Probabilistic graphical models for computational biomedicine

Y. Moreau†, P. Antal, G. Fannes, B. De Moor

Department of Electrical Engineering ESAT-SCD (SISTA)
Katholieke Universiteit Leuven
Leuven, Belgium

† Correspondence to
Yves Moreau
Department of Electrical Engineering ESAT-SCD (SISTA)
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10
B-3001 Leuven
Belgium

Tel: +32/16/32.18.06
Fax: +32/16/32.19.70
Email: yves.moreau@esat.kuleuven.ac.be

SUMMARY

Background: As genomics becomes increasingly relevant to medicine, medical informatics and bioinformatics are gradually converging into a larger field that we call computational biomedicine.

Objectives: Developing a computational framework that is common to the different disciplines that compose computational biomedicine will be a major enabler of the further development and integration of this research domain.

Methods: Probabilistic graphical models such as Hidden Markov Models, belief networks, and missing-data models together with computational methods such as dynamic programming, Expectation-Maximization, data-augmentation Gibbs sampling, and the Metropolis-Hastings algorithm provide the tools for an integrated probabilistic approach to computational biomedicine.

Results and Conclusions: We show how graphical models have already found a broad application in the different fields that compose computational biomedicine. We also indicate several challenges that lie at the interface between medical informatics, statistical genomics, and bioinformatics. As a conclusion we assert that probabilistic graphical models should be a foundation in the curriculum of students of computationally intensive approaches to biology and medicine. From such a foundation, students could then build towards specific computational methods in medical informatics, medical image analysis, statistical genetics, or bioinformatics while keeping the communication open between these areas.

KEYWORDS

Probabilistic graphical models, belief networks, Expectation-Maximization, Gibbs sampling, medical informatics, statistical genetics, bioinformatics, computational biomedicine.

1. TOWARDS COMPUTATIONAL BIOMEDICINE

In medicine, the increasing prevalence of computerized information (medical imaging, electronic patient records, automation of clinical studies) considerably enhances the further progress of medicine as a data-driven evidence-based science, alongside its empirical tradition. As a result, medicine is developing tighter and tighter links to engineering, computer science, and statistics. In biology, faced to the flood of data generated by high-throughput genomics (Human Genome Project, *Arabidopsis* Genome Initiative, microarrays, Single Nucleotide Polymorphism Initiative, and so on), biologists have a pressing need for support, guidance, and collaboration for the analysis of their data. The importance of data management and analysis cannot be underestimated, as it has become a main bottleneck in molecular biology (which itself is a driving force of the pharmaceutical and biotechnological sectors).

Information technology provides a practical platform for a better integration of the different biological and medical disciplines, both for practice and research. As a result, we witness the convergence of the many disciplines relating to the application of computation and information technology to biology and medicine, such as (together with some examples):

- Medical information systems (electronic patient records)
- Biostatistics (design and analysis of clinical studies and clinical trials)
- Medical decision-support systems (diagnosis assistance and critiquing)
- Biomedical image analysis (radiography, nuclear magnetic resonance)
- Biomedical signal processing (electroencephalography, electrocardiography, and also as an essential initial step for image analysis)
- Biomedical systems and control (intelligent prostheses, intelligent drug delivery devices)
- Statistical genetics and epidemiology (gene mapping, single nucleotide polymorphism analysis)
- Computational structural biology (prediction of protein structure from sequence)
- Biological databases and information technology (gene and protein databases)
- Bioinformatics and computational biology (statistical data analysis strategies for molecular biology and *in silico* biology)

We call Computational Biomedicine the general discipline resulting from this convergence. This evolution is a long-term trend that will continue over several decades. This convergence is actually happening by bringing together elements from medical information systems, biostatistics, medical decision support, biological information technology, and bioinformatics for a series of medical and biological applications. Most importantly, we argue here that integration of these different

disciplines is not limited to information technology but that integration will also happen at the level of data analysis thanks to the use of probabilistic graphical models. This integration will enable us to tackle problems that are currently out of reach of each research area on its own. Fig. 1 illustrates a number of research areas that are closely connected and can be integrated within computational biomedicine; it also presents a number of applications within each area where graphical models are already heavily used. As an additional remark, many applications of graphical models are also found in image processing [1] but we do not discuss them here as they fall outside our field of expertise.

