Bioinformatics and Statistical Physics, October 10/11
Saarbrücken, Germany

Talks

Thursday, October 10

Opening

9:00 R. Blossey: Bioinformatics and Statistical Physics

First Session: Bioinformatics: information from data

9:15 E. Meese: Human Cancer Biology and Bioinformatics

10:00 A. Zien: Statistical Analysis of Microarray Data - How many samples are required?

10:25 W. Huber: Robust calibration, error modeling and variance stabilization of microarray gene expression data

10:50 *Coffee Break*

11:20 S. Tornow: Relating gene expression to protein-protein interactions and functional modules using spin-spin correlations

11:45 A. Hartmann: The rare-event statistics of local sequence alignment

Second Session: DNA & RNA: denaturation and mechanical unzipping

12:10 R. Bundschuh: Modeling single molecule RNA force extension experiments

12:55 *Lunch*

14:15 D. Mukamel: Phase transitions in biopolymers: statistical mechanics of interacting loops

15:00 A. Stella: Denaturation of biomolecules and statistical mechanics of polymer networks

15:45 C.J. Benham: Stress-Induced DNA Duplex Destabilization: its statistical mechanical analysis and roles in biology

16:30 *Coffee Break*

Third Session: Biophysics: Molecules and Charges

17:00 R. Kree: Spatially Structured Voltage Signals at CA3 Synapses and LTP

17:45 V. Helms: Diffusional Dynamics of Cytochrome c Molecules in the Presence of a charged surface

18:10 H.M. Harreis: Phase behaviour of columnar DNA assemblies

18:35 *Closing Statement*

Friday, October 11

Fourth Session: Population dynamics

9:00 J.O. Indekeu: Discrete local population model with a carrying capacity

9:25 A. v. Haeseler: Inference of Population History

10:10 J. Krug: Punctuated evolution for the quasispecies model

10:55 *Coffee Break*

Fifth Session: Biomolecules: Structure

11:30 P. Grassberger: Growth Algorithms for Lattice Heteropolymers at Low Temperatures

12:15 *Lunch*

13:15 F. Seno: Learning effective amino acid interactions through iterative stochastic techniques

14:00 O. Kohlbacher: Including experimental data into protein docking algorithms:
NMR-based protein-protein docking

14:25 J.T. Kim: Principles underlying the information content of protein-DNA binding sites

Sixth Session: Biomolecules: Interactions and networks

14:50 P. Uetz: Protein domain interactions: unsolved problems from an experimentalists view

15:35 *Coffee Break*

16:00 M. Lässig: Networks in Molecular Biology

16:45 I.J. Farkas: The topology of the transcription regulatory network in the yeast S. cerevisiae

17:10 F. Iglói: First- and second order phase transitions in scale-free networks

17:35 *Closing remarks*

Bioinformatics and Statistical Physics
Saarbrücken (Germany)
October 10/11

Book of Abstracts

Craig J. Benham
UC Davis Genome Center
Davis, CA (USA)

Stress-Induced DNA Duplex Destabilization (SIDD) -
Its Statistical Mechanical Analysis and Roles in Biology

DNA within cells is topologically constrained into loops by periodic attachments to a matrix. This fixes the linking number Lk of each loop, the number of times either strand of the DNA duplex links through the closed loop formed by the other strand. The value of Lk is controlled by processes involving transient strand breaks. This in turn regulates the stresses imposed on the DNA within individual loops. Untwisting torsional stresses can destabilize the DNA duplex, causing its strands to separate at specific positions where the thermodynamic stability is low. This phenomenon is biologically important because strand separation is an obligatory step in the initiation of both replication and gene expression, the two main jobs of DNA. It is complicated to analyze because the topological constraint globally couples together the transition behaviors of all the base pairs in the loop.

This talk will briefly describe how DNA is topologically constrained within cells. A recently developed mathematical method will be presented for analyzing the strand separation transition within DNA loops where Lk is fixed. This method performs a formally exact statistical mechanical calculation of the equilibrium distribution of a population of identical DNA loops among all the available conformational states. The energy and conformational parameners used in this analysis are all taken from experimental measurements, so there are no free parameters in this model. Yet when it is applied to the analysis of specific DNA sequences, the results of this method are in quantitatively precise agreement with experimental measurements of the locations and extents of local strand separations. This justifies its use to predict the duplex destabilization properties of other DNA base sequences, on which experiments have not been performed.

