Bayesian methods in structural bioinformatics

Hamelryck, Thomas Wim; Mardia, Kanti; Ferkinghoff-Borg, Jesper

Link to article, DOI:
10.1007/978-3-642-27225-7

Publication date:
2012

Document Version
Early version, also known as pre-print

Statistics for Biology and Health

For other titles published in this series, go to
http://www.springer.com/series/2848

Thomas Hamelryck • Kanti Mardia
Jesper Ferkinghoff-Borg
Editors

Bayesian Methods in Structural Bioinformatics

Springer

Editors

Thomas Hamelryck
University of Copenhagen
Department of Biology
Bioinformatics Centre
Copenhagen
Denmark

Jesper Ferkinghoff-Borg
Technical University of Denmark
Department of Electrical Engineering
Lyngby
Denmark

Kanti Mardia
University of Leeds
School of Mathematics
Department of Statistics
Leeds
United Kingdom

Statistics for Biology and Health Series Editors

M. Gail
National Cancer Institute
Bethesda, MD
USA

A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC
USA

Klaus Krickeberg
Le Châtelet
Manglieu
France

W. Wong
Department of Statistics
Stanford University
Stanford, CA
USA

Jonathan M. Samet
Department of Preventive Medicine
Keck School of Medicine
University of Southern California
Los Angeles, CA
USA

ISSN 1431-8776
ISBN 978-3-642-27224-0 e-ISBN 978-3-642-27225-7
DOI 10.1007/978-3-642-27225-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012933773

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Thomas Hamelryck dedicates his contributions to his mother Gib, and to the memory of his father Luc (1942–2011).

Kanti Mardia dedicates his contributions to his grandson Sashin Raghunath.

Foreword

The publication of this ground-breaking and thought-provoking book in a prestigious Springer series will be a source of particular pleasure and of stimulus for all scientists who have used Bayesian methods in their own specialized area of Bioinformatics, and of excitement for those who have wanted to understand them and learn how to use them but have never dared ask.

I met the lead author, Dr. Hamelryck, at the start of his career, when, as part of his PhD in Protein Crystallography, he determined and proceeded to analyze in great detail the 3D structures of several tight protein-carbohydrate complexes. His attention was drawn to Bayesian and related statistical methods by the profound impact they were having at the time on the two workhorses of macromolecular crystallography, namely experimental phasing and structure refinement. In both cases, recourse to the key Bayesian concept of marginalisation with respect to the phases (treated as nuisance parameters) freed those techniques from the limitations inherent in their old, least-squares based implementations, and gave access, through a shift to a maximum-likelihood approach, to much higher quality, much less biased electron-density maps and atomic models. Preferring the world of computational methods to the biochemist's bench, Dr. Hamelryck went for a post-doc in Copenhagen, in the laboratory where he subsequently developed a highly productive structural extension to the already prominent Bioinformatics Department headed by Prof. Anders Krogh. The contents of this book include a representative sample of the topics on which he and his collaborators have focussed their efforts since that time.

The use of advanced statistical methods in Bioinformatics is of course well established, with books such as *Biological Sequence Analysis – Probabilistic Models of Proteins and Nucleic Acids* by Durbin, Eddy, Krogh and Mitchison setting a very high standard of breadth, rigour and clarity in the exposition of the mathematical techniques involved and in the description of their implementations. The present book, however, is the first of its kind in directing the arsenal of Bayesian methods and their advanced computational tools towards the analysis, simulation and prediction of macromolecular (especially protein) structure in three dimensions.

This is an extremely ambitious goal, as this area comprises no less than the Protein Folding Problem, arguably the most fundamental riddle in the whole of the Life Sciences.