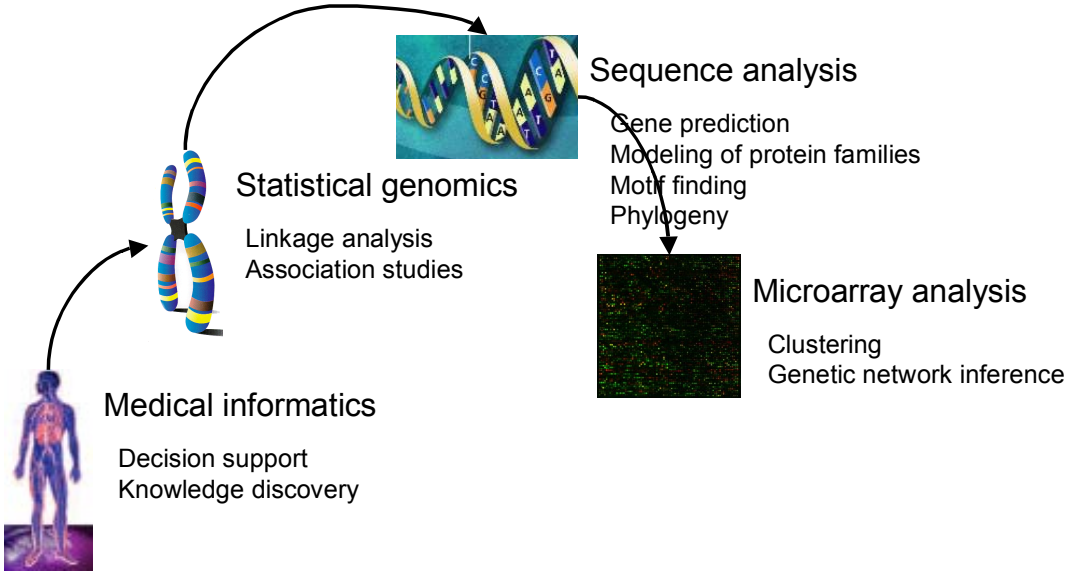

Fig. 1 This drawing illustrates the probabilistic framework for computational biomedicine. For a better understanding of the mechanisms underlying complex pathologies, we must integrate clinical, genetic, and molecular biological knowledge. At the level of data analysis, this means for example linking medical informatics, statistical genomics, sequence analysis, and microarray analysis.

2. PROBABILISTIC GRAPHICAL MODELS

In bioinformatics, probabilistic graphical models have emerged as a dominant approach for data analysis [2]. By probabilistic models, we mean here models that express the probability of some observations given a set of model parameters (i.e., the likelihood). Such models are graphical when this probability can be broken down into the combination of several elementary contributions and the probability can then be represented as a graph. Examples of probabilistic graphical models (or graphical models for short) are Hidden Markov Models (HMMs) (for example, for the modeling of protein families) and belief networks (for the reconstruction of gene networks from expression data). Once the probabilistic model has been set up, the goal is to find

good sets of model parameters matching the observed data. This goal can be achieved by maximum likelihood or maximum a posteriori estimation or by Bayesian inference. In Bayesian inference, we use the data to update a prior probability distribution over the parameters into a posterior probability distribution over the parameters given the data. While this approach is computationally intensive and has only recently become really practical, Baldi and Brunak [2] have convincingly argued that this Bayesian framework offers distinctive advantages, such as a systematic way of "incorporating prior knowledge and constraints into the modeling process" and such as the fact that probability distributions over parameters or observations are more informative than optimal point estimates. After the modeling criterion has been chosen, a variety of algorithms are available for estimating the model, such as gradient descent, Expectation-Maximization (EM), or Markov Chain Monte Carlo (MCMC) methods (Gibbs sampling, the Metropolis-Hastings algorithm, or simulated annealing).

The application of graphical models in bioinformatics is extremely broad. For example, DNA, RNA, and protein sequences lend themselves to simple probabilistic modeling thanks to their sequential structure and their discrete alphabet. In fact, and this is essential to our argument, probabilistic graphical models are not limited to sequence analysis. In medical informatics, belief networks provide a powerful tool for decision support in diagnosis. Another domain where probabilistic graphical models play an important role is statistical genomics. The goal here is to use patterns of genetic inheritance to determine relationships between genes or genetic loci, or relationships between genes and traits or diseases. One important application is the identification of disease-causing genes (which means genes for which some variants contribute to a disease) from affected families (linkage analysis) or populations (association studies). Similarly, graphical models are powerful tools for phylogeny (which is the reconstruction of the tree of evolution based on genomic sequences) thanks to the graphical description of evolutionary trees and of DNA and protein sequences. Furthermore, the patterns of expression of genes and proteins can be efficiently analyzed with graphical models for clustering and with belief networks.