The sites of predicted duplex destabilization within natural DNA sequences do not occur at random, but instead are closely associated with specific types of DNA regulatory elements. Examples include sites controlling the initiation and termination of gene expression, origins of replication, and positions where the DNA is attached to the chromosomal matrix. These associations suggest that stress-induced duplex destabilization may be involved in the mechanisms of function of several types of regulatory elements. A selection of predictions of the destabilization properties of regulatory regions will be presented, as time permits. The results from experiments that were performed to test these predictions will also be described.

Ralf Bundschuh
The Ohio State University
Columbus, OH (USA)

Modeling single molecule RNA force extension experiments

RNA molecules are fundamental building blocks of the molecular biology of a cell. Similar to proteins they fold into very specific shapes determined by their sequence. While it is easy to obtain the sequence of an RNA molecule, it is much harder to determine the structure it folds into. However, it is often the structure that determines the biological function of a molecule. Recently, single molecule experiments have been developed that probe the structure of an RNA molecule by unfolding it through an external force. For simple structures, these experiments reveal detailed information on the secondary structure. I present a quantitative computer model of such force extension experiments. The model is in good agreement with the experimental data on simple molecules. However, it shows that extracting structure information from force extension measurements becomes more difficult as the molecules become more complex. Since within the computer model all microscopic quantities can be directly observed, the model helps understanding why force extension curves of RNA molecules show the less signature of the secondary structure the more complex the molecule becomes. I will discuss several modifications of the experimental approach to overcome these limitations.

I.J. Farkas
Eötvös University (ELTE)
Budapest (Hungary)

The topology of the transcription regulatory network in the yeast S. cerevisiae

In the majority of single gene deletion Saccharomyces cerevisiae mutant strains the expression of a variable number of other genes is altered. This suggests the presence of a set of direct and indirect regulatory interactions between genes that together comprise a "transcription regulatory" network. Here, we quantitatively analyze the characteristics of this network revealed by microarray expression data from 287 different yeast gene deletion mutants. We find that gene expression interactions form a continuum hierarchy among the deleted genes, suggesting that similarly to biochemical reaction- and protein interaction webs, as a whole they form a robust, error-tolerant scale-free network. The current results imply uniformity in the higher-level features of biological organization with implications for the integration of cell components into a coordinated molecular machine.

coauthors: H. Jeong, T. Vicsek, A.-L. Barabasi, Z.N. Oltvai

Peter Grassberger
FZ Jülich
Jülich (Germany)

Growth Algorithms for Lattice Heteropolymers at Low Temperatures

We first give a review of the HP model as a toy model for protein folding. Next we discuss sequential importance sampling methods with re-sampling (SISR) as applied to this model. Finally, we discuss two improved versions (nPERMss and nPERMis) of the pruned-enriched-Rosenbluth method (PERM), which is a depth-first implementation of the SISR. Although these versions should have a much wider applicability, we discuss only their application to finding lowest energy configurations. nPERMis outperforms all known stochastic methods in finding the lowest known energies in all test cases but one, with CPU times which are in general lower than those of its competitors. In several cases, it found also new lowest energies. The only method comparable to nPERMis in efficiency is the "core directed chain growth" (CG) method of Beutler and Dill. But while the CG method heavily uses heuristics for the hydrophilic nucleus of proteins, no such heuristics is used in nPERMis. Finally, in contrast to claims made in the literature we find that fold prediction is speeded up by formally including a repulsive H-P interaction which favours more open (less compact) configurations, with more hydrophobic residues exposed to water.

