PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Bioinformatics tools in predictive ecology: applications to fisheries

Allan Tucker and Daniel Duplisea

Phil. Trans. R. Soc. B 2012 **367**, 279-290
doi: 10.1098/rstb.2011.0184

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 279–290
doi:10.1098/rstb.2011.0184

Research

Bioinformatics tools in predictive ecology: applications to fisheries

Allan Tucker[1,*] **and Daniel Duplisea**[2]

[1]*School of Information Systems, Computing and Maths, Brunel University, Uxbridge, Middlesex UB8 3PH, UK*
[2]*Fisheries and Oceans Canada, Institut Maurice-Lamontagne Mont Joli, Quebec, Canada*

There has been a huge effort in the advancement of analytical techniques for molecular biological data over the past decade. This has led to many novel algorithms that are specialized to deal with data associated with biological phenomena, such as gene expression and protein interactions. In contrast, ecological data analysis has remained focused to some degree on off-the-shelf statistical techniques though this is starting to change with the adoption of state-of-the-art methods, where few assumptions can be made about the data and a more explorative approach is required, for example, through the use of Bayesian networks. In this paper, some novel bioinformatics tools for microarray data are discussed along with their 'crossover potential' with an application to fisheries data. In particular, a focus is made on the development of models that identify functionally equivalent species in different fish communities with the aim of predicting functional collapse.

Keywords: bioinformatics; Bayesian networks; classification; dynamic models; fisheries management

1. INTRODUCTION

Bioinformatics has revolutionized the way we analyse molecular biological data. Owing to the explosion in data collection and storage made available since the dawn of parallel sequencing, there has been a demand for specialist techniques to analyse and model data such as microarray experiments, which measure the expression of thousands of genes simultaneously. The advance of research in fields including machine learning [1], data mining [2] and intelligent data analysis [3,4] has resulted in many novel tools for the analysis of such data. In bioinformatics, techniques such as clustering were initially extremely popular for identifying groups of genes with similar expression profiles [5,6]. This allowed biologists to identify the function of previously unknown genes through 'guilt by association'. It also allowed these groups to be treated as single modules [7,8] in order to reduce the massive number of variables when building models for prediction. Classification of disease outcome [9] has also been very popular with many approaches being developed, including methods to identify relevant biomarkers through feature selection [10]. Modelling time-series microarray data has been useful in understanding the underlying dynamics of microarray time-series, and cell-cycle data have been a popular topic of study [11]. One particular development in these areas is the adoption of graph-based models in the form of genetic regulatory networks (GRNs) [12,13]. These approaches allow biologists to explore the complexities of gene interaction on a large scale and therefore take a *systems* approach to modelling.

In contrast, ecological data analysis has been rather less explorative to date when compared with bioinformatics and systems biology. There are of course exceptions, and in the study of Hochachka *et al.* [14] a discussion of the potential of using data-mining techniques is explored for situations where there is little or no prior knowledge about a system. In this paper, we investigate the cross-over potential of techniques used in bioinformatics, such as feature selection, classification, Bayesian networks (BNs) and in particular an adaptation of an algorithm that we previously developed for exploiting the availability of multiple datasets. This is applied to fisheries data in order to identify species that perform similar functional roles in different fish communities. These equivalent species are used to predict functional collapse in their respective regions through the use of dynamic Bayesian models with latent variables.

In the remainder of this section, BNs are introduced in the context of bioinformatics research, and recent relevant work on specialist bioinformatics techniques that have cross-over potential is discussed. The use of these techniques applied to ecological data is also discussed with a focus on fisheries. In §2, the fisheries data and the 'functional equivalence' algorithm are described. Results in §3 demonstrate how models learned from data in one region can be used to identify and predict the biomass of 'functionally similar' species and as a result, the functional collapse in other regions. Finally, the use of the techniques explored in this paper (namely, BNs for feature selection and classification, the functional equivalence algorithm and dynamic models

* Author for correspondence (allan.tucker@brunel.ac.uk).

One contribution of 16 to a Discussion Meeting Issue 'Predictive ecology: systems approaches'.

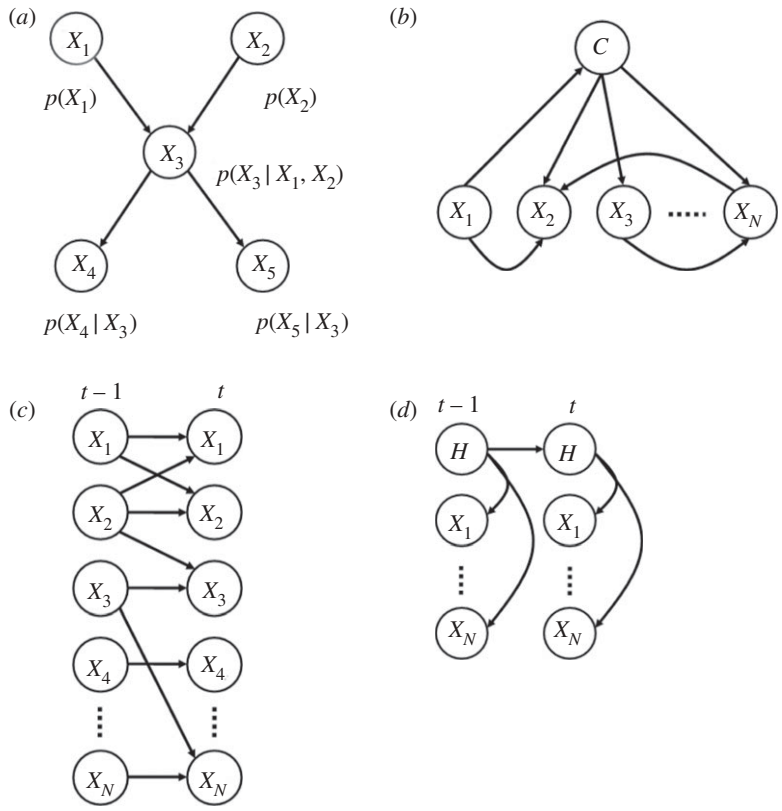

Figure 1. (*a*) A Bayesian network (BN) that encodes a joint distribution using a graphical structure and local conditional distributions. Links between variables represent conditional independences. (*b*) A BN classifier where C denotes a class node to predict. (*c*) A dynamic BN where nodes represent variables at a point in time and (*d*) a hidden Markov model, where H denotes an unmeasured (hidden or latent) variable.

with latent variables) are discussed in §4 in terms of the wider ecological literature.

(a) *Bayesian networks for bioinformatics*

BNs have become a popular method for computational modelling of GRNs from microarray expression data [15–17]. A BN describes the *joint distribution* (which is a way of assigning probabilities to every possible outcome over a set of variables, $X_1 \ldots X_N$) by exploiting conditional independence relationships represented by a directed acyclic graph (DAG). See figure 1*a* for an example of BN with five nodes. Each node in the DAG is characterized by a state which can change depending on the state of other nodes and information about those states propagated through the DAG. This kind of inference facilitates the ability to ask 'what if?' questions of the data by entering evidence (changing a state or confronting the DAG with new data) into the network, applying inference and inspecting the *posterior distribution* (which represents the distributions of the variables given in the observed evidence). For example, one could ask, what is the probability of seeing gene A 'switch on' (through high expression) given that we have observed a low expression in genes B and C?