2.1 Belief networks for ovarian tumors diagnosis

To make our discussion of graphical models more concrete, we briefly present a belief network that we have used to assess ovarian tumors using clinical information and ultrasonography [3,4]. The preoperative discrimination between malignant and benign tumors is a crucial issue in gynecology. The International Ovarian Tumor Analysis (IOTA) consortium [5], which is led by the University Hospitals of Leuven, is collecting the world's largest database of ultrasonographic case reports from patients with ovarian tumors (about 1000 cases per year) and aims at developing predictive models based on statistics and artificial intelligence for the preoperative assessment of such tumors. We present the belief network resulting from this study in Fig. 2.

A belief network (also called Bayesian network) represents a joint probability distribution over a set of variables. The model consists of a qualitative part (a directed

graph) and quantitative parts (dependency models). Directed graphical models are not allowed to have directed cycles and have a complicated notion of independence, which takes into account the directionality of the edges. For a particular domain, the vertices of the graph represent the domain variables and the directed edges describe the probabilistic dependency-independency relations among the variables. There is a dependency model for every vertex (i.e., for the corresponding variable) to describe its probabilistic dependency on the parents (i.e., on the corresponding variables). If the variables are discrete, a common dependency model is the table model, which contains the conditional distribution of the child variable conditioned on its parents.

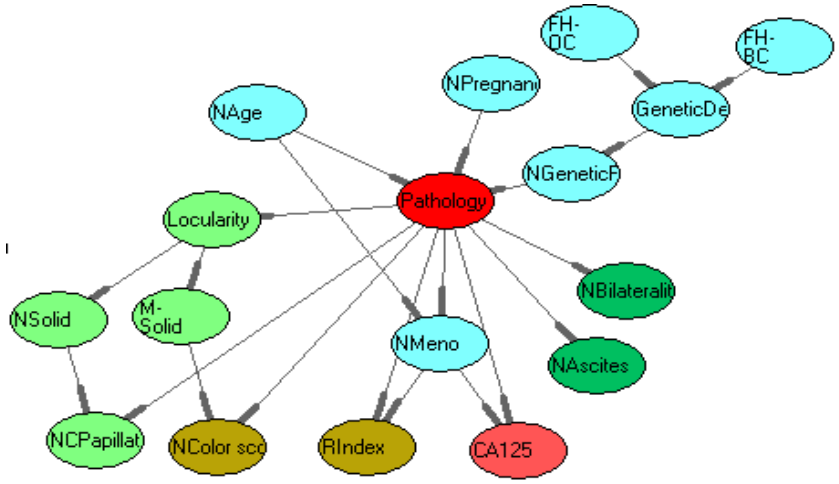

Fig. 2 This belief network represents the joint probability distribution of the measurements in the record of a patient with an ovarian tumor. Nodes represent the variables, such as age, pathology (benign vs. malignant), and CA125 serum level. Nodes are ordered so that each variable has a set of predecessors and a set of successors. Edges represent the probabilistic conditional dependency between variables. For example, the probability of the CA125 level being low, medium, or high given the menopausal status and pathology is independent of all its predecessors. Edges are quantified by an elementary probabilistic model, such as a probability table. For example, the presence or absence of a genetic defect (GeneticD = 0 or 1) is a probability value for each of the configurations of the family history of ovarian cancer (FH-OC = 0 or 1) and family history of breast cancer (FH-BC = 0 or 1).

We have developed such belief networks alongside more classical techniques (such as logistic regression and neural networks) [6]. These belief networks can be used for prediction (in our case the prediction of malignancy in ovarian tumors) by deriving the probability of the pathology variable given the observations from the joint probability distribution of all the variables. Important advantages of belief networks over other methods (such as neural networks) are the easy handling of missing variables and the possibility to incorporate prior knowledge in the model (in the form of the model structure or prior parameterizations of the dependency models). As far as classification performance is concerned, belief networks delivered performances similar to that of other methods [3,4].