Arndt von Haeseler
Heinrich-Heine Universität Düsseldorf and FZ Jülich
Düsseldorf and Jülich (Germany)

Inference of Population History

We introduce an approach to reveal the likelihood of different population histories that utilize an explicit model of sequence evolution for the DNA under study. Asuming this more realistic model of sequence evolution, we suggest to use a likelihood approach to estimate population history parameters. More precisely the likelihood of the data is based on the mean pairwise differences between DNA sequences and the number of variable sites in the sample. The use of likelihood ratios enables us to compare the plausibility of different demographic scenarios. The method is applied to show that the population of the Basques has expanded, whereas population size of the Biaka pygmies is most likely decreasing. Th Nuu-Chah-Nulth are conistent with a model of constant population.

coauthor: Gunter Weiss

H.M. Harreis
Heinrich-Heine Universität Dsseldorf
Düsseldorf (Germany)

Phase behavior of columnar DNA assemblies

The interaction between two stiff parallel DNA molecules depends not only on the distance between their axes but also on their azimuthal orientation. The positional and orientational order in columnar B-DNA assemblies in solution is investigated, on the basis of the electrostatic pair potential that takes into account DNA helical symmetry and the amount and distribution of adsorbed counterions. A phase diagram obtained by lattice sums predicts a variety of positionally and azimuthally ordered phases and bundling transitions strongly depending on the counterion adsorption patterns.

coauthors: A. A. Kornyshev, C. N. Likos, H. Löwen and G. Sutmann

---

Alexander Hartmann
Universität Göttingen
Göttingen (Germany)

The rare-event statistics of local sequence alignments

A new technique to obtain probability distributions in low probability regions (e.g. $p \sim 10^{-40}$) is presented. The basic idea is to map the underlying model on a physical system. The system is simulated on a computer at low temperature, such that preferably configurations with originally low probabilities are generated. Since the distribution of such a physical system is known, the original unbiased distribution can be obtained.

As an application, local alignment of protein sequences is studied. The deviation of the distribution $p(S)$ of optimum scores from the extreme-value distribution is quantified. This deviation decreases with growing sequence length.

Volkhard Helms
Max Planck Institute of Biophysics
Frankfurt (Germany)

Diffusional Dynamics of Cytochrome c Molecules in the Presence of a Charged Surface

A new Brownian Dynamics code was developed which is capable of computing trajectories of several charged spherical particles in the presence of a charged planar surface. The code takes into account electrostatic, van der Waals and hydrodynamic interactions. In this work we describe the methods used in the program and show results from calculations for Cytochrome c molecules interacting with a negatively charged lipid bilayer. This system is of particular biological interest since these molecules play a major role as electron carriers e.g. in photosynthesis. The shape and charge distribution of cytochrome c molecules can be well approximated as spherical particles with an embedded monopole and dipole and can therefore easily be handled by the program. That level of approximation makes it possible to study large systems with many (up to 100) particles over time scales up to milliseconds which would be computationally too expensive using detailed atomistic models.

coauthor: Christian Gorba

Wolfgang Huber
German Cancer Research Center
Heidelberg (Germany)

Robust calibration, error modeling and variance stabilization of microarray gene expression data

Intensities measured in microarray experiments aim at quantifying relative transcript abundance. However, the measurements are subject to multiple sources of experimental bias and variation, and statistical processing of the intensities is necessary to obtain good estimates of relative abundance. We introduce a statistical approach that comprises both the treatment of systematic biases, which may be corrected through calibration transformations (also called normalization), and of random variations, which may be accounted for in a stochastic framework. As a result, we obtain (i) estimates of the calibration parameters for each array, and (ii) a quantification of differential transcription that has more favorable statistical properties than the naive log-ratio. On several example data sets, we show that this estimate leads to superior sensitivity and specificity for the identification of differentially transcribed genes.

The approach is applicable to data from one or multiple cDNA slides, to series of Affymetrix genechips, and to series of cDNA membranes. It is implemented as a package for the statistical language R and is free for academic use.

Starting point for this work is the observation from replicate experiments that the variance of the measured intensities depends on their expected value. Usually, microarray data are analyzed on a logarithmic scale, which typically leads to larger variability in the small intensity range. This heteroskedasticity makes logarithms of ratios of intensities between different biological samples difficult to compare. Furthermore, it complicates the parameter estimation of the regression coefficients for the calibration. Our approach for dealing with these problems is based on a simple model of measurement error which comprises additive and multiplicative components. From this model, we derive a variance-stabilizing transformation of the form $h = arsinh(ax + b)$, and an estimate of differential abundance Δ_h. For large intensities, for which the multiplicative error is dominant, h coincides with the commonly used logarithmic transformation and Δ_h with the log-ratio. For smaller intensities, the transformation diminishes the relative fluctuations of the intensities, and leads to more stable estimates.