There are numerous ways to infer both network structure and parameters from data. Constraint-based approaches such as the PC [18] and IC* [19] algorithms both work by applying independence tests between variables and building networks that reflect these independences. However, these do not scale well for high-dimensional datasets and are prone to getting stuck in local minima. Search-and-score methods to infer BNs from data have been used frequently in learning GRNs [15]. These methods involve performing a search through the space of possible networks and scoring each structure. A variety of search strategies can be used [20–23]. BNs are capable of performing many data analysis tasks including feature selection and classification (performed by treating one node as a class node and allowing the structure learning to select relevant features [24] (figure 1*b*)). Modelling time series can be achieved by using an extension of the BN known as the dynamic Bayesian network (DBN) [25,26], where nodes represent variables at particular time slices (figure 1*c*). Closely related to the DBN is the Hidden Markov Model (HMM) which models the dynamics of a dataset through the use of a latent variable [27]. This latent variable is used to infer some underlying state of the series and through an autoregressive link that can capture relationships of a higher order (figure 1*d*).

BNs offer a natural mechanism for incorporating prior knowledge relating to the network structure through informative structure priors [28]. There has been substantial work in using priors to build more robust GRNs. Steele *et al.* [22] use concept profiles learned from abstracts in the biological literature (Medline) to bias BN learning algorithms and found that lesser studied systems generally gain more from updating priors with new data. Imoto *et al.* [29] use energy functions to incorporate prior knowledge sources

including literature-based knowledge extracted from regulatory interactions that are recorded in the Yeast Proteome Database (YPD). In the study of Werhli & Husmeier [30], the approach was extended to multiple sources of prior knowledge, applied to combining protein–protein interactions and pathways from KEGG (Kyoto Encyclopedia of Genes and Genomes) with expression data.

(b) *Consensus and functional models*

Comparing apparently similar multivariate datasets is often problematic owing to differences in collection methods. Such often is the case for microarray data which have methodological and laboratory dependencies [31] and similar issues occur with ecological community data collected for different systems. Though data normalization is the logical solution to such problems, it is neither straightforward nor a complete solution [32,33]. A post-learning aggregation framework called *consensus BNs* was developed for microarray datasets [34] to overcome some of these issues by combining datasets generated by different platforms, research groups and laboratories without requiring normalization. In this framework, learnt models that are generated from each dataset are aggregated, producing a combined model that represents prominent features which occur in all, or a subset of, the individual dataset models. The problem with this approach is the need to pre-select higher quality datasets to prevent the 'dumbing down' of networks from lower quality data resulting in an 'average' network rather than a 'best-of'. A reliable method to identify these higher quality datasets prior to the consensus algorithm was found to be the *predictive accuracy* of models learned from one dataset and tested on other available independent sets [35]. This approach resulted in consensus models that were consistently more parsimonious to biologically validated networks and was extended by Anvar *et al.* [36,37]. It is this idea of exploiting independent datasets that shapes the work in this paper.

In summary, the success of bioinformatics methods such as feature selection, classifiers and HMMs has led to many novel discoveries including the identification of biomarkers, the prediction of disease outcome and GRNs built at a systems level. What is more, the exploitation and integration of multiple data sources allow more robust regulatory mechanisms to be identified and predictions to be made across very different platforms and organisms. We now demonstrate the transfer of some of these methods to ecological data with an application in fisheries.

(c) *Fisheries and ecoinformatics*

In this paper, the focus is on the application of bioinformatics techniques described in §1 to biomass data from Georges Bank (GB in figure 2), the East Scotian Shelf (ESS) and the North Sea (NS) between the years 1960 and 2007. Data are typically noisy and there are similar data quality issues as found with many microarray datasets. There are also multiple studies carried out throughout the world and prior expertise available much similar to bioinformatics datasets. For example, food webs that describe

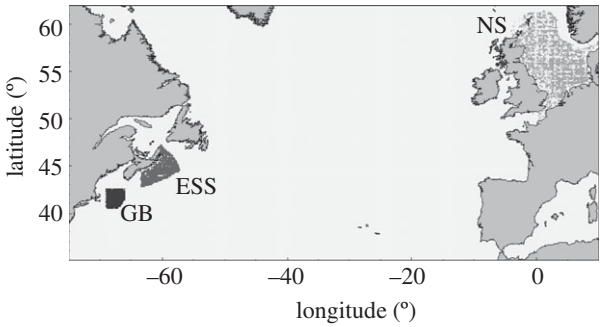

Figure 2. Georges Bank (GB), the East Scotian Shelf (ESS) and the North Sea (NS). The focus of the empirical analysis.

predator–prey and competitor species are available. Some of these are more detailed than others and may include the results of stomach surveys [38], where the diet of specific species can be determined.

The experiments carried out in this paper focus on cod biomass. Some spectacular collapses in fish stocks have occurred in the past 20 years but the most notable is the once largest cod stock in the world, the Northern cod stock off eastern Newfoundland, which experienced a 99 per cent decline in biomass. Cod, unfortunately, is not alone and there are stocks of various species that have been reduced to only a small percentage of stock sizes in recent history. Much of this effect is due to direct mortality on fish through fishing and subsequent indirect effects and weak linkages with other species. Some of these regions may have moved to an 'alternative stable state' or experienced a 'regime shift' and are unlikely to return to a cod-dominated community without some chance event beyond human control [39].

Different species may have similar functional roles within a system depending on the region. For example, one species may act as a predator of another which regulates a population in one location, but another species may perform an almost identical role in another location. If we can model the function of the interaction rather than the species itself, data from different regions can be used to confirm key functional relationships, to generalize over systems and to predict impacts of forces such as fishing and climate change. The approach concerns functional network topology and avoids the necessity of describing the specifics of network nodes. For example, the 'wasp waist' (WW) is a common structure present in many temperate and boreal fish community food webs [40]. The WW functional structure is characterized by few or just one mid-trophic level species preying upon several lower trophic level species, while several high trophic species prey upon the mid-trophic level species. In this way, energy flow from low to high trophic level species is constricted at the mid-trophic level species analogous to a WW. These WW species exert undue influence on aquatic community structure by top-down control of lower trophic levels through predation and bottom-up control of higher trophic levels by restricting energy flow. The WW effect is found in populations in the northwest Atlantic and the northeast Atlantic. This functional structure is identical in the two regions but the species involved are different (in the northwest Atlantic one of

the WW is the capelin and in the northeast one is known to be the sand eel).

This focus on critical sub-systems through the exploration of functionally equivalent species across different populations is a novel approach to fish population modelling. This approach to modelling fish populations will explore functional relationships (such as predator, prey and WW) that are generalizable between different oceanic regions allowing more robust models to be built and predictions to be made about future biomass. There is some research into using BNs for ecological modelling [41] and in particular for modelling fish populations [42,43]. There is also considerable literature on integrating heterogenous data within the data-warehousing community including the environmental data [44], but no exploration of integrating or comparing different variables under a single function as we do with species. In the study of Thrush *et al.* [45], functions are explored by investigating weights in hidden nodes of neural network models. Here, we focus on DBNs with latent variables that can be used in conjunction with human expertise to predict functional collapse in different dynamic systems.

A number of questions are posed based upon fish interaction:

— Can we use bioinformatics-style analysis (in particular, feature selection) to identify species that are relevant to some event such as cod functional collapse?
— Can we model the temporal and dynamic nature of fish interactions?
— Can we identify species in different oceans that perform similar functions, and therefore predict functional collapse in their respective regions?