The development of such models is an extensive process of interaction with the medical expert. Also, belief networks are computationally intensive. Moreover, we wanted to be able to perform Bayesian inference on belief networks – a facility that is

not readily available. For these reasons, we developed a software environment for the easy development of belief networks. A screenshot of this environment is presented in Fig. 3.

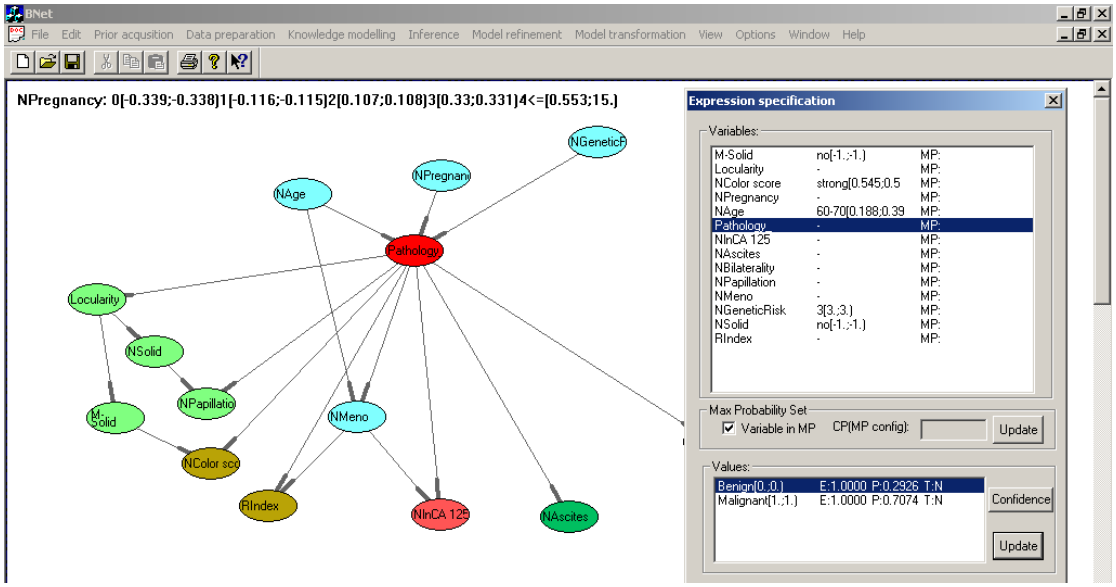

Fig. 3 Screenshot of the Bayesian network tool that was developed for the diagnosis of ovarian tumor malignancy. Updating the value of any variable changes the probability of the whole configuration.

3. COMPUTATIONAL METHODS

Several methods are available to select good sets of parameters θ for a graphical model M given some data set D. We review briefly the main criteria and algorithms for such model estimation.

3.1 Maximum likelihood and maximum a posteriori estimation and Bayesian inference

Usually the graphical model let us express the likelihood $P(D|\theta,M)$ (or some closely related quantity) easily. A reasonable option is then to follow the maximum likelihood principle and choose the set of parameters that maximizes this likelihood:

$$\theta^{ML} = \underset{\theta}{\operatorname{argmax}}\, P(D \mid \theta, M).$$

Another effective strategy is to select the maximum a posteriori parameters:

$$\theta^{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(\theta \mid D, M).$$

This approach focuses more directly on optimizing the set of parameters of the data. However, the graphical model describes the likelihood $P(D|\theta,M)$ while we consider here the posterior probability $P(\theta|D,M)$. Thanks to Bayes' rule

$$P(\theta \mid D, M) = P(D \mid \theta, M)P(\theta \mid M) / P(D \mid M),$$

we can go from the likelihood to the posterior. Additionally, the prior $P(\theta|M)$ let us introduce the prior knowledge we may have about the problem solution. If we further take into account that the data prior $P(D|M)$ is independent of the parameters θ we want to estimate, maximum a posteriori estimation then amounts to

$$\theta^{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(D \mid \theta, M)P(\theta \mid M).$$

Yet, the likelihood function contains much more information about the data than just the maximum-likelihood point estimate. In fact, the posterior distribution provides a more accurate representation of which parameter values are good candidates to describe our data. For example, if the posterior is multimodal, the modes provide very different models that describe the data well. Also, we can construct confidence intervals for the parameters based on this distribution while we do not get this information from an optimal point estimate. Thus it can be advantageous to work with the full probability distribution instead of limiting ourselves to a point estimate. This is the approach taken in Bayesian inference:

$$\begin{aligned} P(\theta \mid D, M) &= \frac{P(D \mid \theta, M)P(\theta \mid M)}{P(D \mid M)} \\ &= \frac{P(D \mid \theta, M)P(\theta \mid M)}{\int_{\theta} P(D, \theta \mid M)P(\theta \mid M)d\theta}. \end{aligned}$$

Note that as long as we can solve the integral in the denominator, the posterior distribution can be entirely expressed in term of the likelihood and the prior. Note also that Bayesian inference does not provide as such a model that fits our data best, but rather a description of the fit of each set of parameters to the data. If we then wish to determine a point estimate, we can achieve this easily in a post-processing step.