The book begins with an up-to-date coverage, in Part I, of the concepts of Bayesian statistics and of the versatile computational tools that have been developed to make their use possible in practice. A particular highlight of the book is found in Part II, which presents a review of the time-honored use of knowledge-based potentials obtained by data mining, along with a critical reassessment of their exact relationship with potentials of mean force in statistical physics. Its conclusion is a highly original and welcome contribution to the solution of this long-standing problem. Another highlight is the combined use in Part V of the type of directional statistics that can be compiled by the methods of Part IV with the technology of Bayesian networks to produce very efficient methods for sampling the space of plausible conformations of protein and RNA molecules.

It is fitting that the book should return in its final Part VI to the interface between Bayesian methods of learning the structural regularities of macromolecules from known 3D structures on the one hand, and experimental techniques for determining new 3D structures on the other. As is well known, most of these techniques (with the exception perhaps of ultra-high resolution X-ray crystallography with plentiful sources of experimental phase information) need to be supplemented with some degree of low-level a priori knowledge about bond lengths and bond angles to enforce sensible stereochemistry in the resulting atomic models. The Bayesian picture turns this conventional approach on its head, making the process look instead like structure prediction assisted by X-ray (or NMR) data, with the measurements delivered by each experimental technique providing the likelihood factor to supplement the prior probability supplied by previous learning, in order to cast the final result of this inference into the form of a posterior distribution over an ensemble of possible structures. It is only by viewing structure determination in this overtly Bayesian framework that the classical problems of model bias and map overinterpretation can be avoided; and indeed crystallographers, for example, are still far from having made full use of the possibilities described here. Dr. Hamelryck is thus likely to have paved the way for future improvements in his original field of research, in spite of having since worked in a related but distant area, using tools first developed and applied to address problems in medical diagnosis. This instance of the interconnectedness of different branches of science through their methods, and this entire book, provide a splendid illustration of the power of mathematical approaches as the ultimate form of re-useable thought, and of Bayesian methods in particular as a repository of re-useable forms of scientific reasoning.

I am confident that the publication of this book will act as a rallying call for numerous investigators using specific subsets of these methods in various areas of Bioinformatics, who will feel encouraged to connect their own work to the viewpoints and computational tools presented here. It can therefore be expected that this will lead to a succession of enlarged new editions in the future. Last but not

least, this book should act as a magnet in attracting young researchers to learn these advanced and broadly adaptable techniques, and to apply them across other fields of science as well as to furthering the subject matter of the book itself.

Global Phasing Ltd.,
Cambridge, UK *Gerard Bricogne*

Preface

The *protein folding problem* is the loose denominator for an amalgam of closely related problems that include protein structure prediction, protein design, the simulation of the protein folding process and the docking of small molecules and biomolecules. Despite some, in our view, overly optimistic claims,[1] the development of an insightful, well-justified computational model that routinely addresses these problems is one of the main open problems in biology, and in science in general, today [461]. Although there is no doubt that tremendous progress has been made in the conceptual and factual understanding of how proteins fold, it has been extraordinary difficult to translate this understanding into corresponding algorithms and predictions. Ironically, the introduction of CASP[2] [400], which essentially evaluates the current state of affairs every two years, has perhaps lead to a community that is more focussed on pragmatically fine-tuning existing methods than on conceptual innovation.

In the opinion of the editors, the field of structural bioinformatics would benefit enormously from the use of well-justified machine learning methods and probabilistic models that *treat protein structure in atomic detail*. In the last 5 years, many classic problems in structural bioinformatics have now come within the scope of such methods. For example, conformational sampling, which up to now typically involved approximating the conformational space using a finite set of main chain fragments and side chain rotamers, can now be performed in continuous space using graphical models and directional statistics; protein structures can now be compared and superimposed in a statistically valid way; from experimental data, Bayesian methods can now provide protein ensembles that reflect the statistical uncertainty. All of these recent innovations are touched upon in the book, together with some more cutting edge developments that yet have to prove the extent of their merits.

[1] See for example *Problem solved* (*sort of)* in the news section of the August 8th, 2008 issue of Science, and the critical reply it elicited.