Techniques such as those described in §1 will be employed to answer these questions within a BN framework. In particular, a novel algorithm—the *functional equivalence*—search is introduced to make inferences between the different geographical systems.

2. MATERIAL AND METHODS

(a) *Data description*

GB fish community data come from the National Marine Fisheries Service autumn multi-species trawl survey from 1963 to 2008. About 80 randomly selected stations were sampled on GB each year and annual averages of biomass of each species were calculated and used in this analysis. About 220 have been caught in the survey but most infrequently and with low statistical power; therefore, analyses were confined to a subset of 39 species filtered from the dataset for which we have confidence in their quantitative estimates of abundance each year. ESS and NS data were collected via a similar methodology as on GB and this resulted in subsets of 34 and 45 species, respectively. The sources of these datasets are outlined in the acknowledgements of this paper.

GB is a relatively small productive fishing bank historically supporting large catches of common groundfish such as cod and haddock and also with a very valuable sea scallop fishery. Fish on GB tend to have ideal growing conditions and mature quickly. GB is relatively self-contained with deep channels to the northeast and ocean currents containing waters on the bank giving the region a distinct character. However, the GB community does have seasonal migrants such as mackerel and dogfish which affect the community. Drastic changes occurred on GB in the late 1980s, where groundfish were much less abundant. We have termed 1988 as the collapse year for GB. The ESS, though geographically not far from GB is a much different system with lower productivity, diversity and more open to both the northwest and the southwest biologically and oceanographically. A key characteristic of the ESS is the presence of a small sandy arc 200 km offshore called Sable Island, which is the largest grey seal breeding colony in the world and has been growing exponentially since the mid-1980s. The ESS showed drastic declines in cod and some other groundfish in the early 1990s to almost undetectable levels. We consider 1992 to be the collapse year for ESS. The NS is a shallow warm sea with high fish community diversity and productive multi-species fisheries. The NS has supported very large groundfish and pelagic fisheries and despite extremely high fishing pressure, it is difficult to see a sudden change in the system that might be termed a collapse as seen in GB and ESS. The NS fish community always seems to respond positively to curtailment of fishing effort, while the equivalent is not true for GB and ESS.

(b) *Experiments*

The experiments undertaken in this paper involve applying classification. This involves the prediction of a pre-selected variable (here functional collapse) based on the values of other variables (here species biomass). Feature selection is used to identify the relevant species for optimal classification. There are two approaches to feature selection: filter selection that simply scores variables (species) independently, and wrapper selection that builds models and selects combinations of variables (thus identifying interactions between them). These experiments adopt the BN classifier approach, where the class node is a binary variable that represents functional collapse in GB. The K2 search algorithm [20] is used to build the BN classifiers. This involves a greedy search technique where links are incrementally added to an initially unconnected graph and scored using the metric given in equation (2.1), where n is the number of nodes, F_{ijk} is the frequency of occurrences in the dataset that the node x_i takes on the value vik (where there are r_i possible instantiations) and the parent nodes π_i take on the instantiation w_{ij} (where there are q_i possible instantiations). This metric is based on equation (2.2), which calculates the probability of observing a structure G and a set of data D, $p(G,D)$, where c is a constant prior probability $p(G)$. For simplicity, we assume a step change in functional structure in 1988 for GB data and 1992 for ESS. Further work will explore using hidden variables with more states and continuous variables to explore intermediate stages prior to collapse. A bootstrap [46] approach is employed to repeat the following 1000 times:

1. Score each species using the likelihood score given in equation (2.1) and take the mean over the bootstrap. This is known as *filter* feature selection [10] and scores each variable independently.
2. Learn BN structure with the (greedy) K2 algorithm and score the proportion of times that links are associated with the class node during the bootstrap (the confidence). This is known as *wrapper* feature selection [10] and scores each variable by taking into account their interaction with other variables through the use of a classifier model.

$$\log \sum_{i=1}^{n} \sum_{j=1}^{q_i} \frac{(r_i - 1)!}{(F_{ij} + r_i - 1)!} \sum_{k=1}^{r_i} F_{ijk} \tag{2.1}$$

$$\max_{G}[p(G, D)] = c \prod_{i=1}^{n} \max_{\pi_i} \left[\prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \right] \tag{2.2}$$

and

$$\log p(G|D) \approx \log p(D|G, \hat{\theta}_G) - \frac{\log M}{2} \mathrm{Dim}_G \tag{2.3}$$

We rank species based upon these two feature selection approaches and examine their relevance to functional collapse in GB. In order to explore the functionally equivalent species in the NS and ESS data, we use species identified using feature selection from GB in conjunction with the *functional equivalence* search algorithm (which is fully documented in algorithm 1). This is applied to both the NS and ESS to identify equivalent species. Finally, we use dynamic models, specifically DBNs (as seen in figure 1*d*) but with a single dynamic hidden variable to identify functional collapse. These networks are built from the GB data (using the REVEAL algorithm [47], which is a greedy search applied to DBNs) to predict cod biomass and functional collapse (using the hidden variable).

The functional equivalence algorithm uses a simulated annealing approach [48] to search for an optimal combination of variables that fit the given function. This is where a random allocation of selected variables is initialized and scored. Within each iteration, a single replacement is made to the selected variables and the new selection is scored. Here, we demonstrate the approach using a BN model, where the given function is in the form of a predefined BN structure, BN_1, and set of variables, $vars_1$ that is parametrized from a dataset, $data_1$. This model is then used to search for the variables in another dataset, $data_2$ that fits best. The algorithm gives as output the set of variables that best fits the given model. We use the Bayesian Information Criterion which penalizes overly connected networks to avoid overfitting. It is given in equation (2.3), where M is the number of samples, Dim_G is the dimension of the model, and $\hat{\theta}_G$ is the maximum-likelihood estimate of the parameters. The first term is essentially the log-likelihood and the second is a penalty for model complexity. We set iterations = 1000 and $t_{\mathrm{start}} = 1000$ as these were found through experimentation to allow convergence to a good solution.

Algorithm 1. The functional equivalence search algorithm.

Input: t_{start}, *iterations*, $data_1$, $data_2$, $vars_1$, BN_1
Parametrize Bayesian Network, BN_1 from $data_1$
Generate randomly selected variables in $data_2$: $vars_2$
Use $vars_2$ to score the fit with selected model BN_1 using equation 2.2: *score*
Set *bestscore* = *score*
Set initial temperature: $t = t_{start}$
for $i = 1$ to *iterations* **do**
  Randomly replace one selected variable in $data_2$ and rescore using equation 2.2: *rescore*
  dscore = *rescore* − *bestscore*
  if *dscore* ≥ 0 OR *UnifR and* (0,1) < $\exp^{(dscore/t)}$ **then**
    bestscore = *rescore*
  else
    Undo variable switch in $vars_2$
  end if
  Update the temperature: $t = t \times 0.9$
  end for
Output: $vars_2$

3. RESULTS

Figure 3 displays the rankings for filter and wrapper feature selection for differentiating between pre- and post-functional collapse in GB (1988). From both feature selection approaches, it is clear that there are a relatively small number of key players in this collapse and these are known to be involved with cod. For example, the likelihood approach strongly implicates two zooplankton species (*Calanus* and *Pseudocalanus*) as key to the functional collapse and it is known from other sources that there were relatively large changes then [49], and these changes can have bottom-up effects which affect species such as cod [50]. Herring (*Clupea harengus*) was also identified as a key species and its abundance changes in the late 1980 may have changed the predation environment of juvenile cod whose recruitment to adult stages may, in some systems, be significantly controlled by herring abundance [51]. Thorny skate (*Amblyraja radiata*) became more abundant at the time of the cod collapse on GB and although some attribute this to an ecosystem regime shift [52] others attribute this to immigration from the ESS [53].

