Objective Bayesian Nets for Integrating Cancer Knowledge: a Systems Biology Approach

Sylvia Nagl[1], Matt Williams[2], Nadjet El-Mehidi[1], Vivek Patkar[2], and Jon Williamson[3]

[1]Cancer Systems Biology & Biomedical Informatics, Royal Free & University College Medical School, Rowland Hill Street, London NW3 2PF, [2]Advanced Computation Laboratory, Cancer Research UK, 44 Lincoln's Inn Fields, London WC2A 3PX, [3]Philosophy, Cornwallis Building, University of Kent, Canterbury CT2 7NF, United Kingdom
s.nagl@ucl.ac.uk

Abstract. According to objective Bayesianism, an agent's degrees of belief should be determined by a probability function, out of all those that satisfy constraints imposed by background knowledge, that maximises entropy. A Bayesian net offers a way of efficiently representing a probability function and efficiently drawing inferences from that function. An objective Bayesian net is a Bayesian net representation of the maximum entropy probability function. In this paper we apply the machinery of objective Bayesian nets to breast cancer prognosis. Background knowledge is diverse and comes from several different sources: a database of clinical data, a database of molecular data, and quantitative data from the literature. We show how an objective Bayesian net can be constructed from this background knowledge and how it can be applied to yield prognoses and aid translation of clinical knowledge to genomics research.

Keywords: Cancer systems biology, objective Bayesianism, Bayesian networks, breast cancer, karyotype evolution, prognosis, clinico-genomic predictive models

1 Introduction

Cancer treatment decisions should be based on all available knowledge. But this knowledge is complex and varied: it includes not only the patient's symptoms and expert knowledge of the relevant causal processes, but also clinical databases relating to past patients, databases of observations made at the molecular level, and knowledge encapsulated in scientific papers and medical informatics systems. Objective Bayesian nets offer a principled path to knowledge integration. This is important from the systems biology perspective, which needs to integrate data that concern different levels of analysis, and is also important from the point of view of medical informatics.

Although many risk factors for different cancer types are known, most of the genetic background and molecular mechanisms still remain to be elucidated. Systems biology seeks to address the complexity of cancer by drawing on a conceptual framework based on the current understanding of complex adaptive systems [1; and

refs therein]. Progression from normal tissue to malignancy is associated with the evolution of neoplastic cell lineages with multiple genomic lesions (abnormal karyotypes). In cancer, dynamic structural changes of the genome occur at dramatically increased frequency and tumour cell – microenvironment interactions drive the selection process. Mutations, affecting individual oncogenes and tumour suppressors, and extensive chromosomal alterations are selected and give rise to rapidly proliferating cell variants that may acquire an aggressive phenotype and resistance to drug exposure. Genomic rearrangements, such as amplifications and deletions, can affect several megabases of DNA and include a large number of genes. Large-scale changes in genome content can be advantageous to the cancer cell by simultaneous activation of oncogenes, elimination of tumour suppressors, and generation of (bystander) gene copy number changes which can have profound effects on the cancer phenotype.

Genomes are dynamic molecular systems and selection acts on cancer karyotypes as integrated wholes, not just on individual oncogenes or tumour suppressors. Given the irreversible nature of evolutionary processes, the randomness of mutations and rearrangements relative to those processes, and the modularity and redundancy of complex systems, there potentially exists *a multitude of ways to 'solve' the problems of achieving a survival advantage in cancer cells*. Since each patient's cancer cells evolve through an independent set of genomic lesions and selective environments, the resulting heterogeneity of cell populations within the same tumour, and of tumours from different patients, is a fundamental reason for differences in survival and treatment response.

Since the discovery of oncogenes and tumour suppressors, a reductionist focus on single, or a small number of, mutations has resulted in cancer being conceptualized as a 'genetic' disease. More recently, cancer has been recast as a 'genomic' or 'systems' disease. Here, we apply a systems framework to karyotype evolution and employ Bayesian networks (BN) to generate models of non-independent rearrangements at chromosomal locations from comparative genome hybridisation data. Furthermore, we present a method for integration of genomic BN models with BNs learnt from clinical data.

The method enables the construction of multi-scale nets from BNs learnt from independent datasets, with each of the BNs representing the joint probability distributions of parameter values obtained from different levels of the biological hierarchy, i. e., the genomic and tumour level in the application presented here (together with treatment and outcome data). BN integration allows one to capture 'more of the physiological system' and to study dependency relationships across scales.

1.1 Breast Cancer

The mainstay of treatment for breast cancer remains surgery and radiotherapy, with hormonal and chemotherapeutic agents often used to treat presumed micro-metastatic disease. Surgery involves removing tumour from the breast, as well as taking out a sample from the axillary lymph nodes. These lymph nodes often act as first destination for spreading cancer cells, and their removal not only removes any spread

that may have occurred, but also helps to predict the degree of distant metastatic spread. The two main aims of treatment are to prevent breast cancer recurrence and to prevent breast cancer related death. Therapy planning seeks to match treatment with the risk of recurrence and/or death. Thus those at high risk should be treated aggressively while those at low risk could be treated less aggressively.

