Construction and Validation of Food Webs Using Logic-Based Machine Learning and Text Mining

Alireza Tamaddoni-Nezhad[*,1], Ghazal Afroozi Milani[*],
Alan Raybould[§], Stephen Muggleton[*], David A. Bohan[†]

*Computational Bioinformatics Laboratory, Department of Computing, Imperial College London, London, United Kingdom
[†]UMR 1347 Agroécologie, AgroSup/UB/INRA, Pôle Ecologie des Communautés et Durabilité de Systèmes Agricoles, 21065 Dijon Cedex, France
[§]Product Safety, Syngenta, Jealott's Hill International Research Centre, Bracknell, Berks, United Kingdom
[1]Corresponding author: e-mail address: a.tamaddoni–nezhad@imperial.ac.uk

Contents

Abstract

Network ecology holds great promise as an approach to modelling and predicting the effects of agricultural management on ecosystem service provision, as it bridges the gap between community and ecosystem ecology. Unfortunately, trophic interactions between most species in agricultural farmland are not well characterised empirically, and only partial food webs are available for a few systems. Large agricultural datasets of the nodes (i.e., species) in the webs are now available, and if these can be enriched with information on the links between them then the current shortage of network data can potentially be overcome. We demonstrate that a logic-based machine learning method can be used to automatically assign interactions between nodes, thereby generating plausible and testable food webs from ecological census data. Many of the learned trophic links were corroborated by the literature: in particular, links ascribed with high probability by machine learning corresponded with those having multiple references in the literature. In some cases, previously unobserved but high probability links were suggested and subsequently confirmed by other research groups. We evaluate these food webs using a new cross-validation method and present new results on automatic corroboration of a large, complex food web. The simulated frequencies of trophic links were also correlated with the total number of literature 'hits' for these links from the automatic corroboration. Finally, we also show that a network constructed by learning trophic links between functional groups is at least as accurate as the species-based trophic network.

1. INTRODUCTION

1.1. Ecosystem services and agricultural management

The requirements from agricultural land are becoming more diverse. In many parts of the world, there has been a change from the immediate post–WWII focus on increasing crop productivity from agricultural land, to a view that managed agricultural landscapes provide humans with a suite of ecosystem services, some of which are essential for the long-term productivity of agriculture (Raffaelli and White, 2013). Sustainable intensification of agriculture (SIA) describes the process whereby crop productivity increases, but not at the expense of other services provided by non–crop biodiversity that maintain crop production and the intrinsic cultural value that humans place on biodiversity (Baulcombe et al., 2009). The policy shift from increasing crop productivity towards SIA means that decisions about the best way to manage agricultural land are becoming more complicated. To help decision-making for SIA, we need to predict the effects of agricultural management on biodiversity, and to predict the effects of changed biodiversity on the delivery of ecosystem services. However, ongoing changes in management practice, climate, water, etc., mean that decisions cannot always wait for the full complement of empirical evidence or data to be gathered.

The Millennium Ecosystem Assessment (MEA, 2005) considered the services provided by nature as humanity's 'life-support system'. It concluded that human activity is having a significant and escalating impact on the biodiversity of world ecosystems, reducing both their long-term resilience and productivity. The MEA developed a framework for evaluating the effects of changes in biodiversity on ecosystems, and the value of those changes in service to humanity. It posited four key types of service that ecosystems provide, all of which are relevant to agriculture: provisioning services, such as the production of food, fuel and fibre; regulation services, such as pollination, biological control and carbon sequestration; cultural services, such as cultural heritage, education and recreation spaces; and, support services, including the cycling of nitrogen and carbon. There have been numerous refinements and modifications to this schema in recent years, but these general principals and terms are now widely used in both the natural and social sciences (Raffaelli and White, 2013).

Changing in biodiversity, and particularly its loss, affects ecosystem structure and resilience, and may result in the partial or complete loss of ecosystem services (Chapin et al., 1998; Costanza et al., 1997), which are ultimately provided by processes driven by particular functional groups within the biota (Bennett et al., 2009). Economically and socially, biodiversity provides functions that become ecosystem services only if a human 'beneficiary' can be identified (Busch et al., 2012; Fisher et al., 2009), as illustrated in the 'cascade model' (De Groot, 2010; Haines-Young and Potschin, 2010).

Although the links between species, functioning and services are neither simple nor linear, it has become increasingly evident that the loss of biodiversity has affected agricultural ecosystem services (e.g., biological control, Cardinale et al., 2003). Human-driven management, whether at local or landscape scales, will alter ecosystem services (Lavorel et al., 2010; Mulder et al., 2012). The use of pesticides and modern intensive farming techniques over the past century, directed to increasing provisioning services in agriculture, have been used at the expense of local decreases in regulation and cultural services, and farmland biodiversity in general (Raudsepp-Hearne et al., 2010).

This picture is complicated by interactions between services; for example, the provisioning service of food production is directly affected by the regulation service of biological control of agricultural pests. Food production is also indirectly affected by support and regulation services, such as soil erosion and preservation of genetic resources, which affect long-term sustainability. Cultural services, including landscape aesthetics and the conservation of threatened and flagship species, may also affect production (Tilman et al., 2002).

Furthermore, increasing provisioning services from land already culti-
vated may protect other ecosystem services on uncultivated land (Green
et al., 2005).

1.2. Biodiversity, ecosystem functioning and services

Species-based, autecological studies of regulatory processes and services in
agriculture, such as control of the slug *Deroceras reticulatum* and weeds by
the carabid beetle *Pterostichus melanarius* (Bohan et al., 2000, 2011), have long
been a staple subject of applied ecology. The benefits of studying interactions
between two species are clear: the interaction is often defined by a specific
problem, such as damage caused by slugs to a particular crop; and questions
about the interaction can be broken down into conceptually simple steps
that are experimentally tractable. There is also a considerable body of eco-
logical theory within which any experimental results can be placed, for
instance to gauge how and when ecological regulation occurs.