coauthors: Anja von Heydebreck, Holger Sueltmann, Annemarie Poustka, Martin Vingron

---

Ferenc Iglói
Research Institute for Solid State Physics and Optics
Budapest (Hungary)

First- and second-order phase transitions in scale-free networks

We study first- and second-order phase transitions of ferromagnetic lattice models on scale-free networks, with a degree exponent γ. Using the example of the q-state Potts model we derive a general self-consistency relation within the frame of the Weiss molecular-field approximation, which presumably leads to exact critical singularities. Depending on the value of γ, we have found three different regimes of the phase diagram. As a general trend first-order transitions soften with decreasing γ and the critical singularities at the second-order transitions are γ-dependent.

coauthor: Loïc Turban

J.O. Indekeu
Katholieke Universiteit Leuven
Leuven (Belgium)

Discrete local population model with a carrying capacity distribution

A discrete model for the growth of a local biological population N(x,t) is derived from a hierarchical random deposition process previously studied in statistical physics. Two biologically relevant parameters, the probabilities of birth, B, and of death, D, determine the carrying capacity K. Due to the randomness the population depends strongly on position, x, and there is a distribution of carrying capacities. This distribution has self-similar character owing to the hierarchy. The most probable carrying capacity and its probability are studied as a function of B and D. The growth rate decreases with time, roughly as in a Verhulst process. The model is possibly applicable, for example, to bacteria in a spatially inhomogeneous biofilm exposed to random local chemical contamination (antibiotic spray) and an overall temperature change away from optimal growth conditions.

coauthor: K. Sznajd-Weron

Jan T. Kim
Universität Lübeck
Lübeck (Germany)

Principles Underlying the Information Content of Protein-DNA Binding Sites

Empirically, it has been observed in several cases that the information content of transcription factor binding site sequences R_{seq} approximately equals the information content of binding site positions R_{freq}. A general framework for formal models of transcription factors and binding sites is developed to address this issue. Measures for information content in transcription factor binding sites are revisited and theoretic analyses are compared on this basis. These analyses do not lead to consistent results. A comparative review reveals that these inconsistent approaches do not include a transcription factor state space. A state space for mathematically representing transcription factors with respect to their binding site recognition properties is therefore introduced into the modelling framework. Maximum entropy analysis of the resulting comprehensive model shows that the structure of genome state space favours equality of R_{seq} and R_{freq} indeed, but the relation between the two information quantities also depends on the structure of the transcription factor state space. Further investigation and biological arguments show that the effects of the structure of the transcription factor state space on the relation of R_{seq} and R_{freq} are strongly limited for systems which are autonomous in the sense that all DNA binding proteins operating on the genome are encoded in the genome itself. This provides a theoretical explanation for the empirically observed equality.

coauthors: T. Martinetz, D. Polani

Oliver Kohlbacher
Zentrum für Bioinformatik, Universität des Saarlandes
Saarbrücken (Germany)

Including experimental data into protein docking algorithms: NMR-based protein-protein docking.

Protein-protein docking tries to predict the three-dimensional structure of a protein complex from the structure of its constituent proteins. Crucial in all algorithms for protein docking is the accurate prediction of the binding free energy of that complex to separate the true complex structure from decoys. Since current (empirical) energy functions often lack the necessary accuracy to make this distinction, success rates are still not satisfying. The approach we present here includes experimental data (unassigned 1D ^{1}H-NMR spectra of the protein complex) into the docking algorithm. Whereas the complexity of the data itself does not allow the direct determination of the complex structure without further experiments, it still contains a large amount of structural data. By simulating the NMR spectra of a number of tentative complex structure and scoring them with the deviation between experimental and theoretical spectrum, we are able to improve the prediction performance of the docking algorithm significantly.

