approaches.

As an extension, we use this Eulerian path idea to address the multiple sequence alignment problem. In particular, we target at aligning up to thousands of sequences simultaneously, which is computationally beyond the capabilities all existing alignment algorithms. As a beginning, we focus on DNA sequence alignments. Our method can align hundreds of DNA sequences within minutes as well as maintain high accuracy. In addition, it provides a linear growth in computational time as the number of sequences increases. We demonstrate its performance by alignments on simulated sequences and by an application in a base-calling project of *Arabidopsis thaliana*. As a comparison, our method outperforms ClustalW (Thompson et. al. 1994) in all sequences tested.

We conclude that the Eulerian path approach is extremely efficient in computation and accurate in results. Although having some weaknesses, this approach provides a new perspective in solving fragment assembly or multiple alignment problems.

Reference:

- Pevzner, P., Tang, H., and Waterman, M. (2001) An Eulerian path approach to DNA fragment assembly Proc. Natl. Acad. Sci. USA, **98** 9748-9753.

Lecture 4: Algorithms for Haplotype Blocks in Human Chromosomes:

A reference human genome sequence is known with good accuracy. The sequence variations between human individuals is responsible for great differences in physical properties (height and eye color, for example) and in health implications. Therefore there is great interest in collecting and studying those differences.

The human genome is comprised of chromosome pairs, which is referred to as diploid. The individual chromosomes are called haplotypes. In this paper we develop dynamic programming algorithms for haplotype block partitioning to minimize the number of representative single nucleotide polymorphisms (SNPs) required to account for most of the haplotype quality in each block. The block quality is a function of the haplotypes defined by the SNPs in the block. Any measure of haplotype quality can be used in the algorithm and of course the measure should depend on the specific application. The dynamic programming algorithm is applied to analyze the haplotype data on chromosome 21 of Patil et al. ((2001) Science **294**, 1719-1723). Using the same criteria as in Patil et al., we identify a total of 3,582 representative SNPs and 2,575 blocks which are 21.5% and 37.7%, respectively, smaller than those identified using a greedy algorithm of Patil et al.

Single nucleotide polymorphisms (SNPs) are promising markers for population genetic studies and for localizing genetic variations responsible for complex diseases. They are preferred to other genetic markers such as microsatellites because of their high abundance (SNPs with minor allele frequency greater than 0.1 occur once about 600-1000 base pairs) (1), relatively low mutation rate, and easy adaptability to automatic genotyping. It is known that studies using haplotype information generally outperform those using single-marker analysis. Thus it is important to know the haplotype structure of the whole genome in the populations under study.

In their recent paper, Patil et al. studied the global haplotype structure on chromosome 21. There are two general classes of methods to infer haplotype frequencies. One is based on genotypes on large pedigrees and the other is based on statistical methods such as the EM algorithm. Those methods deal with genotype data for diploid copies of the chromosomes. Uncertainties on haplotype phase are generally unavoidable for such data. Unlike previous

studies, Patil et al. studied haplotypes on haploid copies of chromosome 21 isolated in rodent-human somatic cell hybrids allowing the determination of full haplotypes of those chromosomes. They also found that the haplotypes can be divided into blocks of limited haplotype diversity. Only a small fraction of the SNPs are sufficient to uniquely identify the common haplotypes in each block. Those SNPs are referred as representative SNPs. The techniques to perform the haplotype block partitioning to minimize the total number of representative SNPs for the entire chromosome is the topic of this lecture.

Reference:

- Zhang, K., Deng, M., Chen, T., Waterman, M., and Sun, F. (2002) A dynamic programming algorithm for haplotype block partitioning *Proc. Natl. Acad. Sci. USA*, **99** 7335-7339.

2) Denis Thieffry:

Genetic Regulatory networks: from genetic and molecular data to qualitative dynamical modelling

Abstract:

This course will introduce the emerging theme of gene network modelling. It will encompass four parts:

- 1) Introduction to gene regulation, covering various kinds of genetic and molecular experimental data.
- 2) Overview of the main mathematical approaches used to model the dynamical behaviour of genetic regulatory networks.
- 3) Logical modelling of gene networks. Applications to gene networks involved in the control of viral and bacterial gene expression.
- 4) Modelling of the segmentation gene network during *Drosophila* early embryogenesis.

Readings:

1. A short introduction to molecular biology, with a special emphasis on gene regulation (appendix for the non biologist accompanying my HDR thesis):
http://www.esil.univ-mrs.fr/~thieffry/Texte_Appendix.pdf
http://www.esil.univ-mrs.fr/~thieffry/Figures_Appendix.pdf
2. de Jong H: Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* 2002, 9: 67-103.
An excellent extensive review of many gene networks modelling studies.
3. Reinitz J, Kosman D, VanarioAlonso CE, Sharp DH: Stripe forming architecture of the gap gene system. *Dev Genet* 1998, 23: 11-27.
Using computationally intensive algorithm (simulated annealing), the authors were able to fit a set of generic differential equations with experimental in situ expression patterns the gap segmentation module. This led them to reverse engineer gap cross-regulatory interactions, both qualitatively (signs) and quantitatively (values of the corresponding parameters), and to provide new insight regarding the regulation of pair-rule genes by gap genes.
4. Sánchez L., Thieffry D.: A logical analysis of the *Drosophila* gap gene system. *J. Theor. Biol.* 2001, 211: 115-141.