A limitation of these autecological approaches is that they are explicitly
non-general. They are species-based, and species are often limited to particular
systems. In arable agricultural systems, there are marked changes in species com-
position between crops (e.g., Smith et al., 2008). Hence making statements
about likely regulatory effects in a novel system, even between closely related
and previously studied species, can be difficult. Working through all combina-
tions of species in a system, across all systems, is not feasible for describing or
managing ecosystem services; it is too slow given the urgent need. Furthermore,
autecological approaches may not capture large-scale processes effectively, many
of which underpin ecosystem services. This is because interactions among
species, processes and services all play an important role in the behaviour of
the ecosystem and ultimately how it will respond to management.

Extrapolating from species to functioning, and ultimately to ecosystem ser-
vices, is not straightforward. Functional approaches hold that resource use effi-
ciency by each species is a key determinant of ecological functioning. The null
model might therefore be that each species interaction contributes additively;
each species using the available resources similarly. Replacing any one species
by another does not affect overall resource use and resource use efficiency is, in
effect, independent of species diversity. However, based upon the simple
observation that diverse systems, containing a great many species and conse-
quently interactions, tend to be robust, stable and productive, alternative
hypotheses for the contribution of species to functioning have been proposed
(e.g., Reiss et al., 2009). These theories fall into two broad classifications:
assemblages with greater diversity use resources more efficiently; and

assemblages with greater diversity are more likely to contain at least one species that uses resources more efficiently, all other things being equal.

Differentiation of niches, or complementarity of niche requirements, promotes co-existence: a more diverse community should therefore be able to use resources more efficiently, and thus be both more productive and stable (Fargione et al., 2007; Kraft et al., 2008; Loreau and Hector, 2001). Certain species may also be able to modify the environment, facilitating co-occurring species (Stachowicz, 2001), as observed in the plants where there is competition for nutrients, water and space, and direct competitive interactions with invertebrate herbivores and other plants (e.g., allelopathy). Legumes are regularly used in agriculture as facilitators to enrich soil nitrogen, with clear economic benefit (i.e., an ecosystem service). Intraspecific competition is often more intense than its interspecific form, so substituting individuals of the same species by individuals of different species can reduce the competition experienced by an individual leading to greater productivity (Griffin et al., 2009) resulting in overyielding (Hector et al., 2002). The so-called sampling effect can arise due to an increased probability of including one species with atypically high productivity or efficient resource use, as the total species pool broadens (Fargione et al., 2007). The portfolio effect (Tilman et al., 1998) hypothesises that conditions that are bad for one species might be good for another, so across a portfolio of species (a community), the variation in performance is much lower, and this tends to increase with species diversity. All these examples highlight the inadequacy of considering species in isolation, and how biodiversity does not necessarily scale linearly with functioning, as it is modulated by complex species interactions. In addition, to these indirect competitive effects, trophic interactions in the food web can have powerful, and often counterintuitive effects, and these are best examined using network-based approaches (Reiss et al., 2009), although these are only now starting to be considered in agroecology (e.g., Bohan et al., 2013; Loeuille et al., 2013; Massol and Petit, 2013; Mulder et al., 2013; Tixier et al., 2013; Traugott et al., 2013).

1.3. A network approach to model field-scale agricultural management

To manage agricultural land for the optimal delivery of ecosystem services, we need theories that accurately predict the effects of perturbation on agroecosystems that affect their productivity and stability through impacts on diversity. Until recently, two very different approaches were available: community ecology and ecosystem ecology (Thompson et al., 2012). The former concentrates on the spatial and temporal distribution of individuals, populations and

species, and seeks explanations for distribution patterns in processes such as dispersal, selection and speciation (Vellend, 2010). While it can provide detailed explanations of the behaviour of the basic units of ecosystems, these explanations cannot be combined and scaled up to provide predictions of the behaviour of whole ecosystems (Thompson et al., 2012). Ecosystem ecology focuses on the flows of energy, biomass and nutrients through ecosystems, and explains these processes in terms of chemistry and thermodynamics (Jørgensen and Fath, 2004). It can provide models of dynamics at large scales, for example, global carbon fluxes, but is often weak at predicting the effects of changes in species composition on individual ecosystems.

Both approaches have drawbacks for developing general theories that are useful for agro-ecosystem management. A more fruitful approach may be food web ecology, which bridges the gap between community and ecosystem ecology (Ings et al., 2009; Reiss et al., 2009; Thompson et al., 2012; Woodward et al., 2005a,b). Food webs represent the trophic interactions between organisms in an ecosystem as a network of species and the flow of energy (or biomass) between them. Food webs combine the species distribution elements of community ecology and the energy flows of ecosystem ecology, and as such offer a means of linking changes in species abundance to changes in ecosystem functioning (e.g., Mulder et al., 2012; O'Gorman and Emmerson, 2010). Food web approaches therefore offer a potentially powerful means to predict changes in the delivery of ecosystem services that result from the effects of agricultural management on biodiversity (see Fig. 4.1).