[2] CASP stands for "Critical Assessment of protein Structure Prediction".

A comprehensive treatment of probabilistic methods in structural bioinformatics is, at first sight, something that would require several weighty book volumes. However, upon closer scrutiny, it becomes clear that the use of well-justified probabilistic methods in structural bioinformatics is currently in its infancy, and that their potential is enormous. Many knowledge based methods that claim to be firmly rooted in probability theory or statistical physics are at best heuristic methods that are often only partly understood. A classic example are the so called *potentials of mean force* that make use of pairwise distances in proteins. The validity and scope of these potentials have been topics of hot debate for over twenty years. Indeed, methods that both consider biomolecular structure in atomic detail and have a sound probabilistic justification are currently far and between.

In this book, we therefore focus on methods that have two important features in common. First, the focus lies on methods that are well justified from a probabilistic point of view, even if they are approximative. Quite a few chapters make use of point estimates, such as empirical Bayes, maximum likelihood or even moment estimates. However, in all cases, these are used as valid *approximations* of a true Bayesian treatment. Second, the methods deal with biomolecular structure in atomic detail. In that respect, classic applications of probabilistic reasoning in structural bioinformatics such as secondary structure prediction fall outside the scope of the book.

This book should be of use to both novices and experts, though we do assume knowledge of structural biology. Introductory chapters on Bayesian methods and Markov chain Monte Carlo methods, which play a key role in Bayesian inference, provide methodological background. These chapters will also be useful to experts in structural bioinformatics that are perhaps not so versed in probabilistic modelling. The remaining parts address various timely topics in structural bioinformatics; we give a short overview.

As mentioned before, the first two chapters provide the foundations. In the first chapter, Hamelryck gives a high level overview of the Bayesian interpretation of probability. The chapter also touches upon relevant topics in information theory and statistical mechanics, and briefly discusses graphical models. Despite the fact that Bayesian methods are now firmly established in statistics, engineering and science in general, most university courses on statistics in these disciplines still uniquely focus on frequentist statistics. The underlying reasons are clearly beyond the usual academic inertia, and can probably be identified with two main perceived problems [51]: Bayesian statistics is unjustly seen as inherently subjective and thus unsuitable for scientific research, and the difficulty of integrating the Bayesian view with the – for many applications still dominating – frequentist paradigm. In addition, many frequentist methods can be seen as perfectly valid approximations to Bayesian methods, thereby removing the apparent need for a paradigm shift. Therefore, we believe that the introductory chapter on Bayesian statistics is quite appropriate. Ferkinghoff-Borg provides an overview of Markov chain Monte Carlo methods. These sampling methods are vital for Bayesian inference, especially when it involves the exploration of the conformational space of proteins. Together, these two chapters should provide a decent start for anybody with some knowledge of

structural bioinformatics, but a lacking background in Bayesian statistics. Both chapters focus on concepts rather than mathematical rigor and provide ample references for deeper study. However, these two chapters provide the bedrock on which the rest of the book is founded.

The second part addresses the estimation of so-called knowledge based potentials from data. Recently, dramatic progress has been made in the understanding of the statistics behind these potentials, which were previously justified using rather *ad hoc* physical arguments. The chapter by Borg et al. discusses knowledge based potentials from a physical viewpoint, and explains their statistical background. The chapter by Frellsen et al. overlaps slightly with the previous chapter, but has a more statistical slant. Knowledge based potentials can be formally understood in a Bayesian framework, which justifies, clarifies and extends them. The chapter by Frellsen et al. discusses the recently introduced *reference ratio method*, and highlights its theoretical and conceptual importance for one of the holy grails of structural bioinformatics: a rigorous and efficient probabilistic model of protein structure. Finally, the chapter by Podtelezhnikov and Wild discusses the application of another well-founded machine learning method to the construction of knowledge based potentials, namely *contrastive divergence learning*. The preliminary results reviewed in this chapter establish the method as promising for the future.