3.2 Dynamic programming

The previous criteria let us estimate the fit of different parameter sets to the data as long as we can perform the necessary computations and optimizations. Several techniques let us perform the actual computations. The first important technique is dynamic programming [7], which is for example applied in sequence alignment methods and HMM learning. It lets us perform efficiently maximization or compute sums over a huge set of configurations (such as all possible alignments between two DNA or amino-acid sequences or all possible parameter configurations of an HMM). This computation depends on the possibility of decomposing the configuration set into a chain of subconfigurations to which we can apply Bellman's optimality principle [8]. The importance of dynamic programming arises from its high speed and from the guarantee to find the global optimum for the problems to which it is applicable. This technique is also often a module within the more complex techniques we describe further.

3.3 Expectation-Maximization and missing data problems

Often, the likelihood of the data given the parameters is by itself difficult to compute. As an example, we can consider the discovery of motifs in protein sequences [9]. We then have a set of proteins in which a probabilistic motif (an extended but variable amino-acid sequence) is hidden. The parameters we want to estimate are the parameters of the distribution of the amino acids that form the motif. To compute the likelihood of the full sequences, we need to know where these motifs are located; but in practice this information is not available. So we introduce the positions of the motif as missing data m. When augmented with this missing data, the likelihood $P(\theta,m|D)$ is straightforward to compute. EM is then used to find the maximum of this likelihood for both the motif parameters (true parameters) and positions (missing data).

EM [10] is a two-step iterative procedure for obtaining the maximum likelihood parameter estimates for a model of observed data and missing values. It replaces the maximum likelihood estimate by an iterative procedure for a missing data problem:

$$\theta_{i+1}^{\mathrm{EM}} = \arg\max_{\theta} E_{m|D,\theta_i^{\mathrm{EM}}}(\ln P(D,m\mid\theta)).$$

In the expectation step, the expectation of the data and missing values is computed given the current set of model parameters. In the maximization step, the parameters that maximize the likelihood are computed. EM is guaranteed to converge to a local optimum of the likelihood but not to the global maximum.

3.4 Markov Chain Monte Carlo methods

When performing Bayesian inference, it is in some cases possible to describe the posterior distribution analytically. However, for more complex models such as sequence models, it is impossible to handle the probability distributions analytically. The idea behind MCMC methods is to use sampling for optimization. Several methods are available to generate data according to a complex probability distribution. These are methods such as the Metropolis-Hasting algorithm [11] (which is well known as the foundation of the simulated annealing algorithm for global optimization) and Gibbs sampling [12]). If we assume that we can generate samples according to the posterior distribution, we can use these samples to approximate quantities of interest (possibly using Monte Carlo integration). For example, we can approximate a global solution with maximum posterior probability by tracking the sample with the highest posterior probability if we draw enough samples from the posterior distribution.

MCMC methods are methods to sample from any distribution provided it has an appropriate structure. While sampling from an arbitrary one-dimensional probability density can be achieved simply using a uniform random generator on the [0,1] interval and using the cumulative density function, sampling from high-dimensional probability densities is hard. MCMC methods exploit an important property of Markov chains, which is that data generated by a Markov chain will eventually (after

a number of samples going to infinity) be generated according to a fixed probability distribution called the equilibrium distribution of the Markov chain.

A Markov chain is a stochastic process that generates a sequence of samples according to the fact that any value in the sequence depends only on the previous one and is independent of all the earlier ones. This property is called the Markov property. Markov chains have the remarkable property that, starting from an arbitrary initial condition, the distribution of the samples of the Markov chain will converge to an equilibrium distribution called a stationary distribution (if the chain and the stationary distribution satisfy a condition called detailed balance). Thus in practice, after a sufficient number of transient samples (called burn-in period), the Markov chain will draw approximately from this stationary distribution.