Recent theoretical and empirical work has suggested that there might be a simplifying framework underlying food webs. In a seminal study of Tuesday Lake, Michigan, United States, Cohen et al. (2003) established the food web links between 56 phytoplankton, zooplankton and fish species (see also Carpenter and Kitchell, 1988, 1993). In parallel, the authors used a simple energetic theory for how trophic height in a food web might be correlated with abundance and body mass: essentially, a few big things eat many little things (Cohen, 1991; Elton, 1927; Petchey et al., 2008). All other factors being equal, species at similar trophic heights had similar body mass and abundance: it is possible, then, to infer that food webs might be simplified by an amalgamation of species-based nodes to nodes that combine the three key biological variables of trophic height, individual abundance and body mass with little loss of explanatory power (Cohen et al., 2003). These so-called 'trivariate food webs' have been used in various forms to assess patterns of energy or biomass flux and size-structure, mostly in natural aquatic ecosystems (e.g., Layer et al., 2010, 2011; O'Gorman and Emmerson, 2010;
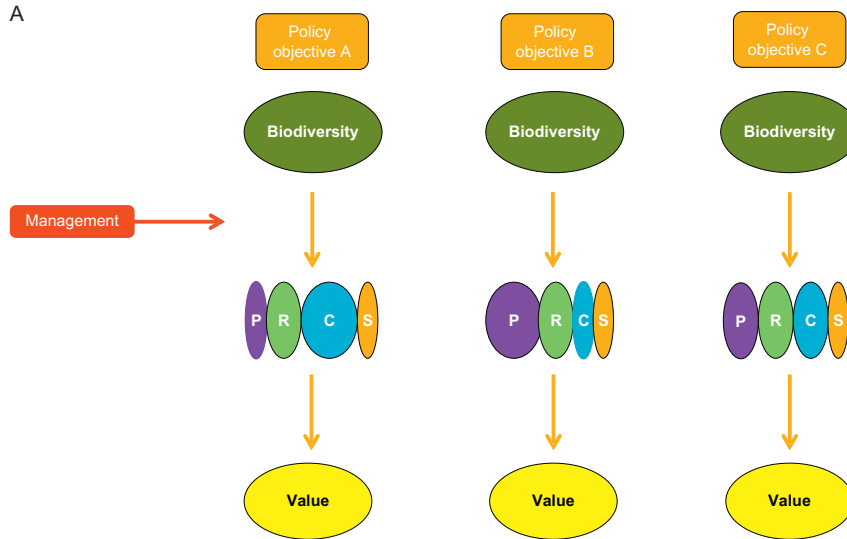
Figure 4.1 (A) Varied management can produce different distributions of ecosystem services. The optimum distribution of provisioning (P), regulating (R), cultural (C) and supporting (S) services will depend on the objectives of environmental and agricultural policy.

(continued)
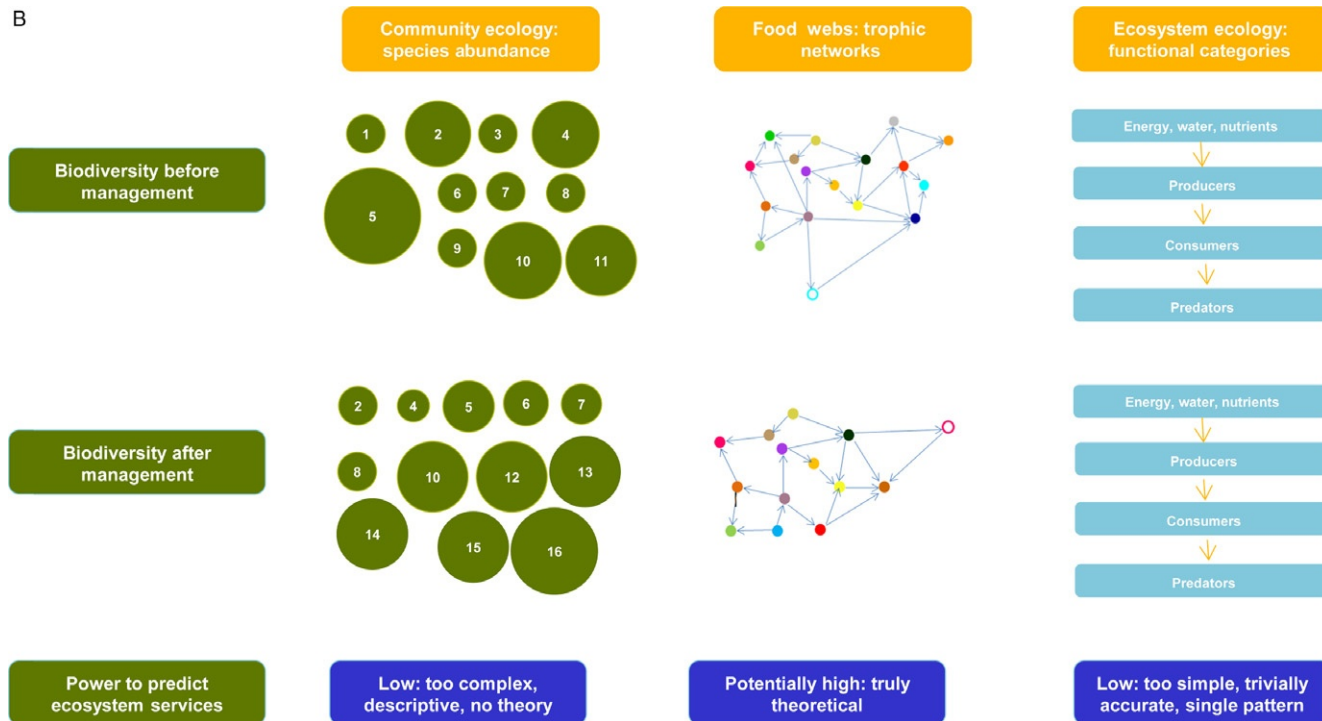
Figure 4.1—Cont'd (B) Delivery of the optimum distribution of ecosystem services will depend on our ability to predict the effects of management on biodiversity, and to predict the relationship between changes in biodiversity and the delivery of ecosystem services. Figure based on categories described by Thompson et al. (2012).

Woodward et al., 2005a,b), but also increasingly in their terrestrial counterparts (e.g., Mulder et al., 2012; Reuman et al., 2009) and also in experimental systems (Ledger et al., 2013; Woodward et al., 2012).

About 200 species of plants and aboveground invertebrates are commonly present in arable ecosystems across the UK (Smith et al., 2008). The crop and weeds compete for available resources. The invertebrates compete feed on the plants, on each other, or both. Conceptually, this system consists of primary producers, which transform resources into biomass, and primary consumer invertebrates that feed on this biomass, converting it to invertebrate biomass (Bohan et al., 2005; Hawes et al., 2003, 2009). Higher trophic levels of secondary consumer invertebrates then feed on and convert this invertebrate biomass (Brooks et al., 2003). Detritivores consume dead organisms from across all trophic levels.