Using the higher ranking species from figure 3, a DBN model was built with a hidden node using the REVEAL algorithm (see §2) to confirm how predictable both cod biomass and the unobserved functional collapse were from the related species. Figure 4 plots these results and shows that a reasonable fit to the GB data is achievable. What is more, the hidden state identifies a noisy underlying process which appears to stabilize somewhere in the late-1980s correlating with the expected functional collapse.

The confidences resulting from the Bayesian wrapper method applied to GB showed a quick decline with species rank, such that thorny skate was the most important species implicated in the decline. When this structure is imposed on the ESS and NS using the functional equivalence search, a small number of functionally equivalent species are identified in both the ESS and the NS with high confidence (figure 5). An interesting thing to note was the species/processes

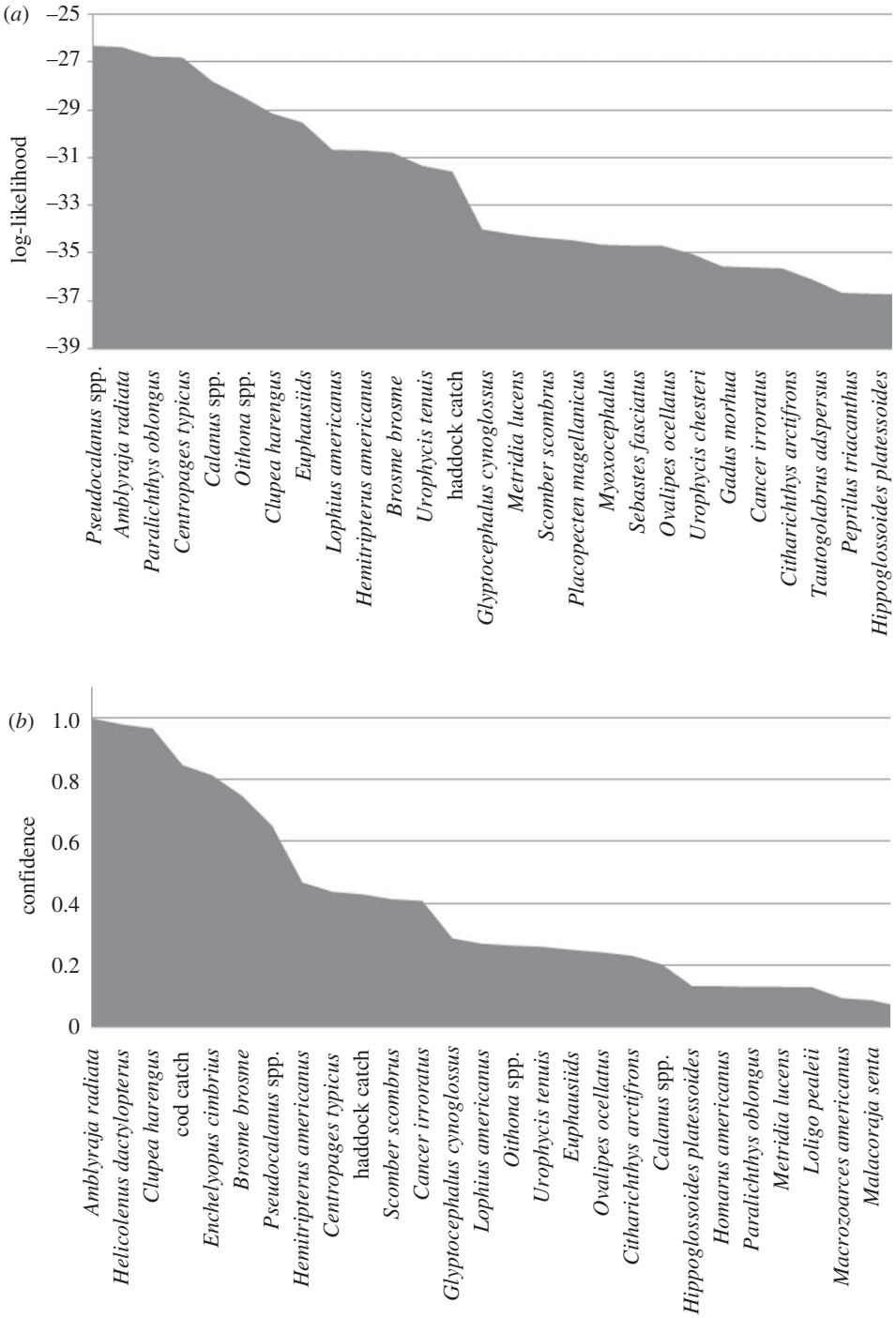

Figure 3. Features selected from GB data using a bootstrap on (*a*) filter selection using log-likelihood and (*b*) a Bayesian network classifier wrapper.

on GB, where the change in the community seemed to be captured by changes in two zooplankton species while in the ESS and the NS, there was no strong indication of zooplankton changes that accompanied fish community change.

Perhaps, the most striking feature of the functional equivalence applied to the ESS is the presence of many deepwater species such as argentine (*Argentina sphyraena*), grenadier (*Nezumia bairdi*) and hakes (*Merluccius bilinearis*). Surprisingly, cod was not implicated in the ESS collapse despite the fact that cod were a highly targeted species prior to collapse. The inclusion of grey seals is also expected as they were implicated in the decline and lack of recovery of many groundfish stocks on the ESS. The largest breeding colony of grey seals in the world is located on Sable Island in the middle of the ESS.

The presence of coldwater-seeking deepwater species on the ESS could be an indication of the water cooling that occurred on the ESS in the late-1980s and early-1990s, which also led to increases in coldwater shrimp and snow crabs. Furthermore, though grey seals increased in abundance at the same time, grey seals are not deep divers and if the deepwater species remained in the shelf basins and slope water, they would be less susceptible to grey seal predation than would cod.

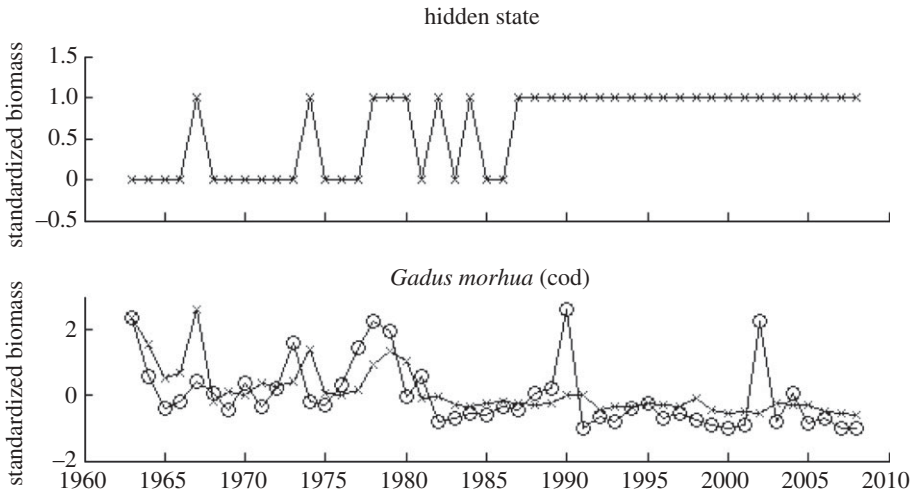

Figure 4. The fit for the model trained on GB data along with the associated discovered hidden state. The series marked with crosses denote the predicted biomass and hidden state as opposed to the observed biomass denoted by circles.

