Annotated Alignments

Abha Singh Bais

May 2007

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Knut Reinert

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Knut Reinert

Tag der Promotion: 12th July, 2007

Abstract

Elucidating the mechanisms of transcriptional regulation relies heavily on the sequence annotation of the binding sites of DNA-binding proteins called transcription factors. With the rationale that binding sites conserved across different species are more likely to be functional, the standard approach is to employ cross-species comparisons and focus the search to conserved regions. Usually, computational methods that annotate conserved binding sites perform the alignment and binding site annotation steps separately and combine the results in the end. If the binding site descriptions are weak or the sequence similarity is low, the local gap structure of the alignment poses a problem in detecting the conserved sites. In this thesis, I introduce a novel method that integrates the two axes of sequence conservation and binding site annotation in a simultaneous approach yielding *annotated alignments* – pairwise alignments with parts annotated as putative conserved transcription factor binding sites.

Standard pairwise alignments are extended to include additional states for binding site profiles. A statistical framework that estimates profile-related parameters based on desired type I and type II errors is prescribed. This forms the core of the tool **SimAnn**. As an extension, I use existing probabilistic models to demonstrate how the framework can be adapted to consider position-specific evolutionary characteristics of binding sites during parameter estimation. This underlies the tool **eSimAnn**.

Through simulations and real data analysis, I study the influence of considering a simultaneous approach as opposed to a multi-step one on resulting predictions. The former enables a local rearrangement in the alignment structure to bring forth perfectly aligned binding sites. This precludes the necessity of adopting post-processing steps to handle errors in pre-computed alignments, as is usually done in multi-step approaches. Additionally, the framework for parameter estimation is applicable to any novel profile of interest. Especially for instances with poor sequence conservation or profile quality, the simultaneous approach stands out. As a by-product of the analysis, I also present a formulation of the annotated alignment problem as an extended pair Hidden Markov Model and illustrate the correspondence between the various theoretical concepts.

Preface

Acknowledgements After the enriching experience at IIT, both at the academic and personal fronts, little did I expect my time in Berlin to be as intense. Working at the Dept. of *Computational Molecular Biology* and interacting with an eclectic set of bright minds, I had the opportunity to grow both intellectually and culturally. Each present or past member of the group deserves thanks and I can not hope to mention all.

First of all my thanks go to Prof. Martin Vingron, my supervisor, who gave me this opportunity and, despite tight schedules, never refused a meeting. I have learned, and still am learning, from his clarity of thought as well as spoken word and his ability to think of the big picture while appreciating the simple ideas. Next, I sincerely thank Steffen Grossmann, my post-doc supervisor, a valuable friend and an excellent teacher who saw me through the most nitty-gritty phases. Through numerous, often heated, discussions with him I learned how to formulate, question and most essentially assert my views on various subjects. His ability to muddle me enough to finally answer my own question is something I would try to emulate, if ever I get the opportunity. Special thanks go to Prof. Knut Reinert for his valuable input and support as a member of my PhD committee. I also thank Stefan Haas, Holger Klein and Dennis Kostka for their helpful comments on various parts of this thesis, Ho-Ryun Chung for his help with biological examples and Aditi Kanhere for reading related work. Birgit Loehmer and Wilhelm Ruesing deserve special thanks for keeping the department machinery running smoothly, everything from cookies to computers.

My time in Berlin would not have been half as memorable if not for the now ex-members of the Dept., thanks go to – Stefan Roepcke for his honest criticism and helpful attitude, Christoph Dieterich for initiating me into the field of computational biology, Eike Staub for those short discussions that mirrored his bright enthusiasm for science, and Shobhit Gupta for encouraging me and being there without much-ado. Special thanks also go to the now ex-Computational Diagnostics Group for all the lively discussions at lunch time.

I also thank the Gene Regulation group members for the stimulating discussions. Particularly I wish to acknowledge the interesting interactions with my nomadic office-mate Utz Pape and with Hannes Luz who always took the time to listen and discuss a problem. Despite the constant electronic music buzz he lends after the first few minutes of conversation, I wish to thank Holger Klein for simply being my friend throughout. I also thank Andrea, Sharif, Dheeraj and Kshitij for making my Berlin stay an unforgettable experience.

Last but not the least, I wish to acknowledge the wonderful times spent with Inge, Arno and Berit Kostka and their words of love and encouragement. I thank my brother Apoorva for being there when I was not, my father for teaching me to strive for the challenging choices and aunty, Ruchi and Shubhi for loving me despite my short, often chaotic, visits to home. Finally, anybody who knows me, knows what a tough time it must have been for Dennis. Can I thank him enough? Guess not.

Published and related work The main contribution of this thesis has been previously published as part of the *European Conference on Computational Biology, (ECCB 2006)*, with proceedings available as a special issue of the journal *Bioinformatics*. An extension of the basic approach was accepted for presentation at the *International Conference on Bioinformatics, (InCoB 2006)*, to be published in the *Journal of Biosciences*. The formulation of the problem in a hidden markov model framework is based on initial ideas by Steffen Grossmann.

Abha Singh Bais Berlin, May 2007

Contents