3.4.1 Gibbs sampling

Gibbs sampling is a MCMC method that was introduced by Geman and Geman [1] in the context of image restoration. Tanner and Wong [12] introduced its use for data augmentation problems. The idea is to describe a complex probability distribution in terms of a Markov chain built with the simpler marginals of the distribution. Suppose we have a set of variables described by a joint probability. Gibbs sampling will consist in drawing sequentially every variable according to the probability of this variable conditioned on all the other variables held frozen to their current values. In many cases, Gibbs sampling is applied to missing data problems and is called data-augmentation Gibbs sampling.

3.4.2 Metropolis-Hastings

Metropolis-Hastings sampling [13] is another MCMC method. The method relies on randomly walking through parameter space according to a Markov chain satisfying the following conditions. First, the walk must use a symmetric transition proposal. Second, each move of the walk is accepted according to an acceptance mechanism (if the probability of the proposed parameter set is higher than of the previous set, the move is always accepted; if it is lower, the move is accepted with a probability equal to the relative probability of the two configurations). After convergence of the chain, this procedure then generates samples from the probability distribution used to score the transition proposals. Also, Metropolis-Hastings sampling is the foundation of simulated annealing for global optimization [14].

4. SOME APPLICATIONS OF GRAPHICAL MODELS

Now that we have a general idea of what graphical models are and how we can use them to model data, we can have a look at the variety of problems from computational biomedicine that can be addressed by these techniques. This summary overview will give us an impression of how prevalent these methods have become.

For the analysis of data from clinical trials, Bayesian techniques have become important because of the limited amount of data available and therefore the need for a

detailed characterization of uncertainty. WinBUGS [15] is a popular software package for the automatic generation of Gibbs sampler for simple graphical models and it has been used in numerous applications [16]. For the knowledge discovery in medical or clinical data, clustering by Expectation-Maximization estimation of mixture models is an effective approach. This approach is available in AutoClass [17] together with a Bayesian technique for estimating the number of clusters. For decision support in diagnosis, belief networks are a promising technique. Kahn et al. [18] describe the development and validation of MammoNet, a belief network for mammographic diagnosis of breast cancer that integrates patient-history features, physical findings, and mammographic features to determine the probability of malignancy. We have also applied belief networks to the prediction of malignancy in ovarian tumors [3,4].

In statistical genetics, Excoffier et al. [19] introduced an EM algorithm leading to maximum-likelihood estimates of molecular haplotype frequencies (a haplotype is a combination of alleles of closely linked loci that are found in a single chromosome and tend to be inherited together) under the assumption of Hardy-Weinberg proportions. They evaluated their method on simulated data representing both DNA sequences and highly polymorphic loci with different levels of recombination. Long et al. [20] introduced an EM algorithm to obtain allele frequencies, haplotype frequencies, and gametic disequilibrium coefficients for multiple-locus systems. They validated their method on three unlinked dinucleotide repeat loci in Navajo Indians and to three linked HLA loci in Gila River (Pima) Indians. Recently, methods using Gibbs sampling have been use to extend these approaches. Niu et al. [21] proposed a Gibbs sampler algorithm that can accurately and rapidly infer haplotypes for a large number of linked single nucleotide polymorphisms. The algorithm is also robust to the violation of Hardy-Weinberg equilibrium, to the presence of missing data, and to occurrences of recombination hotspots. For linking genetic loci with disease, Liu et al. [22] proposed an MCMC method using a stochastic model describing the dependence structure among several variables characterizing the observed haplotypes and the location of the disease mutation. They validated their method cystic fibrosis and Friedreich ataxia data.

For phylogenetic inference (the reconstruction of evolutionary trees from current day sequences), EM methods are by now a classical approach [23] while MCMC are making their appearance [24]. Recently, Friedman et al. [25] proposed a heuristic structural EM that performs more efficient topology searches over the possible phylogenetic trees, resulting in better solutions in a much shorter time thereby enabling phylogenetic analysis of large protein data sets in the maximum likelihood framework.