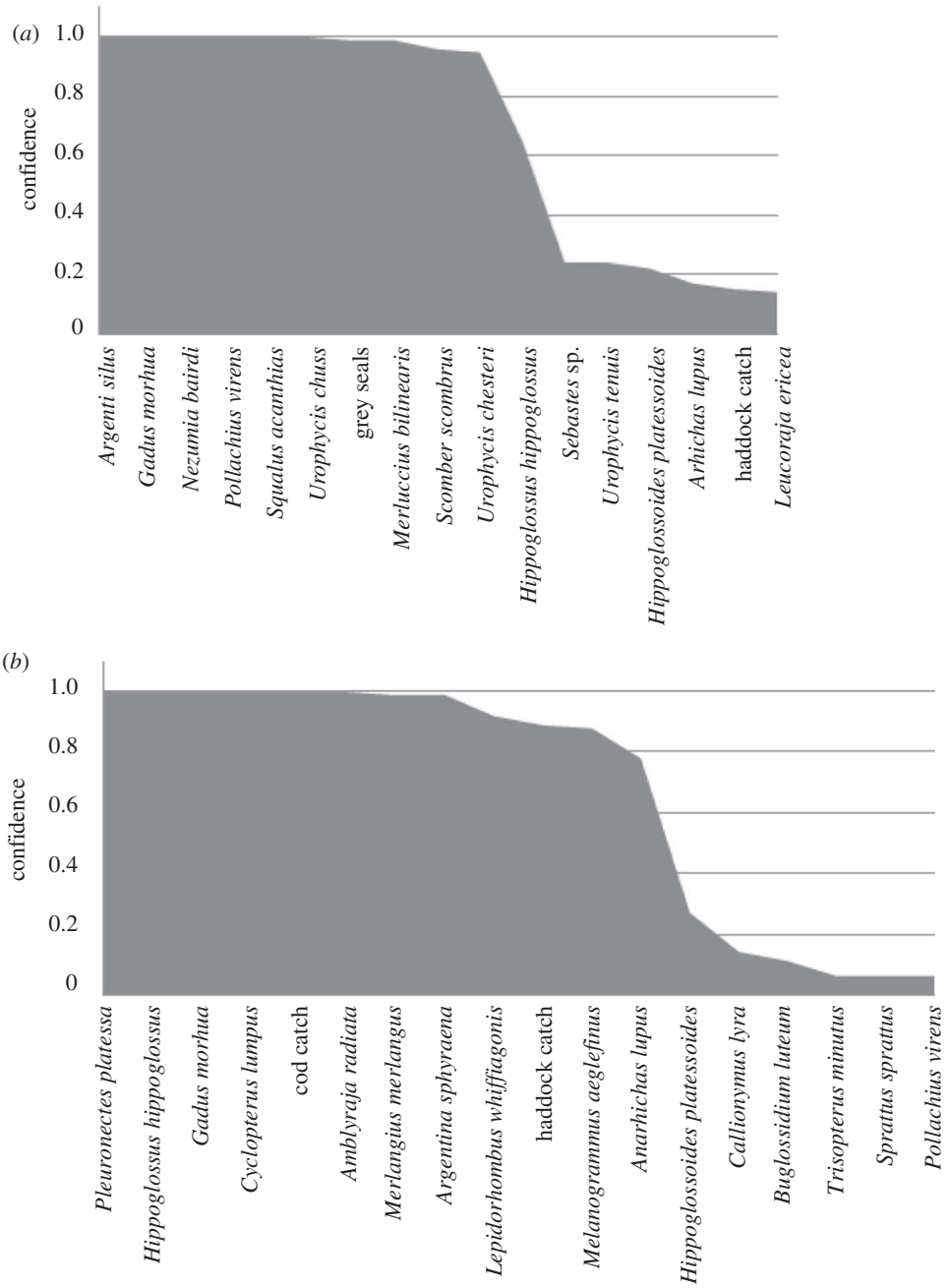

Figure 5. Functionally equivalent species to those selected from GB data identified using the functional equivalence algorithm. (*a*) Shows the equivalent species in the ESS and (*b*) shows the species in the NS.