This ecosystem can be represented using ecological network theory (Dunne et al., 2002). This theory treats the different species as discrete nodes, with particular properties of biomass and abundance, with the links between the nodes representing the interactions between the species. This approach has proved useful for yielding important information on the structure of the system and for evaluating the likelihood of node extinction and network collapse in a changing environment (Pascual and Dunne, 2006). Food webs are the most familiar form of ecological network, in which the links are trophic and represent the flow of energy or biomass between the nodes, and these have been used successfully to explain structural and dynamical properties of ecological systems (Dunne et al., 2002), particularly aquatic systems (Cohen et al., 2003; Emmerson and Raffaelli, 2004; Layer et al., 2010, 2011; Reuman and Cohen, 2004; Woodward et al., 2005a,b). Terrestrial food webs have proved less easy to explain (Polis, 1991; Rott and Godfray, 2000), although notable progress has been made in recent years (Hagen et al., 2012; Mulder and Elser, 2009; Mulder et al., 2011; Petchey et al., 2008). Arable farmland networks have only recently begun to be investigated and this work has concentrated on pollinator and parasitoid webs (Gibson et al., 2006; Van Veen et al., 2008), and increasing efforts have been devoted to understanding how these networks operate in fragmented agricultural landscapes (Hagen et al., 2012; Loeuille et al., 2013).

Recently, Pocock et al. (2012) showed that the species in an arable agricultural system in the UK provide distinct ecological functions and services, and that they are linked together through a network of different types of ecological interactions. There are, therefore, clear routes of interaction

between species providing different ecological functions, which will affect services. Changes in the highly linked species, such as certain flowering plants will affect both the regulation service provided by pollinators (through the ecosystem process of pollination) and the regulation service of biological control (through predation by parasitoids).

1.4. Networks and interactions in agriculture

The agro-ecological, mechanistic underpinning of ecosystem services, their response to change and how they interact are still poorly understood, as exemplified by the so-called 'optimist's scenario' (Pocock et al., 2012), which may be summarised as 'The management of one ecosystem service, for improved outcomes, benefits the outcomes of all ecosystem services'. However, the dependencies of one service and any other are often only poorly understood and the validity of this scenario at system-relevant scales can only be guessed at.

Pocock et al. (2012) tested the applicability of a food web approach to the analysis of delivery of ecosystem services using an extensive ecological network from UK agriculture. They showed considerable linkage between services, with particular plant and invertebrate species being disproportionately involved in the network of links. Managing, and therefore changing, the abundance or presence of certain species for the benefits from one service, such as pollination, would not always necessarily benefit other services, such as biocontrol of aphids. The optimist's scenario could not be maintained in this ecosystem. The large-scale problems presented by juggling the needs of different ecosystem services in the agricultural ecosystem could therefore benefit from some of the solutions that large-scale network ecology could provide.

Since ecosystems are structured by flows of energy (biomass) between primary producer plants (autotrophs) and consumers (heterotrophs), such as invertebrates, mammals and birds (Dickinson and Murphy, 1998; Lindeman, 1942), the food web is key for explaining ecosystem structure and dynamics and understanding and predicting responses to environmental change (Caron-Lormier et al., 2009; Cohen et al., 2009; Odum, 1971; Woodward et al., 2012).

Still relatively few ecosystems have been described and detailed using food webs because establishing predation relationships between the many hundreds of species in an ecosystem is resource intensive, requiring considerable investment in field observation and laboratory experimentation (Ings et al., 2009). Across such large datasets, it is often difficult to relate observational data sampled in protocols that have different, basic metrics; such as

density or activity-density or absolute abundance. Increasing the efficiency of testing for trophic links by filtering out unlikely interactions is typically not possible because of uncertainty about basic background knowledge of the network, such as whether any two species are likely even to come into contact and then interact (Ings et al., 2009). In addition, it may require considerable analysis and interpretation to translate from the ecological 'language' of sample data (count, abundance, density, etc.) to the network language of links within a trophic network. Consequently, of those ecosystems that have been studied using trophic network approaches, component communities that provide known, valuable ecosystem services or those that are experimentally tractable or under threat have most often been evaluated (Ings et al., 2009).

In a recent paper, Bohan et al. (2011) demonstrated that machine learning has the potential to construct realistic food webs, using a logic-based approach called abductive/inductive logic programming (A/ILP), and that it can generate plausible and testable networks from field sample data of the taxa (network nodes) alone. The data Bohan et al. (2011) used came from a national-scale Vortis suction sampling of invertebrates from arable fields in the UK: 45 invertebrate species or taxa, representing approximately 25% of the sample, and about 74% of the invertebrate individuals included in the learning, were hypothesised to be linked. As might be expected, detritivorous Collembola were consistently the most important prey, and generalist and omnivorous carabid beetles were hypothesised to be the dominant predators of the system. One surprising result, however, was the importance of carabid larvae, suggested by the machine learning, as predators of a wide variety of prey. High probability links were also hypothesised for widespread, potentially destabilising, intra-guild predation; predictions that could be experimentally tested.

A review of the literature revealed that many of the high probability links in the model had already been independently observed or suggested for this system, supporting the contention that A/ILP learning could produce plausible food webs from sample data, independent of preconceptions about who-eats-whom. Well-known links in the literature corresponded with links ascribed with high probability through A/ILP. Arguably, therefore, this very general machine learning approach has great power and could be used to extend and test theories of agricultural ecosystem dynamics and function. In particular, it could be used to support the development of a wider theory of ecosystem responses to environmental change.

There were, however, clear problems with the learning and validation methodology in Bohan et al. (2011) that required addressing and further

study. While showing great promise, the review was done by hand and reflects, at best, a partial knowledge of the literature that could lead to some links not being validated. In turn, certain links were likely overrepresented because of good operator knowledge of particular parts of the trophic literature. Finally, some reports came from what might be termed the 'grey literature', information that was not peer-reviewed; this literature should be considered anecdotal information at best.

Further, the network was learned for a single protocol, the Vortis suction sampling, which sampled species of invertebrates at the soil surface and in association with the weed plants. It was also explicitly species-based and therefore limited to the four cropping systems studied, due to the system-specific species compositions of agricultural cropping systems (Smith et al., 2008). The approach produced a testable network, but one that was not likely to be general.

Here, we extend both the work of Bohan et al. (2011) and tackle some of the inherent problems encountered in the initial work. Specifically, we will attempt to: (i) develop a robust (general) validation methodology for networks (originally discussed in Tamaddoni-Nezhad et al., 2012b and Afroozi Milani et al., 2012); (ii) grow the network by learning across disparate sampling protocols; and, (iii) develop a generic approach to networks that might allow us to move between systems, based on a more functional, as opposed to a purely taxonomic, approach.