For the analysis of DNA and amino-acid sequences, HMMs are a method of choice. HMMs have been used successfully for highly sensitive detection of new members of known protein families [26,27]. HMMs (sometimes in combination with neural networks) are a major approach for the prediction of genomes from raw genomic sequences [28,29]. GENSCAN predictions [28] were directly part of the annotation of the human genome produced by the Human Genome Project [30] and this is maybe the most significant application of graphical models to date. For the

discovery of motifs in protein sequences, MEME (multiple EM for motif elicitation) [9] provides an application of EM to motif discovery using multistart optimization to escape local optima. For the discovery of motifs both in DNA and protein sequences, Gibbs sampling [31,32] is a highly sensitive approach. We showed in [33] that the robustness of the method could be enhanced by the use of a higher-order Markov chain, similarly to what is done in many gene prediction algorithms.

Microarrays [34] are a recent technique that allows the measurement of the activity of several thousands of genes in a single biological experiment. This technique is generating a massive amount of data (several orders of magnitude more than genome sequences) and the analysis of such data has become a major challenge for bioinformatics. Clustering of such data for discovery of new classes among samples or of groups of genes that share their expression pattern can be efficiently achieved by EM clustering [35]. Another major challenge for microarray data analysis is to attempt to reconstruct the network of interaction between genes from expression measurements. Promising initial results have been achieved using belief networks [36].

This list of applications of graphical models is certainly not exhaustive, but we hope that it illustrates clearly that these methods cover the scope of computational biomedicine.

5. INTEGRATIVE USE OF GRAPHICAL MODELS

To understand more efficiently the mechanisms underlying pathologies, medical and biological researchers attempt to integrate clinical, genetic, and molecular biology aspects more tightly. This stronger integration increases the number of clues available to identify important genes and proteins and to understand the cascades governing the relevant biological processes. Although alternatives to graphical models exist for each specific application, graphical models are the only approach for which we can envision a unified conceptual framework that encompasses computational biomedicine as a whole. As probabilistic models are already developed in medical informatics, statistical genomics, and bioinformatics, what is now necessary is (1) to master the methodologies available in these different fields (which few people do), (2) formulate these methodologies in a common language, and (3) identify synergies across these fields and develop new analysis methods that integrate different types of data.

Several tracks are possible to initiate this integration. First of all, the development of algorithms that can handle heterogeneous types of data is an important research direction. For example, Segal et al. [37] proposed the use of probabilistic relational models to discover patterns in heterogeneous data such as microarray data coupled to experiment type, putative binding sites, or functional information. The patterns discovered could reveal context-specific relationships that exist only over a subset of the experiments. Another track is the integration of statistical genomics and molecular cell biology. For example, Janssen and Nap [38] proposed to combine microarray expression profiling with marker-based fingerprinting. The resulting setup allows a

precise control over the genetic background of the experiment while measuring results at the level of thousands of genes. With appropriate analysis tools, such approaches are likely to boost the further unraveling of metabolic, regulatory, and developmental pathways.

Finally, and probably most importantly, a major link between medical informatics and bioinformatics is the fact that clinicians have access to patient information, which is the truly relevant phenotype information in understanding complex pathologies. Extrapolation to the clinical level from studies of the molecular behavior of cell cultures is extremely hazardous. Even extrapolation to humans from studies in animals is often inaccurate. Therefore, clinicians and biologists must collaborate so that clinical information can increase the relevance of biological studies to human health.

6. CONCLUSIONS

The different disciplines relating to the application of computational techniques to biology and medicine are converging into a field that we call computational biomedicine. We argued that this convergence takes place not only at the level of information technology but also at the computational level. Probabilistic graphical models provide a systematic framework to handle prior knowledge and uncertainty and are applicable to any of the disciplines composing computational biomedicine. We introduced graphical models (such as HMMs or belief networks) and the computational techniques necessary to perform data modeling with them. We further demonstrated their broad applicability through multiple applications in biostatistics, statistical genomics, sequence analysis, and expression analysis that we briefly surveyed. Furthermore, this computational framework opens up research directions for the integrative analysis of multiple and heterogeneous data types, which will be essential if we want to solve the puzzles behind complex pathologies.

We conclude by asserting that the theory of probabilistic graphical models lies at the foundation of the application of computational techniques to biology and medicine. It should be as such integrated as a common basis in the curriculum of students entering these domains so as to enhance the later communication among these areas.

ACKNOWLEDGMENTS

Dr. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Dr. Yves Moreau is a post-doctoral researcher of the Fund for Scientific Research - Flanders and an assistant professor of the K.U.Leuven. Geert Fannes is a doctoral fellow of the Fund for Scientific Research - Flanders.