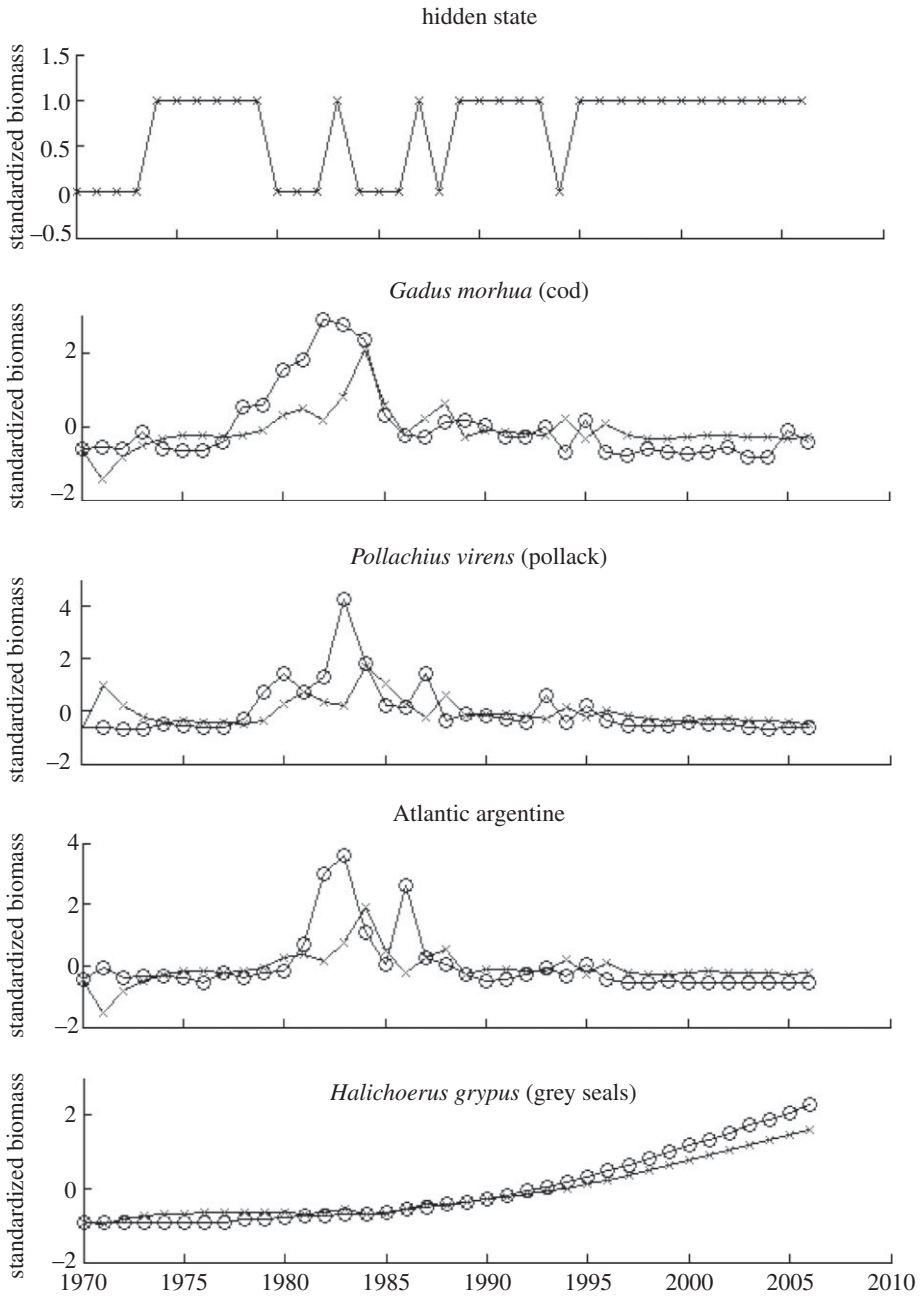

Figure 6. One-step ahead prediction of cod using DBN model trained on GB data and mapping to equivalent species in the ESS (identified using the functional equivalence algorithm along with the associated discovered hidden state). The series marked with crosses denote the predicted biomass and hidden state as opposed to the observed biomass denoted by circles.

In the NS, most of the selected species are commercially desirable and some experienced large declines in biomass in this period, though the nature of the species is not dissimilar to GB when compared with ESS, which showed the appearance of some qualitatively very different species. Catch of haddock and cod appeared to be important in the NS while commercial fish catch seemed less important on the ESS. These factors combined might suggest that catch is one of the most important factors driving change in the NS, while on the ESS, it may be that other factors lead to fundamental changes in the fish community composition.

The final set of results explore how well the functionally equivalent species can predict future biomass and the underlying state of the geographical system. Figure 6 documents these results for the selected functionally equivalent species for ESS (using the DBN trained on GB data and then mapped on equivalent species on ESS). The prediction of many of these species was surprisingly good, with close fits to the observed data. This is impressive considering that the model was parametrized using biomass data from different species in GB. For example, the model predicts the increase in seal numbers year after year based upon parameters determined on the relationship between cod catch and other species in GB. What is more, the hidden state inferred from the predicted data resembles very much what was observed in terms of functional collapse. While the state fluctuates in the period up to the late-1980s/early-1990s, in the period after the collapse the state becomes very

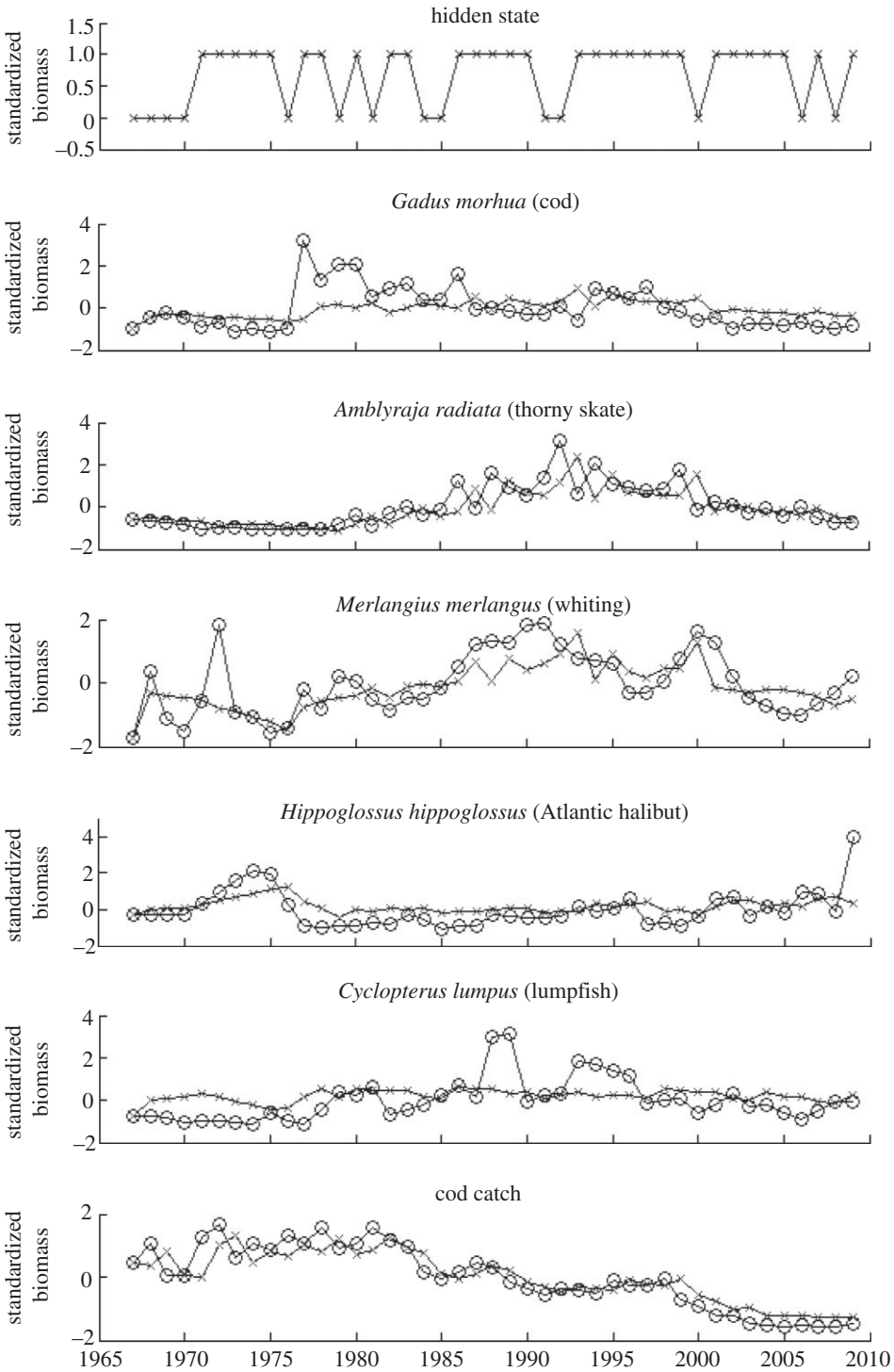

Figure 7. One-step ahead prediction of cod using DBN model trained on GB data and mapping to equivalent species in the NS (identified using the functional equivalence algorithm along with the associated discovered hidden state). The series marked with crosses denote the predicted biomass and hidden state as opposed to the observed biomass denoted by circles.

stable. This further adds credence to the conclusion that the selected species are indeed key to the functional collapse of cod in the ESS.

The same analysis was applied to identifying functionally equivalent species in the NS and testing them for prediction of biomass and identifying changes in the underlying state. Figure 7 illustrates the results. Firstly, note that the hidden state does not appear much less stable than in the ESS results. Rather than identifying no change in state (as was expected as no collapse has been observed), the hidden variable appears to have fitted the states to some noise process that fluctuates throughout the series. This could be due to the hidden state capturing the functional collapse successfully, which is the most influential predictive feature of cod in the ESS dataset, whereas the prediction of cod in the NS is more complex due to the lack of any collapse.

4. DISCUSSION

Since the large-scale fisheries collapses in many different regions globally in the late-1980s and into the 1990s, there has been a search for causal mechanisms (e.g. [54]). This research has included studies of

fisheries on species [55] as well as indirect effects of modifications of food webs and functional structure [56]. They have included simulation studies on functional structures in food webs [54,57], development of static functional structures through covariance techniques [58] or summaries of complicated multivariate data to examine overall temporal trends [59]. The present use of machine-learning techniques and BNs is another method applied to the the problem.