In Part III, we turn to directional statistics. Directional statistics concerns data on unusual manifolds such as the torus, the sphere or the real projective plane, which can be considered as a sphere with its antipodes identified. Examples of data from such manifolds include wind directions and dihedral angles in molecules. The need for directional statics can be understood by considering the fact that biomolecular structure is often expressed in terms of angles, and that the average of, for example, 1° and 359° is not 180 but zero. This is of course due to the fact that such data is naturally represented on the circle, rather than on the line. Hence, directional statistics is becoming vital for the formulation of probabilistic models of biomolecular structure. The two chapters by Mardia and Frellsen and by Kent are of a more technical nature and provide information on parameter estimation and sampling for two distributions from directional statistics that are of particular relevance for biomolecular structure. Kent's chapter discusses the Fisher-Bingham 5 (or Kent) distribution, which can be used to model data on the two-dimensional sphere, that is, data consisting of unit vectors. Frellsen and Mardia discuss the univariate, bivariate and multivariate von Mises distributions, for data on the circle and the torus, respectively. The latter case is of special relevance for modeling the ϕ and ψ angles in proteins, which are well known from the celebrated Ramachandran plot.

Part IV explores the use of shape theory in comparing protein structures. Comparing and superimposing protein structures is one of the classic problems in structural bioinformatics. The chapter by Theobald discusses the superposition of proteins when their equivalent amino acids are known. Classically, this is done using a least squares criterion, but this leads to poor performance in many cases. Theobald describes a maximum likelihood alternative to the classic method, and also discusses a fully Bayesian extension. The problem of superimposing protein structures when

the equivalent amino acids are *not* known is subsequently discussed in the chapter by Mardia and Nyirongo. They review a Bayesian model that is built upon a Poisson process and a proper treatment of the prior distributions of the nuisance variables.

Part V is concerned with the use of graphical models in structural bioinformatics. The chapter by Boomsma et al. introduces probabilistic models of RNA and protein structure that are based on the happy marriage of directionals statistics and dynamic Bayesian networks. These models can be used in conformational sampling, but are also vital elements for the formulation of a complete probabilistic description of protein structure. Yanover and Fromer discuss belief propagation in graphical models to solve the classic problem of side chain placement on a given protein main chain, which is a key problem in protein design.

In the sixth, final part, the inference of biomolecular structure from experimental data is discussed. This is of course one of the most fundamental applications of statistics in structural biology and structural bioinformatics. It is telling that currently most structure determination methods rely on a so-called *pseudo-energy*, which combines a physical force field with a heuristic force field that brings in the effect of the experimental data. Only recently methods have emerged that formulate this problem of inference in a rigorous, Bayesian framework. Habeck discusses Bayesian inference of protein structure from NMR data, while Hansen discusses the case of SAXS.

Many of the concepts and methods presented in the book are novel, and have neither been tested, honed or proven in large scale applications. However, the editors have little doubt that many concepts presented in this book will have a profound effect on the incremental solution of one of the great challenges in science today.

Copenhagen, Leeds

Thomas Hamelryck
Kanti V. Mardia
Jesper Ferkinghoff-Borg

Acknowledgements

We thank Justinas V. Daugmaudis for his invaluable help with typesetting the book in LaTeX. In addition, we thank Christian Andreetta, Joe Herman, Kresten Lindorff-Larsen, Simon Olsson and Jan Valentin for comments and discussion.

This book was supported by the Danish Research Council for Technology and Production Sciences (FTP), under the project "Protein structure ensembles from mathematical models – with application to Parkinson's α-synuclein", and the Danish Research Council for Strategic Research (NaBiIT), under the project "Simulating proteins on a millisecond time-scale".