1.5. Machine learning network models from data

Machine learning is the sub-area of computer science that studies methods for building predictive computational models from observational data (Mitchell, 1997). The area is divided into various sub-topics that are largely related to the form of computational model employed, many of which are presently being applied to complex biological systems (Bernot et al., 2004; Calzone et al., 2006; Chen and Xu, 2004; Dale et al., 2010; Lin et al., 2012; Mazandu and Mulder, 2012; Tamaddoni-Nezhad et al., 2007; Xiong et al., 2006).

Modelling techniques can be broadly divided into three classes:

Logical: These representations (Bernot et al., 2004; Calzone et al., 2006; Lin et al., 2012; Tamaddoni-Nezhad et al., 2007) are typically discrete, with particular strengths in the ease with which models can be understood by domain experts. Existing applications include shape-oriented models of large and small molecules, as well as biochemical network representations.

Probabilistic: These include probabilistic descriptions of complex networks of interactions (Dale et al., 2010; Mazandu and Mulder, 2012), stochastic processes modelling individual molecular and cellular dynamics, and inferential procedures that analyse the underlying evolutionary process.

Mixed: These include approaches such as Bayesian networks (Chen and Xu, 2004) that support both probabilistic and logical interpretation. The advantage of such techniques is that they are readily understandable by domain experts, while still allowing the modelling of uncertainty. Applications include Bayes net models of gene regulation networks as well as complex hierarchical models that integrate transcriptomic, proteomic, metabonomic and phenotypic data.

1.5.1 Inductive logic programming (ILP) and abductive ILP (A/ILP)

Standard forms of machine learning, such as neural nets and support-vector technology, cannot make use of large-scale background knowledge within the learning process. By contrast, techniques from Inductive Logic Programming (ILP) support the inclusion of such background knowledge and allow the construction of hypotheses that describe structure and relationships between sub-parts. ILP (Muggleton, 1991; Muggleton and De Raedt, 1994) is the sub-area of machine learning concerned with inductive inference of logic programs. ILP systems use given example observations E and background knowledge B to construct a hypothesis H that explains E relative to B. The components E, B and H are each represented as logic programs. Since logic programs can be used to encode arbitrary computer programs, ILP is arguably the most flexible form of machine learning, which has allowed it to be successfully applied in a number of complex areas (Tsunoyama et al., 2008; Bohan et al., 2011; Santos et al., 2012).

The main role of abductive reasoning in machine learning of scientific theories is to provide hypothetical explanations of empirical observations (Flach and Kakas, 2000). Then, based on these explanations, we try to inject back into the scientific theory new information that helps complete the theory. This process of generating abductive explanations and updating theory can be repeated as new observational data become available. The process of abductive learning can be described as follows. Given a theory, T, that describes our incomplete knowledge of the scientific domain and a set of observations, O, we can use abduction to extend the current theory according to the new information contained in O. The abduction generates hypotheses that entail a set of experimental observations subject to the extended theory being self-consistent. Here, entailment and consistency

refer to the corresponding notions in formal logic. Abductive logic programming (Kakas et al., 1993) is typically applied to problems that can be separated into two disjoint sets of predicates: the observable predicates and the abducible predicates. In practice, observable predicates describe the empirical observations of the domain that we are trying to model. The abducible predicates describe underlying relations in our model that are not observable directly but can, through the theory T, bring about observable information. Hence, the hypothesis language (i.e., abducibles) can be disjoint from the observation language. We may also have background predicates (prior knowledge), which are auxiliary relations that help us link observable and abducible information.

In many implementations of abductive reasoning, such as that of Progol 5.0 used in this chapter (Muggleton and Bryant, 2000), the approach taken is to choose the explanation that 'best' generalises under some form of inductive reasoning. This link to induction then strengthens the role of abduction to machine learning and the development of scientific theories. We refer to this approach as Abductive ILP (A/ILP). Other frameworks of tight integration of abduction and induction include: ACL (Kakas and Riguzzi, 2000), CF-Induction (Inoue, 2001) and HAIL (Ray et al., 2003). Technically, we refer to induction as a process of taking a set of examples encoded as logical sentences that are free of variables and replacing them with more general hypotheses expressed as logically encoded sentences that contain universally quantified variables. By contrast, in abduction the hypotheses are also free of variables, and thus cannot be viewed as general rules since they do not contain universally quantified variables. A/ILP technology supports both abductive and inductive generalisation. In the present application we use an A/ILP system, Progol5.0, in abductive mode to construct food webs. This system is freely available for academic purposes.

1.5.2 Machine learning of biological networks using A/ILP
A/ILP has been used in a series of studies involving the revision of biological network models from example data. In the Metalog project (Tamaddoni-Nezhad et al., 2006), encoding and revising logical models of biochemical networks was done using A/ILP to provide causal explanations of rat liver cell responses to toxins. The observational data consisted of up- and down-regulation patterns found in high-throughput metabonomic data.

The approach used in Metalog was further extended in the CISBIC project (Sternberg et al., 2013), in which a mixture of linked metabonomic and gene expression data was used to identify biosynthetic pathways for bacterial

polysaccharides. In this case, ILP was shown to provide a robust strategy to integrate results from different experimental approaches.

1.6. Text mining

Text mining is the process of automatic extraction of information from textual resources using different techniques like information retrieval, machine learning, statistics and computational linguistics and natural language processing (NLP). There has been an increasing interest in using text mining in Bioinformatics and Systems biology (Harmston et al., 2010). In general, there are three main approaches (Cohen and Hunter, 2008): the first simply searches for the co-occurrence of concepts in the same textual unit (Jenssen et al., 2001); the second is related to rule-based systems (Blaschke et al., 1999) emphasising on the knowledge of language structure; and the third includes statistical and machine learning based systems (Cohen and Hunter, 2004) that generate classifiers operating on different levels of the text-mining process.

Although the applications of text mining are increasing rapidly, in most cases it can only be viewed as a tool to facilitate and help reveal relevant information hidden in a large volume of text data more efficiently and it cannot replace human in processing and understanding the text (Korhonen et al., 2012; Swanson, 1986): it is a useful means to an end, for processing and filtering huge volumes of information, and not necessarily an end in itself.