Reiner Kree
Universität Göttingen
Göttingen (Germany)

Spatially Structured Voltage Signals at CA3 Synapses and LTP

Synaptic structures in the CNS possess extremely narrow synaptic clefts ($\approx$ 10 nm). By quantitatively calibrated MC simulations we show that under such conditions the post-synaptic potential of a single synapse exhibits characteristic and significant lateral structures, which depend upon the channel distribution within the post-synaptic membrane. Amplitude differences of up to 30 mV may occur. In a CA3 synapse with 30-50 AMPA channels and less than 10 NMDA channels, the laterally structured voltage generally reduces the current response, but the extra voltage is sufficient to remove the Mg^{++} block from NMDA channels. Thus we conclude that the proposed role of NMDA channels as detectors of synchronous incoming, presynaptic signals and outgoing spike activity, which will cause depolarization, is spoilt in active CA3 synapses. In hippocampus, active synapses are accompanied by so called "silent synapses", which lack AMPA channels. There is some evidence that they are transformed into active channels during LTP. Our simulation results suggest, that silent synapses are able to learn in a Hebb-like fashion, if the detection of synchronous acivity generates a metabolic signal to recruit AMPA channels. So we propose that silent synapses do not speak up until they have learnt something, but on the other hand active synapses dont learn.

coauthor: Volker Binding

Joachim Krug
Universität Essen
Essen (Germany)

Punctuated evolution for the quasispecies model

Evolution in a sequence space with random fitnesses is studied within Eigen's quasispecies model. A strong selection limit is employed, in which the population resides at a single point at all times. Evolutionary trajectories start at a randomly chosen sequence and proceed to the global fitness maximum through a small number of intermittent jumps. The distribution of the total evolution time displays a universal power law tail with exponent -2. The evolutionary dynamics is very well represented by a simplified shell model, in which the sub-populations at local fitness maxima grow independently. The shell model allows for highly efficient simulations, and provides a simple geometric picture of the evolutionary trajectories.

Michael Lässig
Universität zu Köln
Köln (Germany)

Networks in Molecular Biology

Molecular networks are an important link between genetic information and phenotype. They play a role, e.g., in gene regulation and as a map of protein interactions. In the latter case, proteins are the network nodes, and a link between two proteins means that they physically interact and can form a bound pair in vivo. Already the currently available data for yeast lead to complex and seemingly random networks with a few thousand nodes and links. We discuss a stochastic evolution model for these networks which is able to predict statistical properties of the observed structures. Understanding the relationship between structure and dynamics will be crucial for discerning the functional properties shaped by selection from random network features.

Eckart Meese
Universität des Saarlandes
Homburg (Germany)

Human Cancer Biology and Bioinformatics

Numerous concepts of cancer pathogenesis have been proposed to further a better understanding of human cancer development. Although there is a general agreement about cancer development as a result of complex alterations involving genetics, cell biology, physiology and environment, the overwhelming majority of these concepts is rather narrowed and strongly biased towards a particular research area. Bioinformatics methods are essential to bridge the different fields and to arrive at more integrated concepts of cancer pathogenesis. The large body of knowledge in cancer biology notably in genetic, cell biology and in immunogenetics has reached a level that allows to meaningful interrelate data of the different fields. However, as of yet attempts to systematically collect cancer biology data are mostly focussing on defined research areas such as cytogenetics where the Catalog of Chromosome aberrations in Cancer offers information of more than 30 000 cases with an aberrant cancer karyotype. In other fields of cancer biology, equivalent databases are still missing let alone databases that bridge the different fields in cancer research. The purpose of the overview is to address promising avenues for bioinformatics in the different areas of cancer biology.

David Mukamel
Department of Physics of Complex Systems, Weizmann Institute
Rehovot (Israel)

Phase transitions in biopolymers: statistical mechanics of interacting loops

Phase transitions which take place in double stranded DNA and in single stranded RNA molecules will be discussed. Examples include thermal denaturation of DNA in which the two strands unbind upon heating, unzipping of DNA in which the unbinding of the two strands is induced by an external pulling force and conformational changes which take place in RNA. Theoretical modelling of these transitions will be presented and the effect of excluded volume interactions, which play an important role in these transitions, will be discussed.