Contents

Contributors

Wouter Boomsma Department of Astronomy and Theoretical Physics, Lund University, Lund, Sweden, wouter@thep.lu.se

Department of Biomedical Engineering, DTU Elektro, Technical University of Denmark, Lyngby, Denmark, wb@elektro.dtu.dk

Mikael Borg The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaløesvej 5, 2200 Copenhagen, Denmark, borg@binf.ku.dk

Jesper Ferkinghoff-Borg Department of Biomedical Engineering, DTU Elektro, Technical University of Denmark, Lyngby, Denmark, jfb@elektro.dtu.dk

Jes Frellsen The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaløesvej 5, 2200 Copenhagen, Denmark, frellsen@binf.ku.dk

Menachem Fromer The Hebrew University of Jerusalem, Jerusalem, Israel, fromer@cs.huji.ac.il

Michael Habeck Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Spemannstr. 35, Tübingen, Germany, michael.habeck@tuebingen.mpg.de

Department of Empirical Inference, Max-Planck-Institute for Intelligent Systems, Spemannstr. 38, Tuebingen, Germany

Thomas Hamelryck The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaløesvej 5, 2200 Copenhagen, Denmark, thamelry@binf.ku.dk

Steen Hansen Department of Basic Sciences and Environment, University of Copenhagen, Faculty of Life Sciences, Thorvaldsensvej 40, DK-1871 FRB C, Frederiksberg, Denmark, slh@life.ku.dk

John T. Kent Department of Statistics, University of Leeds, Leeds LS2 9JT, UK

Kanti V. Mardia Department of Statistics, University of Leeds, Leeds LS2 9JT, UK, k.v.mardia@leeds.ac.uk

Vysaul B. Nyirongo Statistics Division, United Nations, New York, NY 10017, USA, nyirongov@un.org

Alexei A. Podtelezhnikov Michigan Technological University, Houghton, MI, USA, apodtele@gmail.com

Douglas L. Theobald Brandeis University, 415 South St, Waltham MA, USA, dtheobald@brandeis.edu

David L. Wild University of Warwick, Coventry, CV4 7AL, UK, d.l.wild@warwick.ac.uk

Chen Yanover Fred Hutchinson Cancer Research Center, Seattle, WA, USA, cyanover@fhcrc.org

Acronyms

AIC	Akaike information criterion
BIC	Bayesian information criterion
BMMF	Best max-marginal first
BN	Bayesian network
BP	Belief propagation
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CD	Contrastive divergence
CPT	Conditional probability table
DBN	Dynamic Bayesian network
DEE	Dead end elimination
DIC	Deviance information criterion
EM	Expectation maximization
GA	Genetic algorithm
GBP	Generalized belief propagation
GE	Generalized ensemble
GMEC	Global minimum energy configuration
GMH	Generalized multihistogram equations
GIFT	Generalized indirect Fourier transformation
HMM	Hidden Markov model
IFT	Indirect Fourier transformation
ILP	Integer linear programming
ISD	Inferential structure determination
JT	Junction tree
KBP	Knowledge based potential
LP	Linear programming
MAP	Maximum a posteriori
MCMC	Markov chain Monte Carlo
MCSA	Monte Carlo simulated annealing
ML	Maximum likelihood
MM	Max-marginal
MPLP	Max-product linear programming

MRF	Markov random field
MUCA	Multicanonical ensemble
NMR	Nuclear magnetic resonance spectroscopy
OLS	Ordinary least squares
PDB	Protein data bank
PDF	Probability density function
PMF	Potentials of mean force
REMD	Replica-exchange molecular dynamics
RMSD	Root mean square deviation
SANS	Small angle neutron scattering
SAS	Small angle scattering
SAXS	Small angle X-ray scattering
SCMF	Self-consistent mean field
SCWRL	Side-chains with a rotamer library
SLS	Static light scattering
SPRINT	Side-chain prediction inference toolbox
STRIPES	Spanning tree inequalities and partitioning for enumerating solutions
TRMP	Tree-reweighted max-product
tBMMF	Type-specific BMMF
WHAM	Weighted histogram analysis method
WL algorithm	Wang-Landau algorithm