It is desirable to automate the corroboration of hypothetical trophic links, because manual corroboration of a large food web is difficult and requires significant amounts of time. The text-mining method described here can be categorised as a simple co-occurrence-based approach, but it can potentially be extended to more sophisticated approaches. To the best of our knowledge, this is the first attempt for automatic construction and corroboration of food webs from ecological data.

2. METHODS

2.1. Ecological data

2.1.1 The farm-scale evaluations data

The count data for Vortis and Pitfall-trap sampled invertebrate species comes from 66 spring-sown beet, 59 spring maize, 67 spring oilseed rape and 65 winter oilseed rape fields sampled as part of the farm-scale evaluations (FSE) of genetically modified, herbicide-tolerant (GMHT) crops (Bohan et al., 2005; Champion et al., 2003).

In late 1998, varieties of the crops, maize (*Zea mays* L.), beet (*Beta vulgaris* L.), spring oilseed rape and winter oilseed rape, or canola (*Brassica napus* L.) engineered to make them tolerant to broad-spectrum herbicides, had satisfied most of the regulatory requirements for commercial growing in the UK. The maize and oilseed rape varieties were resistant to the herbicide glufosinate-ammonium and the beet to glyphosate. It was argued that such varieties had the potential to allow greater flexibility in the timing of herbicide use (Dewar et al., 2003; Elmegaard and Bruus Pedersen, 2001; Firbank and Forcella, 2000), to facilitate the control of herbicide-resistant weeds (Gressel and Rotteveel, 2000) and to reduce reliance on persistent and relatively hazardous chemicals (Phipps and Park, 2002). The FSE project was established to test both whether GMHT varieties influenced the biodiversity of farmland relative to the management of conventional, non-GMHT, varieties and to what extent farmland biodiversity might change if GMHT crops were introduced commercially (DEFRA 2002; Firbank et al., 1999). For each crop, the FSE aimed to test the null hypothesis that there is no difference between the management of GMHT varieties and that of comparable conventional varieties in their effect on the abundance and diversity of arable plants and invertebrates.

The fields were spread across the geographical regions (Fig. 4.2) and conditions under which these four crops are grown commercially in mainland UK and each field was sampled for one cropping year (Firbank et al., 2003) between 2000 and 2004. Fields ranged in size from 2.7 to 70.8 ha, with an average of 11 ha, approximately 80% of which had hedgerows. Each field was split in half, and the two treatments (GMHT or conventional) were allocated at random to half-fields (Perry et al., 2003). Herbicide management was applied by the farmers to the conventional varieties of the crops at levels designed to achieve cost-effective weed control (Champion et al., 2003). Applications of herbicides varied between fields from 1 to 6 applications of either dicotyledon (broadleaved) or monocotyledon grass-specific herbicides or broad-spectrum herbicides against all weeds. Fields also received up to five applications of insecticide (including seed treatments) that could directly affect invertebrates.

### 2.1.2 Pitfall-trapping soil-surface-active invertebrates

Pitfall traps were used to survey populations of invertebrates active at the soil surface (after Luff, 1996) and consisted of 6 cm diameter plastic cups, sunk in the ground with the cup-lip level with the soil surface. Each was two-thirds filled with a 50:50 mixture of tap water and ethylene glycol as a preservative.
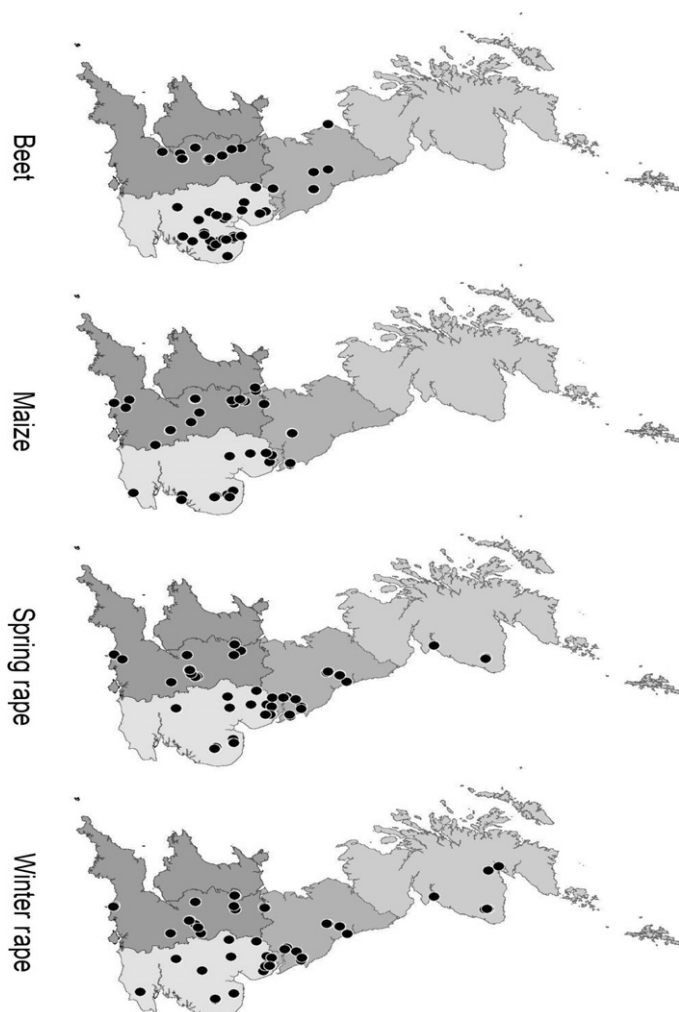
Figure 4.2 Distribution of the 66 spring-sown beet, 59 spring maize, 67 spring oilseed rape and 65 winter oilseed rape fields sampled as part of the farm-scale evaluations (FSE) of GMHT crops (Bohan et al., 2005; Champion et al., 2003).