Flavio Seno
INFM and Dipartimento di Fisica, Universita' di Padova
Padova (Italy)

Learning effective amino acid interactions through iterative stochastic techniques

The prediction of the three dimensional structures of the native state of proteins from the sequences of their amino-acids is one of the most important challenges in molecular biology. An essential ingredient to solve this problem with coarse-grained models is the task of deducing effective interaction potentials between the amino-acids. Over the years several techniques have been developed to extract potentials that are able to discriminate satisfactorily between the native and non-native folds of a pre-assigned protein sequence. In general, when these potentials are used in actual dynamical folding simulations, they lead to a drift of the native structure outside the quasi-native basin. In this talk we present and validate an approach to overcome this difficulty. By exploiting several numerical and analytical tools we set up a rigorous iterative scheme to extract potentials satisfying a pre-requisite of any viable potential: the stabilisation of proteins within their native basin (less than 3-4 A cRMS). The scheme is flexible and is demonstrated to be applicable to a variey of parametrizations of the energy function and provides, in each case, the optimal potentials. In the second part of the talk, we will show how the method can be applied to obtain pairwise contact potentials between amino acid residues in transmembrane helical proteins. With the learnt potentials, the association of the helices in the protein bacteriorhodopsin is successfully simulated. The folding of a second TMP (the helix-dimer glycophorin A) is then accomplished with only a refinement of the potentials from a small number of decoys.

Attilio L. Stella
Università di Padova
Padova (Italy)

Denaturation of biomolecules and statistical mechanics of polymer networks

Conformational transitions like the thermal denaturation of double stranded DNA have attracted continual interest of the statistical mechanics community for over fourty years. This interest increased recently due to the development of biologically motivated physics and to the introduction of single molecule manipulation techniques which, e.g., make mechanical denaturations also feasible. I will review recent advances made in the determination of the nature and the universal features of such transitions for long molecules. The discussion will focus on double strand stiffness, sequence heterogeneity and excluded volume interactions as possible relevant factors in determining such features, and will encompass the case of RNA denaturation. The different scalings encountered can be understood by approximate representations of the macromolecules as homo- or block co-polymer networks of suitable topology, for which field theoretical and even exact results are sometimes available.

Sabine Tornow
GSF German National Center for Health and Environment
Neuherberg (Germany)

Relating gene expression to protein–protein interactions and functional modules using spin–spin correlations

We present a novel technique for an integrative data analysis of genetic networks, like metabolic pathways, protein interaction networks, as well as functional catalogs and expression data. The basic idea is to infer function from different types of information or networks which may generate more meaningful hypotheses than from, e. g., expression data alone. In addition, we are able to score parts of the networks. So far, integrative methods were used based on mean Pearson correlations which are not robust against noise and for which multiple correlations are not defined. Here, we propose to evaluate the correlations with the Super-Paramagnetic (SP) approach which are robust against noise, are naturally expended to multiple genes and allow the definition of their strength. With the SP approach we construct a graph in which genes are nodes assigned with a Potts spin. The nodes are connected with edges weighted with the coupling constant J, a fast decreasing function of the dissimilarity measure of two gene vectors. To assess the correlation strength of genes which are members of the same module of a genetic network we define a p-value which gives the probability that the strength was found by chance. The strength corresponds to the temperature of the transition from the ferromagnetic state of the module to the paramagnetic disordered state. Our method lies between an unsupervised and supervised analysis: we are able to produce new hypotheses but do not disregard related biological knowledge leading us to a more comprehensive picture. We find many significant functional groups, complexes and highly significant protein interactions in various gene expression experiments under different conditions.

coauthor: H.W. Mewes

Peter Uetz
FZ Karlsruhe
Karlsruhe (Germany)

Protein domain interactions:
unsolved problems from an experimentalists view

Despite the large number of protein-protein interactions generated by experimental high-throughput approaches, little is known about the protein domains that mediate those interactions. Even when the structure of protein domains and their ligands is known, interactions of related proteins is hard to predict. We have studied several protein domains of little-known function (such as the PX, FF and DEP domain) experimentally by two-hybrid screening and other methods. In many cases there are few common sequence features among the ligands of a single domain and hence it remains difficult to derive rules for predicting such partners. Finally, even when sufficient data is available, it remains a challenge to visualize protein-domain interactions in a satisfactory way.

Alexander Zien
Fraunhofer SCAI
Bonn (Germany)

Statistical Analysis of Microarray Data - How Many Samples Are Required?

The number of microarrays is estimated that is required in order to gain reliable results from a common type of study: the pairwise comparison of different classes of samples. Current knowlegde seems to suffice for the construction of models that are realistic with respect to searches for individual differentially expressed genes. Such models allow to investigate the dependence of the required number of samples on the relevant parameters: the biological variability of the samples within each class; the fold changes in expression; the detection sensitivity of the microarrays; and the acceptable error rates of the results.