A U S H A N G

FREIE UNIVERSITÄT BERLIN

Fachbereich Mathematik und Informatik

Promotionsbüro, Arnimallee 14, 14195 Berlin

D I S P U T A T I O N

Donnerstag, 16. April 2015, 13.00 Uhr

Ort: Max-Planck Institut für molekulare Genetik, Ihnestraße 63-73, 14195 Berlin, Seminarraum SR I

Disputation über die Doktorarbeit von

Herrn Jonas Maaskola

Thema der Dissertation:

Discriminative Learning for Probabilistic Sequence Analysis

Titel der Disputation:

Graphical Probabilistic Models - Independence, Inference, and Learning

Die Arbeit wurde unter der Betreuung von **Prof. Dr. M. Vingron** durchgeführt.

Abstract: *Lecture: Graphical Probabilistic Models - Independence, Inference, and Learning*
Probability theory provides a sound foundation for reasoning about the true state of the world, allowing us to form conclusions in spite of ubiquitous uncertainty. Yet, the state space of models with numerous variables, required to represent complex systems, quickly grows too large as to still be enumerable. Graphical probabilistic modeling offers a framework for representation, inference, and learning that is fit to deal with the challenges posed by such large state spaces. Bayesian and Markov networks respectively are directed and undirected graphical representations of families of probability distributions. Their graphical structure is intimately related to independence assumptions that hold for the corresponding distributions. These independencies allow for compact factorizations of the distributions, making computation of marginal and conditional probabilities feasible. There is thus a close correspondence between graphical structure, sets of independence assumptions, and families of probability distributions. The framework of graphical probabilistic modeling offers general algorithmic methods of inference. As we will see, the complexity of these algorithms hinges both on topological properties of the graphical representation, as well as on the order in which variables are eliminated. Finally, we will briefly discuss how methods of inference form the basis of parameter learning in models of complex systems. Graphical probabilistic models find pervasive application in science, examples of which will be presented in this lecture, highlighting models of particular importance for biological applications, such as the naïve Bayes model and hidden Markov models.

Abstract *Short Talk: Discriminative Discovery of Sequence Motifs*
Discrover [1] is a discriminative discovery method for binding site patterns in nucleic acid sequences based on IUPAC regular expressions and hidden Markov models. It mines sets of positive and negative example sequences for sequence motifs whose occurrence frequency varies between the sets. Discrover is applicable to genome- and transcriptome-scale data, makes use of available repeat experiments, and aside from binary contrasts also more complex data configurations can be utilized. In a systematic comparison with numerous published motif finding tools, the method achieves the highest motif discovery performance, while being faster than most published methods. Practicality and utility are demonstrated in case studies for embryonic stem cell transcription factors and for RNA-binding proteins with data from various technologies, including ChIP-Seq, RIP-Chip, and PAR-CLIP. For the alternative splicing factor RBM10 our analysis finds motifs known to be splicing-relevant.

[1] Maaskola, J. and Rajewsky, N. (2014). Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. Nucleic Acids Res, 42(21):12995–13011.

Die Disputation besteht aus dem o. g. Vortrag, danach der Vorstellung der Dissertation einschließlich jeweils anschließenden Aussprachen.

Interessierte werden hiermit herzlich eingeladen

Der Vorsitzende der Promotionskommission
Prof. Dr. M. Vingron