Examination of the primary tumour and lymph nodes lets us define certain characteristics of the disease that make local recurrence and death more likely. These characteristics are primarily the grade of the tumour, (which represents the degree of abnormality displayed by the cells, scored 1-3), the size of the tumour (as its maximum diameter, in mm) and the number of involved nodes. There are also newer tests for the presence or absence of certain proteins on the cell surface that may predict tumour behaviour or response to certain drugs.

These prognostic characteristics are currently modelled using statistical techniques to provide an estimate of the probability of survival and local recurrence. Two commonly used systems are the Nottingham Prognostic Index (NPI), which uses data from large UK studies, and results derived from the American Surveillance, Epidemiology and End Results (SEER) database [2]. Both techniques rely on multivariate analyses of large volumes of data (based on over 3 million people for SEER) to calculate prognostic formulae.

These tools, and others like them, are effective at providing estimates of risk of death and local recurrence. However, they have two major weaknesses. Whilst effective, they lack explanatory power in a human-readable form. Extra knowledge that has not been captured by the statistical analysis (such as the presence and impact of other co-existing conditions) cannot be easily incorporated. Secondly, knowledge that post-dates the formation of the formulae (such as the discovery of Her-2neu, a cell-surface protein that is a marker for more aggressive disease) is very difficult to incorporate. Therefore, while they excel at providing an accurate assessment of population-based risk, they have weaknesses in the individualisation of that risk.

2 Assembling an Obnet from Molecular and Clinical Sources

Objective BNs allow one to integrate various knowledge sources [3-5]. According to *objective Bayesianism*, an agent's beliefs should be representative of her knowledge in the sense that they should commit to everything that is warranted by her knowledge but nothing that is unwarranted. Formally this is explicated via the *maximum entropy principle*: an agent's rational degrees of belief are represented by the probability function that has maximum entropy, from all those that satisfy constraints imposed by background knowledge. A BN is a representation of a probability function: a directed acyclic graph together with the probability distribution of each variable conditional on its parents in the graph determines a probability function on the whole domain, as long as the Markov condition - which says that each variable is probabilistically independent of its non-descendants conditional on its parents - holds. An *objective BN* or *obnet* is a BN representation of the probability function produced by the maximum entropy principle. Thus an obnet is representative of the knowledge sources that are

used to produce it. The net can be used to calculate prognostic or diagnostic probabilities by entering evidence from individual cases.

The general procedure for constructing an obnet runs as follows. Represent background knowledge as a set of quantitative constraints on a probability function. Next join variables that occur in the same constraint by an edge to form an undirected constraint graph. Transform that undirected graph into a directed acyclic graph; the Markov condition is guaranteed to hold. Finally maximise entropy to determine the probability distribution of each variable conditional on its parents.

In our context, knowledge takes the form of a clinical database [2] and two molecular databases (data from www.progenetix.de). The clinical database determines a probability distribution c over the clinical variables and imposes the constraint $p|C = c$, i.e. the agent's probability function, when restricted to the clinical variables, should match the distribution determined by the clinical dataset. Similarly the first molecular database imposes the constraint $p|M = m$. An additional molecular dataset and a study [6] contain observations that define the probability distribution s of three variables S, two of which occur in the clinical dataset and the other of which occurs in the molecular dataset; it imposes the constraint $p|S = s$.

Given constraints of this form, an obnet on the variables in C and M can be constructed in the following way (Fig. 1). First use standard methods, such as Hugin software, to learn a Bayesian net from the clinical dataset that represents c, subject to the condition that the 'linking' variables (positive LN, HR status) are root variables. Similarly learn a Bayesian net that represents m, ensuring that the 'linked' variable (22q12) is a root of the net. Finally join the linking variables, and add the corresponding conditional probability distribution determined by s.

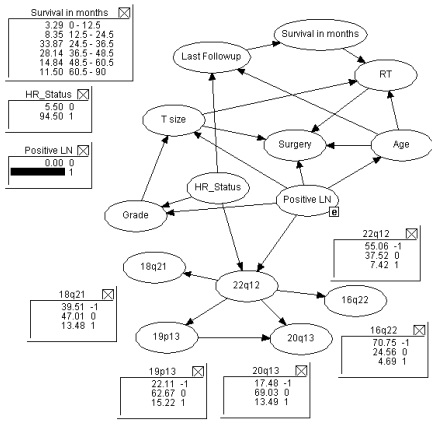

Fig. 1. Integrated obnet. Probability for positive lymph node status is set to 1 (black bar), and the calculated probability distributions for selected nodes are shown (HR status: 0 negative, 1 positive; bands: -1 loss, +1 gain, 0 no rearrangement; RT radiotherapy).

3 Significance and Interpretation

We have presented a general method for merging two different Bayesian networks which model knowledge in different areas. This method was applied to an example application linking knowledge about breast cancer genomics and clinical data. As a result, we have been able to examine the influence of karyotype pattern on clinical parameters (e. g., tumour size, grade, receptor status, likelihood of lymph node involvement) and vice versa (Fig. 1).