Twelve traps were distributed across each half–field, with single traps positioned at 2, 8 and 32 m from the crop edge along 4 of 12 transects (Fig. 4.3). Transect numbers 1, 5, 7 and 12 were used for pitfall trapping, and these were chosen to avoid the proximity of other experimental protocols where possible. Trapping was conducted in the spring (April/May) and summer (June/July), and in the late summer (August) for the spring–sown crops and in the autumn (September/October), spring (April/May) and summer
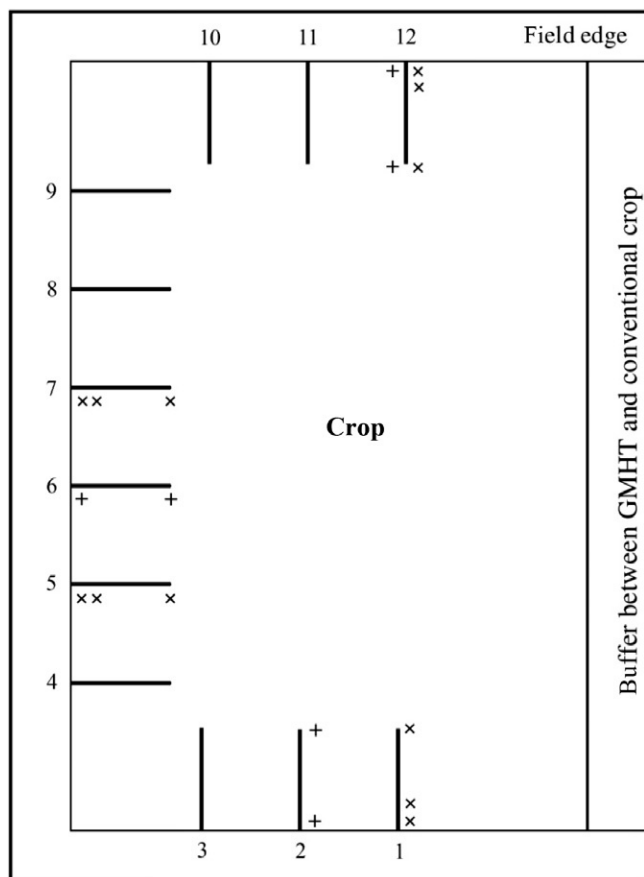
Figure 4.3 Locations of sample points in a standard half-field. Solid lines, transects numbered 1–12; +, Vortis suction sampling; x, Pitfall-trap sampling.

(June/July) for winter oilseed rape. On each occasion traps were opened for two weeks. These dates were chosen both to reflect the invertebrate phenologies and the timing of herbicide management. When returned to the laboratory, the samples were preserved by freezing or placement in 70% alcohol, before identification under a binocular dissecting microscope. All species of carabid beetles were counted and identified, as were five taxa of Araneae: the families Linyphiidae and Lycosidae, the genus *Pardosa*, *Erigone* agg. (consisting of *E. atra* and *E. dentipalpis*) and the species *Lepthyphantes tenuis*. Collembola and staphylinid beetles were counted at the family level. Nomenclature followed Lindroth (1974), Forsythe (2000), Speight et al. (1986) and Aukema (1990) for Carabidae; Fjellberg

(1980) for Collembola; Unwin (1988) for Staphylinidae; and Roberts (1993) for Araneae. Data were then pooled, by summation to give in each field a year-total estimate of the pitfall count of each species sampled.

2.1.3 Vortis suction sampling invertebrates on or around the weeds

In-field Vortis suction sampling for invertebrates living either on the weeds or on the underlying soil surface was conducted after Haughton et al. (2003). Five 10-sec suction samples, spaced 1 m apart, were taken at 2 and 32 m along transects 2, 6 and 12 (Fig. 4.3). Samples for each position were bulked together. Samples were taken in June and August for the spring-sown crops and on one occasion in the autumn (September/October) and one in the summer (May/June) for winter oilseed rape. Identification and counting of the invertebrates were done to the taxonomic levels specified by Haughton et al. (2003) and Roy et al. (2003) and followed Roberts (1993) for Araneae, Fjellberg (1980) for Collembola, Southwood and Leston (1959) for Heteroptera and Lindroth (1974) and Forsythe (2000) for Carabidae. The count data were then pooled, by summation to give in each field a year-total estimate of the Vortis count of each species sampled.

2.1.4 Trophic-functional type classifications for the invertebrates

Detailed measurements of life history and resource acquisition strategies are not available for many species sampled in the Vortis and pitfall protocols. Consequently, we classified the sampled invertebrates into 'trophic-functional types' using more general traits that reflect their functional, primarily resource acquisition and attributes. Following Hawes et al. (2009) and the simplifications made by Caron-Lormier et al. (2009), primary and secondary consumers were subdivided according to three characteristics:

1. *Trophic behaviour.* Phytophagous insects were divided into leaf chewers, sap feeders and pollen or seed feeders. Predators were defined as specialists (mono- and oligophagous species) or generalists (polyphagous species). The remaining groups were detritivores/fungivores, omnivores and mixed feeding groups (taxa that include both herbivorous and predatory species) and parasitic wasps.
2. *Body size.* Invertebrates were classified into four size classes (<4 mm, 4–5.9 mm, 6–8 mm and >8 mm) that are highly correlated with both prey selection behaviour and vulnerability to different predators.
3. *Activity pattern.* As with body size, activity level and period influences foraging behaviour and apparency. Activity level was divided into mobile or sessile categories. Mobile invertebrates were further separated

into the activity periods: diurnal, nocturnal and both. As the Vortis sampling was carried out during daylight hours only, invertebrates in the strictly nocturnal category were not found in this protocol. Sessile organisms such as aphids are difficult to place into an activity category because they are essentially inactive in their apterous form (at least as sampled in the FSEs).

The invertebrate taxa recorded in the Vortis and pitfall samples were assigned to these functional types using information from Southwood and Leston (1959), Roberts (1993), Luff (1978), Fjellberg (1980) and the Handbooks for the Identification of British Insects, Royal Entomological Society of London (see Appendix B). Although other general ecological traits, such as dispersal mechanism and reproduction were not included, these types reflect a range of feeding strategies important in considering trophic relations, and encompass the main types of invertebrates sampled in the FSE. It should be noted that this particular functional type representation is only one of many that could be conceived for the invertebrates.