The use of bioinformatics techniques in this paper is unique because it exploits functional equivalence between different datasets and uses the identified species in conjunction with a dynamic model that uses latent variables to predict functional collapse (and future biomass). The recognition of a latent variable is important in fish community change studies of this nature because it allows causes of change which are not purely found within the constrained model structure. This is very different from mass balance model approaches whose fitting is conditioned completely upon the model structure. The latent variable therefore may partially represent something external to the fish community such as oceanographic conditions. We intend to explore this further by using data of likely factors such as temperature, nutrients and fishing mortality. Changes in conditions external to the fish community may be responsible for collapse in GB and ESS. The longer runs of similar estimates for the hidden state compared with NS could suggest different processes occurring there. Oceanographic conditions are a contender for ESS. For GB, what is occurring is less clear. NS, being highly exploited but shallow and dynamic, may naturally be more variable and able to cope with disturbances that would send the other two systems into collapse. Further work is warranted and exploration of other processes such as system variability before and after collapse [60] may prove to be useful predictors of collapse.

BN models also facilitate the direct incorporation of expertise into the structures and parameters. While this has not been explored fully here (the use of food webs have been used mostly for validation), using informative priors in the network models based upon available expertise will be investigated. The modelling approach also differs from other methods in how correlative structures, which are assumed to represent causal functional relationships, discovered in one system can be imposed upon another system. The components of the other system which best fit these structures can then be found in other systems. The topology of the BN allows us to explore these structures explicitly and a follow-up study will explore them prior to and after suspected regime changes. Though most ecosystem studies recognize the functional relation approach between species, most cannot deal with it in as pure a sense. Essentially, what this approach assumes is that there are only a few ways for similar ecosystems to organize themselves functionally even though the components may have different qualities; our analysis suggests that there may be similar ways to collapse. This can provide real insights into why fished ecosystems collapse and why they sometimes do not recover when a perturbation stops. Most importantly, it may give us an insight into signs of an imminent collapse perhaps while there is still time to prevent it.

We would like to thank Jerry Black DFO-BIO Halifax for assistance with the ESS survey data, Alida Bundy DFO-BIO Halifax for the ESS food web, the ICES datras database for the North Sea IBTS data, Bill Kramer NOAA–NMFS Woods Hole for providing the Georges Bank survey, Jason Link NOASS–NMFS for the Georges Bank food web, Jon Hare NOAA–NMFS for NE USA plankton data, SAHFOS for North Sea plankton data and Mike Hammill for ESS grey seal data.

REFERENCES

1 Bishop, C. M. 2006 *Pattern recognition and machine learning*. New York, NY: Springer.
2 Hand, D. J., Mannila, H. & Smyth, P. 2001 *Principles of data mining*. Cambridge, MA: MIT Press.
3 Berthold, M. R., Borgelt, C., Hoppner, F. & Klawonn, F. 2010 *Guide to intellgent data analysis*. Berlin, Germany: Springer.
4 Peek, N., Combi, C. & Tucker, A. 2009 Biomedical data mining. *Methods Inform. Med.* **48**, 225–228.
5 Eisen, M., Spellman, P., Brown, P. & Botstein, D. 1998 Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 8–25. (doi:10.1073/pnas.95.25.14863)
6 Swift, S., Tucker, A., Liu, X., Martin, N., Orengo, C. & Kellam, P. 2004 Consensus clustering and functional interpretation of gene expression data. *Genome Biol.* **5**, R94.
7 Segal, E., Friedman, N., Koller, D. & Regev, A. 2004 A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* **36**, 1090–1098. (doi:10.1038/ng1434)
8 Bar-Joseph, Z. *et al.* 2003 Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **21**, 1337–1342. (doi:10.1038/nbt890)
9 Pittman, J. *et al.* 2004 Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl Acad. Sci. USA* **101**, 8431–8436. (doi:10.1073/pnas.0401736101)
10 Inza, I., Larrañaga, P., Blanco, R. & Cerrolaza, A. J. 2004 Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.* **31**, 91–103. (doi:10.1016/j.artmed.2004.01.007)
11 Kim, S., Imoto, S. & Miyano, S. 2003 Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief. Bioinform.* **4**, 228–235. (doi:10.1093/bib/4.3.228)
12 Pe'er, D., Regev, A., Elidan, G. & Friedman, N. 2001 Inferring subnetworks from perturbed expression profiles. In *Proc. 9th Int. Conf. Intelligent Systems for Molecular Biology (ISMB 2001), Copenhagen, Denmark, July 2001*. Oxford, UK: Oxford Journals.
13 Li, H., Xuan, J., Wang, Y. & Zhan, M. 2008 Inferring regulatory networks. *Front. Biosci.* **13**, 263–275. (doi:10.2741/2677)
14 Hochachka, W. M., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D. & Kelling, S. 2007 Data-mining discovery of pattern and process in ecological systems. *J. Wildlife Manage.* **71**, 2427–2437. (doi:10.2193/2006-503)
15 Friedman, N., Linial, M., Nachman, I. & Pe'er, D. 2000 Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620. (doi:10.1089/106652700750050961)
16 Pe'er, D., Tanay, A. & Regev, A. 2006 MinReg: a scalable algorithm for learning parsimonious networks in yeast and mammals. *J. Mach. Learning Res.* **7**, 167–189.