Analysis of these relationships may facilitate (i) discovery of genomic markers and signatures, and (ii) translation of clinical data to genomics research and discovery of novel therapeutic targets:

(i) Since hormone receptor status and lymph node involvement are well-known prognostic factors for survival and disease recurrence in patients with breast cancer, the ability to link karyotype patterns to this is clearly of great potential significance. Previous tumour genotyping in breast cancer has already shown the usefulness of genomic rearrangements as prognostic indicators (see, e.g. [7]).

For clinical decision making, this technique may also be useful when applied to integrate karyotype or other molecular data with parameters that cannot be observed in routine clinical practice, but are of clinical significance. An example might be the presence of distant metastasis on PET-CT, an imaging modality that may be present in the research setting but is not widely available in the clinic, but which may have prognostic significance for breast cancer recurrence. The use of such a net would then allow practitioners, where PET-CT is not available, to use genomic data to estimate the likelihood of a positive scan. There are of course, many different possible options for such networks, and it remains an open question as to which will, in clinical terms, prove to be the most useful.

ii) The dependence between 22q12 status and lymph node involvement was followed up by analysis of genes with known function on this chromosomal band. Significantly, KREMEN1 encodes a high-affinity dickkopf homolog 1 (DKK1) transmembrane receptor that functionally cooperates with DKK1 to block wingless (WNT)/beta-catenin signalling, a pathway which promotes cell motility [8]. Loss of 22q12 may therefore contribute to cancer cell migration through loss of the inhibiting KREMEN1 protein. The probability distribution for 22q12 is consistent with this hypothesis (Fig 1). In addition, eight candidate genes which are also associated with cell migration and metastatic potential were identified on the other bands. The complex pattern of dependency relationships between them as revealed by BN modelling, and the potential of the gene products as novel targets for therapy, will be presented in a forthcoming paper.

Large clinical data sets are extremely expensive and difficult to collect. This is particularly true in diseases such as breast cancer, where the risk of recurrence extends up to at least 10 years, and hence requires long-term follow-up for accurate estimation. However, the generation of potential new predictive markers, such as genomic information or cell surface proteins, for exploration is currently a significant area of research. The correlation of such markers with better known clinical markers is (relatively) simple, in that it does not require long-term follow-up, and can be estimated following standard surgical treatment. However, for such information to be

useful, it must be integrated with the existing databases on long-term outcomes, and it is this that we have demonstrated here.

The technique enables progressive integration of BNs learnt from independently conducted studies and diverse data types, e. g., mRNA or proteomic expression, SNP, epigenetic, tissue microarray, and clinical data. New knowledge, and new data types, can be integrated as they become available over time. Application of our integrative BN knowledge discovery method can be envisaged to be valuable in the clinical trials arena which is undergoing profound changes with steadily increasing incorporation of molecular profiling. It is our aim to assess the potential of the technique for integration of different types of clinical trial datasets (with and without molecular data). The BN methods described here are highly complementary to ongoing research initiatives, such as the Cancergrid project (www.cancergrid.org) and caBIG (cabig.nci.nih.gov) which are already addressing pressing informatics requirements that result from these changes.

References

1. Nagl, S.: A Path to Knowledge: from Data to Complex Systems Models of Cancer. In: Nagl, S. (ed.): Cancer Bioinformatics. John Wiley & Sons, London (2006) 3-27
2. Ries, L.A.G., Eisner, M.P., Kosary, C.L., Hankey, B.F., Miller, B.A., Clegg, L., Mariotto, A., Feuer, E.J., Edwards, B.K.: SEER Cancer Statistics Review 1975-2001. National Cancer Institute, 2004
3. Williamson, J.: Maximising Entropy Efficiently. Electronic Transactions in Artificial Intelligence Journal, 6 (2002). www.etaij.org
4. Williamson, J.: Bayesian Nets and Causality: Philosophical and Computational Foundations. Oxford University Press, Oxford (2005)
5. Williamson, J. : Objective Bayesian nets. In: Artemov, S., Barringer, H., d'Avila Garcez, A. S., Lamb, L. C., and Woods, J. (eds.): We Will Show Them! Essays in Honour of Dov Gabbay, Vol. 2. College Publications, London. (2005) 713–730
6. Fridlyand, J., Snijders, A.M., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B.M., Jain, A.N., McLennan, J., Ziegler, J., Chin, K., Devries, S., Feiler, H., Gray, J.W., Waldman, F., Pinkel, D., Albertson, D.G.: Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer 6 (2006) 96
7. Al-Kuraya, K., Schraml, P., Torhorst, J., Tapia, C., Zaharieva, B., Novotny, H., Spichtin, H., Maurer, R., Mirlacher, M., Kochli, O., Zuber, M., Dieterich, H., Mross, F., Wilber, K., Simon, R., Sauter, G.: Prognostic relevance of gene amplifications and coamplifications in breast cancer. Cancer Research 64 (2004) 8534-8540
8. Mao, B., Wu, W., Davidson, G., Marhold, J., Li, M., Mechler, B.M., Delius, H., Hoppe, D., Stannek, P., Walter, C., Glinka, A., Niehrs, C.: Kremen proteins are Dickkopf receptors that regulate Wnt/beta-catenin signalling. Nature 417 (2002) 664-667