This paper describes the (multilevel and asynchronous) logical analysis of the gap regulatory module, focusing on the role of its various feedback circuits. In particular, one single positive circuit, composed of the cross-inhibitory interactions between giant and Krpple, was found to play the most crucial role, as their mutual inhibitions preclude any significant overlap between the expression domains of these two genes. This approach allows the qualitative reproduction of the wild-type patterns of gap gene expression, as well as the simulation of various types of perturbations.

5. Thieffry D., Sánchez L.: Alternative epigenetic states understood in terms of specific regulatory structures. *Ann N Y Acad Sci* 2002, 981: 135-153.

Following up with the logical analysis of the gap regulatory module, the authors provide here an extensive set of simulations, encompassing single and multiple loss-of-function mutations, cis-regulatory mutations, as well as ectopic gene expression. Many of these simulations lead to results that still await experimental testing.

6. Von Dassow G, Meir E, Munro EM, Odell GM: The segment polarity network is a robust developmental module. *Nature* 2000, 406: 188-192.

In this paper, the authors aim at modelling the spatial temporal behaviour of the segment-polarity gene network in terms of ordinary differential equations. To overcome the lack of information about parameter values to be associated with each macro-molecular interaction, the authors have designed tools to systematically explore the parameter space. Strikingly, they found that many parameter combinations can lead to expression pattern qualitatively consistent with experimental data, revealing a remarkable robustness of this genetic module.

3) Bernard Prum:

Markov Models and Hidden Markov Model in genome analysis

Abstract:

Biological sequences essentially consist in DNA chains, the chromosomes which transmit the information from a generation to the following one, and proteic chains, the proteins being the essential component of all phenomena in living cells. The first ones are written in a 4 letters alphabet $\{a, c, g, t\}$ while the second ones contain 20 letters, the amino-acid. Daily, more than 20 millions of new deciphered letters arrive in the data banks and a challenge for the statisticians is to help the biologist for finding the relevant information in this huge amount of data.

A first topic we are interested by consists in searching words whose frequency is too high to let believe it results from pure randomness. As an example, in bacterial genomes exists some signal (called CHI) which participates to their defenses and must therefore be sufficiently frequent to be efficient. Hence CHI's role is irrelevant with the usual genetic code but has another importance for the organism.

To search for these "exceptionnal" words, we look for a modelisation which could be both satisfactory for the biologist and tractable for the mathematician. One has to take into account the frequencies of the letters, of the 2-letters words, 3-letters words, etc..., hence to work conditionally to the sufficient statistics of a Markov chain model. In these models for each word W , using a conditional approach, we compute the expectation and the variance of the number of occurrences and give result about its (asymptotic) law.

A very relevant criticism done against this modelisation is that it assumes the homogeneity of the sequence, and this hypothesis is worst and worst admitted by the biologists when they deal with larger and larger sequences. One way for answering these criticisms consists in allowing the simultaneous existence of more than one markovian model and this led us to work with Hidden Markov Models (HMM). These models quickly turn out to be statistical tools permitting much more than the separate analysis of regions chose to be homogeneous. The fact that, at the beginning of the algorithm, we must nor fix the markovian transition in each state nor the positions of the various states implies that adjusting a HMM on a sequence produces its "segmentation" by allocating a common characteristic to all the segments related to a same state. An important drawback of the 'classical' modelisation by HMM is that it implies that the areas corresponding to a same state must have length distributed according to an exponential law, and this is not at all verified in the reality of genomes. Semi-markovian models solve this difficulty : they allow every law for the length of the various area.

Joined with the use of characteristics of the biological context, these methods must significantly improve the performances of the predictions of homogeneous regions. We will present a few applications as search of "horizontal transfers" and "annotation"

Since some 10 years, it is admitted that beside the vertical transmission (from parents to offsprings), a phenomenon of horizontal transmission of genetical information plays an important role in the evolution of life. For example some viruses may copy a part of the genome of some individual and transport and incorporate it in the genome of another individual - may be of an other species. The potential profit of this phenomenon is obvious : through such tranposons, a new beneficial gene can spread in a great number of species. As it is well known that each species leads to a different adjustment of a Markov model (frequencies of words change from a species to another), modelisation using HMM is perfectly adapted for searching tranposons.

The matter of "annotation" is to contribute to an automatic research in DNA sequences of coding parts, and within these of exons ans introns (in "eucaryotes" - essentially every species except bacteriae - genes contain two kinds of regions : exon message is in fine translated into the proteins, while introns desappear during the 'maturation' process). HMM is also a successful approach for this problem.