Member of The Claremont Colleges

KECK GRADUATE INSTITUTE

of Applied Life Sciences

Yves Brun,
Systems Biology/Microbiology Faculty Search,
Department of Biology,
Indiana University,
Jordan Hall 142, 1001 E 3rd St,
Bloomington IN 47405-7005

Dear Dr. Brun,

I am writing to apply for the tenure-track faculty position in computational systems biology which you are currently advertising, at the rank of Associate/Full Professor. I attach a curriculum vitae, list of publications, statements of research and teaching interests and some representative publications. As you can see from these, I have extensive international experience, both academic and industrial, in computational and structural biology, gained at both world-class research institutes and in the biotech and pharmaceutical industry.

My interests and experience encompass the analysis of data arising from new high throughput molecular technologies using computational approaches and currently focus on applications in systems biology, functional genomics and proteomics. These projects involve close collaboration and communication with colleagues from both the molecular biology and informatics communities, and include NSF funded research to elucidate the gene and metabolic regulatory networks important to the mechanisms of bacterial cell function. I am keen to extend and apply these and related techniques to other data of biological interest and would actively seek to establish interdisciplinary collaborations with colleagues across the University research community, as well as maintaining existing and emerging collaborations.

In 1999, I joined the Keck Graduate Institute as one of its founding faculty, to contribute to the development of a pioneering graduate degree aimed at preparing future leaders in the bioscience industry. I have played a leading role in developing the computational biology and bioinformatics aspects of this degree: a unique program which combines training in computational and systems biology and bioengineering with aspects of management, pharmaceutical development and bioscience business awareness. A joint Ph.D. program in Computational and Systems Biology has also recently been initiated with the School of Mathematical Sciences at Claremont Graduate University. As Director of Computing at KGI, I have also been responsible for the development of the research computing infrastructure, including the development of local databases to support experimental research and the supervision of information technology support personnel.

As a biophysicist by training, with a strong track record of working in cross-disciplinary groupings, I am fluent in both biology and computational science. I believe I can play a key role in the development of an innovative and integrative research program at the interface of biology and computation in the Department.

Yours sincerely,
D L Wild
David Wild, D.Phil

535 Watson Drive
Claremont, California 91711
T: 909 607-7855 F: 909 607-8086
www.kgi.edu

Statement of Research Interests

Recent advances in genome sequencing, microarrays, proteomics and functional and structural genomics have been creating a huge amount of data which needs to be analyzed and understood. Machine learning (data mining) methods have resulted in high impact applications in a variety of domains ranging from marketing to science. There is great scope to increase the interaction between machine learning and computational biology. Machine learning methods have the potential to provide powerful tools for analyzing, predicting and understanding data from high-throughput post-genomic technologies. Most of my recent research has aimed to bridge the gap between machine learning and bioinformatics and has focused on the application of computational probabilistic modeling techniques to problems in systems biology, functional genomics and proteomics.

Modeling gene regulatory networks

This research aims to combine functional genomics and computational modeling into a novel integrative systems approach, based on a probabilistic modeling technique (Bayesian state-space models), to identify key components of the regulatory networks involved in cell physiology. We aim to learn networks integrating transcriptional data with the production of proteins and metabolites with well-defined biological activity.

State-space models are a simple class of probabilistic graphical model used for time series analysis. We are investigating the use of these models for the inference of genetic regulatory networks from high-throughput microarray, proteomics and metabolomics data. In the context of genetic regulatory networks, the hidden states of a state-space model can represent unmeasured factors, such as genes that have not been included in the microarray, levels of regulatory proteins, and the effects of mRNA and protein degradation. We have used state space models to reverse engineer transcriptional networks from highly replicated gene expression profiling time series data obtained from a well-established biological model of T cell activation. The resulting networks reflect many of the dynamics of an activated T cell and provide a methodology for the development of rational and experimentally testable hypotheses. In particular, they reveal the integrated activation of cytokines, proliferation, and adhesion following activation and place JunB and JunD at the center of the mechanisms that control apoptosis and proliferation (Rangel et al., 2001; Rangel et al., 2004a,b; Beal et al., 2005).