17 Hartemink, A. J., Gifford, D., Jaakkola, T. & Young, R. 2002 Bayesian methods for elucidating genetic regulatory networks. *IEEE Intell. Syst.* **17**, 37–43.
18 Spirtes, P., Glymour, C. & Scheines, R. 2000 *Causation, prediction, and search*. Cambridge, MA: MIT Press
19 Pearl, J. 2001 *Causality: models, reasoning, and inference*. Cambridge, UK: Cambridge University Press
20 Cooper, G. F. & Herskovitz, E. 1992 A Bayesian method for the induction of probabilistic networks from data. *Mach. Learning* **9**, 309–347. (doi:10.1007/BF00994110)
21 Janžura, M. & Nielsen, J. 2006 A simulated annealing-based method for learning Bayesian networks from statistical data: research articles. *Int. J. Intell. Syst.* **21**, 335–348. (doi:10.1002/int.20138)
22 Steele, E., Tucker, A., Hoen, PAC.'t & Schuemie, M. J. 2009 Literature-based priors for gene regulatory networks. *Bioinformatics* **25**, 1768–1774. (doi:10.1093/bioinformatics/btp277)
23 Larrañaga, P., Poza, M., Yurramendi, Y., Murga, R. H. & Kuijpers, C. M. H. 1996 Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters. *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 912–926. (doi:10.1109/34.537345)
24 Langley, P., Iba, W. & Thompson, K. 1992 An analysis of Bayesian classifiers. In *Proc. 10th Natl Conf. Artificial Intelligence, San Jose, CA, July 1992*, pp. 223–228. Menlo Park, CA: AAAI Press
25 Ghahramani, Z. 1998 Learning dynamic Bayesian networks. In *Adaptive processing of sequences and data structures. Lecture Notes in Artificial Intelligence*, pp. 168–197. Berlin, Germany: Springer.
26 Friedman, N., Geiger, D. & Goldszmidt, M. 1997 Bayesian network classifiers. *Mach. Learning* **29**, 131–163. (doi:10.1023/A:1007465528199)
27 Munch, K., Gardner, P. P., Arctander, P. & Krogh, A. 2006 A hidden Markov model approach for determining expression from genomic tiling microarrays. *BMC Bioinform.* **7**, 239. (doi:10.1186/1471-2105-7-239)
28 Castelo, R. & Siebes, A. 2000 Priors on network structures: biasing the search for Bayesian networks. *Int. J. Approximate Reasoning* **24**, pp. 39–57. (doi:10.1016/S0888-613X(99)00041-9)
29 Imoto, S., Goto, T. & Miyano, S. 2002 Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. In *Pacific Symposium on Biocomputing, Lihue, Hawaii, January 2002*, vol. 7, pp. 175–186.
30 Werhli, A. V. & Husmeier, D. 2007 Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.* **6**, art. 15 (doi:10.2202/1544-6115.1282)
31 Yauk, C., Nerndt, M. L., Williams, A. & Douglas, G. 2004 Comprehensive comparison of six microarray technologies. *Nucleic Acids Res.* **32**, e124. (doi:10.1093/nar/gnh123)
32 Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L. & Kohane, I. S. 2002 Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412. (doi:10.1093/bioinformatics/18.3.405)
33 Jarvinen, A. K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallionoemi, O. P. & Monni, O. 2004 Are data from different gene expression microarrays comparable? *Genomics* **83**, 1164–1168. (doi:10.1016/j.ygeno.2004.01.004)
34 Steele, E. & Tucker, A. 2008 Consensus and meta-analysis regulatory networks for combining multiple microarray gene expression datasets. *J. Biomed. Inform.* **41**, 914–926. (doi:10.1016/j.jbi.2008.01.011)
35 Steele, E. & Tucker, A. 2009 Selecting and weighting data for building consensus gene regulatory networks. In *Advances in intelligent data analysis VIII*, vol. LNCS 5772, pp. 190–201. Berlin, Germany: Springer.
36 Anvar, S. Y., 't Hoen, PA. & Tucker, A. 2010 The identification of informative genes from multiple datasets with increasing complexity. *BMC Bioinform.* **11**, 32. (doi:10.1186/1471-2105-11-32)
37 Anvar, Y., Tucker, A., Vinciotti, V., Venema, A., van Ommen, G. J. B., van der Maarel, S. M., Raz, V. & 't Hoen, P. A. C. In press. Interspecies translation of disease networks increases robustness and predictive accuracy. *PLOS Comput. Biol.*
38 Garrison, L. P. & Link, J. S. 2000 Dietary guild structure of the fish community in the northeast united states continental shelf ecosystem. *Mar. Ecol. Prog. Ser.* **202**, 231–240. (doi:10.3354/meps202231)
39 Jaio, Y. 2009 Regime shift in marine ecosystems and implications for fisheries management, a review. *Rev. Fish Biol. Fish.* **19**, 177–191. (doi:10.1007/s11160-008-9096-8)
40 Bakun, A. 2006 Wasp-waist populations and marine ecosystem dynamics: navigating the 'predator pit' topographies. *Prog. Oceanogr.* **68**, 271–288. (doi:10.1016/j.pocean.2006.02.004)
41 Marcot, B. G., Steventon, J. D., Sutherland, G. D. & McCann, R. K. 2006 Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Can. J. For. Res.* **36**, 3063–3074. (doi:10.1139/x06-135)
42 Hammond, T. R. & O'Brien, C. M. 2001 An application of the Bayesian approach to stock assessment model uncertainty. *ICES J. Mar. Sci.* **58**, 648–656. (doi:10.1006/jmsc.2001.1051)
43 Marcot, B. G., Holthausen, R. S., Raphael, M. G., Rowland, M. M. & Wisdom, M. J. 2001 Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *For. Ecol. Manage.* **153**, 29–42. (doi:10.1016/S0378-1127(01)00452-2)
44 SEIS 2008. European commission: towards a shared environmental information system. See http://ec.europa.eu/environment/seis/index.htm.
45 Thrush, S., Giovani, C. & Hewitt, J. E. 2008 Complex positive connections between functional groups are revealed by neural network analysis of ecological time-series. *Am. Nat.* **171**, 669–677.
46 Efron, B. & Tibshirani, R. 1995 Cross-validation and the bootstrap: estimating the error rate of a prediction rule. Technical Report no. TR-477, Department of Statistics, Stanford University, Stanford, CA, USA.
47 Liang, S., Fuhrman, S. & Somogyi, R. 1998 Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing, Maui, Hawaii, January 1998*, vol. 3, pp. 18–29.
48 Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. 1983 Optimization by simulated annealing. *Science* **220**, 671–680. (doi:10.1126/science.220.4598.671)
49 Kane, J. 2007 Zooplankton abundance trends on Georges Bank, 1977–2004. *ICES J. Mar. Sci.* **64**, 909–919. (doi:10.1093/icesjms/fsm066)
50 Beaugrand, G., Brander, K., Lindley, J. A., Souissi, S. & Reid, P. C. 2000 Plankton effect on cod recruitment in the north sea. *Nature* **426**, 661–664. (doi:10.1038/nature02164)
51 Swain, D. P. & Sinclair, A. 2000 Pelagic fishes and the cod recruitment dilemma in the northwest atlantic. *Can. J. Fish. Aquat. Sci.* **57**, 1321–1325. (doi:10.1139/f00-104)

52 Fogarty, M. J. & Murawski, S. A. 1998 Large-scale disturbance and the structure of marine systems: fishery impacts on Georges Bank. *Ecol. Appl.* **8**, S6–S22.
53 Frisk, M. G., Miller, T. J., Martell, S. J. D. & Sosebee, K. 2008 New hypothesis helps explain elasmobranch 'outburst' on Georges Bank in the 1980s. *Ecol. Appl.* **18**, 234–245. (doi:10.1890/06-1392.1)
54 Bundy, A. 2001 Fishing on ecosystems: the interplay of fishing and predation in Newfoundland and Labrador. *Can. J. Fish. Aquat. Sci.* **58**, 1153–1167.
55 Myers, R. A. & Worm, B. 2003 Rapid worldwide depletion of predatory fish communities. *Nature* **423**, 280–283. (doi:10.1038/nature01610)
56 Frank, K., Petrie, B., Choi, J. & Leggett, W. 2005 Trophic cascades in a formerly cod-dominated ecosystem. *Science* **308**, 1621–1623. (doi:10.1126/science.1113075)
57 Petrie, B., Frank, K. T., Shackell, N. L. & Leggett, W. C. 2009 Structure and stability in exploited marine fish communities: quantifying critical transitions. *Fish. Oceanogr.* **18**, 83–101. (doi:10.1111/j.1365-2419.2009.00500.x)
58 Duplisea, D. E. & Blanchard, F. 2005 Relating species and community dynamics in an heavily exploited marine fish community. *Ecosystems* **8**, 899–910. (doi:10.1007/s10021-005-0011-z)
59 Choi, J. S., Frank, K. T., Petrie, B. D. & Leggett, W. C. 2005 Integrated assessment of a large marine ecosystem: a case study of the devolution of the Eastern Scotian shelf. *Oceanogr. Mar. Biol.* **43**, 47–67.
60 Carpenter, S. R. *et al.* 2011 Early warnings of regime shifts: a whole-ecosystem experiment. *Science* **27**, 1079–1082. (doi:10.1126/science.1203672)