Our research is supported by grants from several funding agencies and sources:

- Research Council KUL: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several Ph.D., post-doc, and fellow grants

- Flemish Government: Fund for Scientific Research Flanders (several Ph.D. and post-doc grants, G.0115.01 (bio-i and microarrays), G.0407.02 (support vector machines), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary), IWT (STWW-Genprom (gene promoter prediction), GBOU-McKnow (Knowledge management algorithms), several Ph.D. grants)
- Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006): Dynamical Systems and Control: Computation, Identification and Modelling)

REFERENCES

1. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984;6(6):721-41.

2. Baldi P, Brunak S. Bioinformatics: the machine learning approach. Cambridge, MA: MIT Press; 1998.

3. Antal P, Verrelst H, Timmerman D, Moreau Y, Van Huffel S, De Moor B, Vergote I. Bayesian networks in ovarian cancer diagnosis: potentials and limitations. In: Computer-Based Medical Systems. Los Alamitos, CA: IEEE Computer Society; 2000. p. 103-8.

4. Antal P, Fannes G, De Moor B, Vandewalle J, Moreau Y, Timmerman D. Extended Bayesian regression models: a symbiotic application of belief networks and multilayer perceptrons for the classification of ovarian tumors. In: S. Quaglini, P. Barahona, S. Andreassen, editors. Lecture Notes in Artificial Intelligence 2101. Berlin: Springer; 2001. p. 177-87.

5. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet Gynecol. 2000 Oct;16(5):500-5.

6. Timmerman D, Verrelst H, Bourne TH, De Moor B, Collins WP, Vergote I, Vandewalle J. Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. Ultrasound Obstet Gynecol. 1999 Jan;13(1):17-25.

7. Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, UK: Cambridge University Press; 1998.

8. Bellman R. The theory of dynamic programming. Bull Am Math Soc. 1954;60:503-15.

9. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28-36.

10. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via EM algorithm. J R Stat Soc Ser B. 1977;39:138.

11. Neal RM. Bayesian Learning for Neural Networks. Lecture Notes in Statistics 118. New York: Springer-Verlag; 1996.

12. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. J Am Stat Assoc. 1987;82:528-50.

13. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. Equation of state calculations by fast computing machines. J Chem Phys. 1953;21(6): 1087-92.

14. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science. 1983;220:671-80.

15. Spiegelhalter DJ, Thomas A, Best NG. Computation on Bayesian graphical models. In: Bayesian Statistics 5. Oxford, UK: Oxford University Press; 1996. p. 407-25.

16. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo Methods in Practice. CRC Press; 1996.

17. Cheeseman P, Stutz J. Bayesian Classification (AutoClass): Theory and Results. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press/MIT Press; 1996. p. 153-180.

18. Kahn CE Jr, Roberts LM, Shaffer KA, Haddawy P. Construction of a Bayesian network for mammographic diagnosis of breast cancer. Comput Biol Med. 1997 Jan;27(1):19-29.

19. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995 Sep;12(5):921-7.

20. Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 1995 Mar;56(3):799-810.

21. Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet. 2002 Jan;70(1):157-69.

22. Liu JS, Sabatti C, Teng J, Keats BJ, Risch N. Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res. 2001 Oct;11(10):1716-24.

23. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368-76.

24. Mau B, Newton MA, Larget B. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics. 1999 Mar;55(1):1-12.

25. Friedman N, Ninio M, Pe'er I, Pupko T. A structural em algorithm for phylogenetic inference. J Comput Biol. 2002;9(2):331-53.

26. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997 Jul;28(3):405-20.

27. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol. 1998 Dec 11;284(4):1201-10.

28. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997 Apr 25;268(1):78-94.

29. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 1998 Feb 15;26(4):1107-15.

30. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921.

31. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993 Oct 8;262(5131):208-14.

32. Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci. 1995 Aug;4(8):1618-32.

33. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouze,P., and Moreau,Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17, 1113-1122.

34. Brown,P.O. and Botstein,D. (1999). Exploring the new world of the genome with DNA microarrays. Nat. Genet. 21, 33-37.

35. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics. 2001 Oct;17(10):977-87.

36. Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. Bioinformatics. 2001;17 Suppl 1:S215-24.

37. Segal E, Taskar B, Gasch A, Friedman N, Koller D. Rich probabilistic models for gene expression. Bioinformatics. 2001;17 Suppl 1:S243-52.

38. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet. 2001 Jul;17(7):388-91.