In collaboration with the University of Birmingham, UK, we aim to develop a computational framework to reconstruct transcriptional and metabolic networks representative of the response of *E. coli* to acid stress. This experimental system is extremely suitable for understanding the physiology of a bacterial pathogen and, importantly, is amenable to rapid experimental manipulation. The proposed project therefore makes use of the genomic resources, the experimental tools and the biological knowledge available for this well-established model system to address a problem of great biological interest. This project is funded by the NSF and the Biotechnology and Biological Sciences Research Council (UK).

In collaboration with the Cardiovascular Research Center at the University of Hawaii we propose to apply these algorithms to identify new and functionally relevant gene regulatory networks important to the pathobiology of aging human skin, using a biological model of solar elastosis represented as primary human skin fibroblasts. We expect that this research will lead to new and exciting insights into the pathobiology of dermal aging. Moreover, this integrative, interdisciplinary approach to validating methods to reverse engineer gene regulatory networks will establish a paradigm for studying a variety of different, but similarly complex and multifactorial clinical phenotypes.

Analysis of Microarray and Proteomic Data

The development of microarray and mass spectrometry proteomic technologies has facilitated the growth of genomic and proteomic studies in clinical trials and epidemiology. For instance, in cancer clinical trials a key aim is to identify genomic and proteomic factors that are prognostic for survival or relapse-free survival and which predict those patients who respond to treatment. There is therefore a need to develop the current methodology in this area to ensure that the data being produced within genomic and proteomic studies are

appropriately analyzed. The increasing use of gene and protein expression profiles in these types of study requires computational methods of high accuracy for solving clustering, feature selection and classification problems with these data.

In clinical practice, pathological phenotypes are often labeled with ordinal scales rather than binary, e.g. the Gleason grading system for tumor cell differentiation. However, in the literature of microarray analysis, these ordinal labels have been rarely treated in a principled way. We have recently developed a gene selection algorithm based on Gaussian processes to discover consistent gene expression patterns associated with ordinal clinical phenotypes (Chu et al., 2005). The technique of automatic relevance determination is applied to represent the significance level of the genes in a Bayesian inference framework. The usefulness of the proposed algorithm for ordinal labels has been demonstrated by the gene expression signature associated with the Gleason score for prostate cancer data. Our results demonstrate how multi-gene markers that may be initially developed with a diagnostic or prognostic application in mind are also useful as an investigative tool to reveal associations between specific molecular and cellular events and features of tumor physiology. Our algorithm can also be applied to microarray data with binary labels with results comparable to other methods in the literature.

The use of clustering methods has rapidly become one of the standard computational approaches to understanding microarray gene expression data. In clustering, the patterns of expression of different genes across time, treatments, and tissues are grouped into distinct clusters (perhaps organized hierarchically), in which genes in the same cluster are assumed to be potentially functionally related or to be influenced by a common upstream factor. Such cluster structure can also be used to aid the elucidation of regulatory networks. We are employing a novel Bayesian clustering method (infinite Gaussian mixture models) to model uncertainty in the clustering of gene or protein expression profiles. With this method there is no need to make arbitrary choices about how many clusters there are in the data; nevertheless, after modeling one can ask questions such as; how probable is it that two genes belong to the same cluster?

In related work we have developed a Bayesian approach to identify protein complexes and their constituents from high-throughput protein-protein interaction screens (Chu et al., 2005). An advantage of this model is that it again places no prior constraints on the number of complexes and automatically infers the number of significant complexes from the data. Validation results using affinity purification/mass spectrometry experimental data from yeast RNA-processing complexes indicate that our method is capable of partitioning the data in a biologically meaningful way.

Protein Fold Prediction and Remote Homolog Detection in Genomic Sequence Data

A major obstacle to the exploitation of the large volume of genome sequence data is the functional characterization of the gene products. A large proportion, typically 30-40% of the predicted protein coding regions of most organisms' genomes code for proteins of unknown function. For many organisms we therefore do not have a complete 'parts list', which is usually considered to be a prerequisite for a systems biology approach.

Functional annotation is normally inherited from database matches to similar sequences for which the function is known. New algorithms that make use of the information contained within alignments of multiple sequences are very effective at identifying distant sequence relationships. But even using sensitive sequence similarity detection methods, a significant proportion of gene products cannot be reliably assigned function. Recently, large-scale protein structure determination projects have got underway. These initiatives are variously referred to as 'structural genomics' or 'structural proteomics'. Since protein three-dimensional structure is more conserved than sequence, these initiatives also open up the possibility of biochemical or biophysical functional characterization via structure.

In collaboration with University College London, UK, we are developing novel advanced machine learning techniques for protein fold and remote homolog prediction (Raval et al, 2002; Chu et al., 2004,2005). Our current efforts focus on the development of a probabilistic model (a segmental semi-Markov model), which incorporates multiple sequence alignment profiles. Extensive benchmarking indicates that our model is competitive with other contemporary methods for protein secondary structure prediction. Distal interaction

information can also be incorporated into this framework, which has the potential to carry out inference on contact maps, and we are currently extending the approach towards full tertiary structure prediction. We have developed a new atomic-resolution model and Metropolis Monte Carlo sampling procedure with the state-of-the-art efficiency (Podtelezhnikov and Wild, 2005). This method is the first to our knowledge to allow exhaustive and interpretable sampling of polyalanine conformation, characterized by kinks and bulges between helices and stabilized by double hydrogen bonds; features which are all observed in real protein structures. This basic polyalanine model is a foundation for the extension of our methods to improve *ab-intio* protein structure prediction.

These techniques for protein fold prediction and remote homology recognition may be applied to the structural and functional characterization of proteins of unknown function in microbial genomes of clinical interest, with a view to the identification of novel therapeutic targets. They may also be used to assist in metabolic pathway reconstruction, by matching predictions of protein folds and functions to the list of biochemical functions which should exist to make the metabolic pathway model complete and consistent, but for which no gene has been assigned.

A second project involves a novel approach to the problem of automatically clustering protein sequences and discovering protein families, subfamilies etc. using a Bayesian clustering method (Dubey et al., 2004). Unlike previous approaches, this method allows the data itself to dictate how many clusters are required to model it, and provides a measure of the probability that two proteins belong to the same cluster. The consistency of the clusters obtained indicates that our method is producing biologically meaningful results, which provide a very good indication of the underlying families and subfamilies. Applications to the clustering of the important G-protein coupled receptor sequence family (which constitutes around 50% of existing drug targets) lead to some novel functional classifications for orphan receptors.

These projects are funded through a collaborative NIH award to UCSD, KGI and the Burnham Institute to develop tools and data resources in support of structural genomics.

Statement of Teaching Philosophy and Interests:

At the Keck Graduate Institute of Applied Life Sciences I have played a leading role in developing, coordinating and teaching Master's level courses in Computational Biology and Bioinformatics as part of an innovative, interdisciplinary Master of Bioscience degree. All students in this two-year program, with diverse undergraduate backgrounds, take common science core courses in the first year and more specialist courses in the second year. First year courses are currently organized in a series of intensive 3-day workshops and explore a particular area or topic from the perspective of biological systems, computational biology and bioengineering. I have taught first year workshops on bioinformatics databases, gene finding and protein bioinformatics, as well as more specialist second year courses in Computational Molecular Biology and Perl for Bioinformatics. A joint Ph.D. program in Computational and Systems Biology has also recently been initiated with the School of Mathematical Sciences at Claremont Graduate University.

I would plan to take a leading role in the development of multi-disciplinary undergraduate, postgraduate and professional development education in computational and systems biology. My proposed research will provide valuable interdisciplinary training opportunities in computational and systems biology for graduate students and postdoctoral fellows. In the post-genomic era new approaches to research training are needed for studying problems in genomics and molecular biology. The traditional one-person or student-advisor study is unlikely to be successful. Instead, a multidisciplinary approach must be devised, with mathematical and physical scientists involved with experimentalists to understand the motivations for the experiments and the limitations of the studies, and vice-versa. Post-doctoral and graduate students from the biological, physical and mathematical sciences must learn the skills of cross-disciplinary communication and collaboration. My research program will be specifically structured to provide practical and continuing training in this process.

The flow of quantitative and computational approaches to biology into undergraduate curricula, in particular, is essential if we are ever to succeed in the training of a new generation of truly quantitative biologists and biologically literature physical and mathematical scientists and engineers.