2.2. Machine learning of probabilistic food webs from data

Machine learning, in the form of A/ILP, has recently been used to automatically generate a probabilistic food web from FSE data (Bohan et al., 2011; Tamaddoni-Nezhad et al., 2012a). Here, we review the machine learning approach for generating probabilistic food webs from data and extend the initial study by: (1) presenting new food webs generated from pitfall and merged Vortis and pitfall data; (2) describing a new approach for cross-validation of food webs across different crops; and (3) describing a novel approach for learning trophic links between functional groups and evaluating new functional food webs learned from Vortis data.

2.2.1 Abduction of trophic relations from FSE data

The FSE data include information about the change in the abundance of invertebrates between the current, conventional herbicide management of spring-sown maize (M), beet (B) and oilseed rape (R) and winter-sown oilseed rape (W), and the herbicide management of GMHT varieties of the same crops using a split-field design. We used FSE data to measure a treatment effect ratio: counts from each conventional and GMHT half-field pair were converted to a geometric treatment ratio, as used in Haughton et al. (2003). Counts were log-transformed, using formula $L_{ij} = \log_{10}(C_{ij} + 1)$, where C_{ij} is the count for a species or taxon in treatment i at site j. Sites where $(C_{1j} + C_{2j}) \leq 1$ were removed from the learning dataset (as in

Haughton et al., 2003). The treatment ratio, R, was then calculated as $R = 10^{d}$ where $d = (L_{2j} - L_{1j})$. Following the rationale in Squire et al. (2003), important differences in the count between the two treatments were considered to be greater than 50%. Thus, treatment ratio values of $R < 0.67$ and $R > 1.5$ were regarded as important differences in count with direction of *down* (decreased) and *up* (increased) in the GMHT treatment, respectively. This information on *up* and *down* abundances has been considered as our observational data for the learning that can be represented by predicate *abundance(X, S, up)* (or *abundance(X, S, down)*) stating the fact that the abundance of species X at site S is *up* (or *down*).

The existence of a trophic relationship between species is the knowledge gap that needs to be filled by machine learning. This can be represented by an abducible predicate *eats(X, Y)* capturing the hypothesis that species X eats species Y. It is clear that this problem has properties that require an abductive learning approach such as A/ILP: firstly, the given background knowledge is incomplete; and secondly, the problem requires learning in the circumstance in which the hypothesis language is disjoint from the observation language.

In order to use abduction, we also need to provide the rules that describe the observable predicate in terms of the abducible predicate. An example of such a rule is shown below:

abundance(X, S, up) **if**
    predator(X) and
    co_occurs(S, X, Y) and
    bigger_than(X, Y) and
    abundance(Y, S, up) and
    eats(X, Y).

Similarly, a rule for *abundance(X, S, down)* can be defined. This rule expresses the inference that following a management-driven perturbation in the ecosystem, the increased abundance of species X at site S can be explained by the fact that (1) species X eats species Y, and (2) the increased abundance of species Y. This rule also includes additional conditions to constrain the search for abducible predicate *eats(X, Y)*. These constraints are (1) X should be a predator, (2) X and Y should co-occur, and (3) X should be bigger than Y. Predicates *predator(X)* and *bigger_than(X, Y)* are provided as part of the background knowledge and *co_occurs(S, X, Y)* is compiled directly from FSE data. This model describes at an appropriately high level the possible transitive effect of management leading to increased (or decreased) abundance of species.

Given the model described above and the observational data, a set of abduced hypotheses, in the form of '*eats*' relations between species, can be
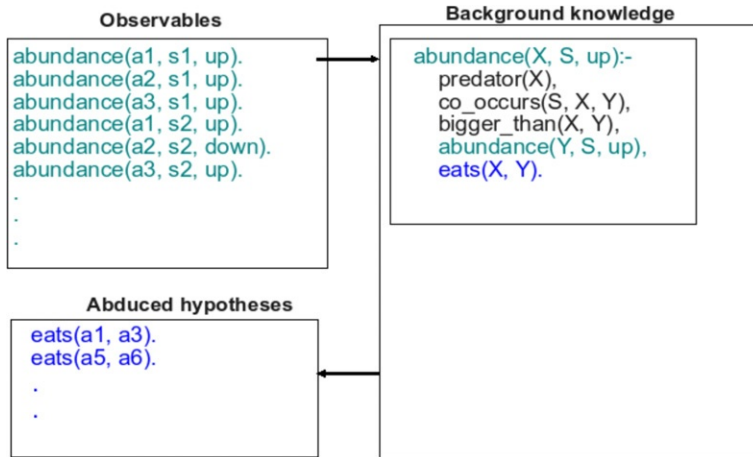
Figure 4.4 Machine learning of trophic links from ecological data using abductive ILP.

generated using A/ILP as shown in Fig. 4.4. An A/ILP system such as Progol 5 (Muggleton and Bryant, 2000) can generate abductive hypotheses by matching observable input against the background knowledge (which includes rules describing the observable predicate in terms of abducible predicate). In general, many choices for matching could be made, leading to a variety of alternative hypotheses and a preference is imposed by Progol 5 using an information–theoretic criterion known as compression (Muggleton and Bryant, 2000). Here, compression can be defined as $p–n–h$, where p is the number of observations correctly explained by the hypothesis, n is the number incorrectly explained and h is the length of the hypothesis (e.g., 1 for a single fact such as a trophic link).

As shown in Fig. 4.4, the output of the learning is a set of abduced hypotheses, in the form of '*eats*' relations between species. This can be visualised as a network of trophic links (the food web) in which the abductive hypothesis that a particular species a eats a particular species b (*eats(a, b)*) is represented by a trophic link from b to a. Figure 4.5 shows the trophic network hypothesised by A/ILP from Vortis sampled invertebrates in the FSE dataset as appeared in Bohan et al. (2011) and Tamaddoni–Nezhad et al. (2012a). Each link between a species or taxon represents a learnt '*eats*' relationship that could be tested either against the literature or by experimentation. The thickness of the link indicates the estimated probability of occurrence, based on the relative frequency from 10 random permutations
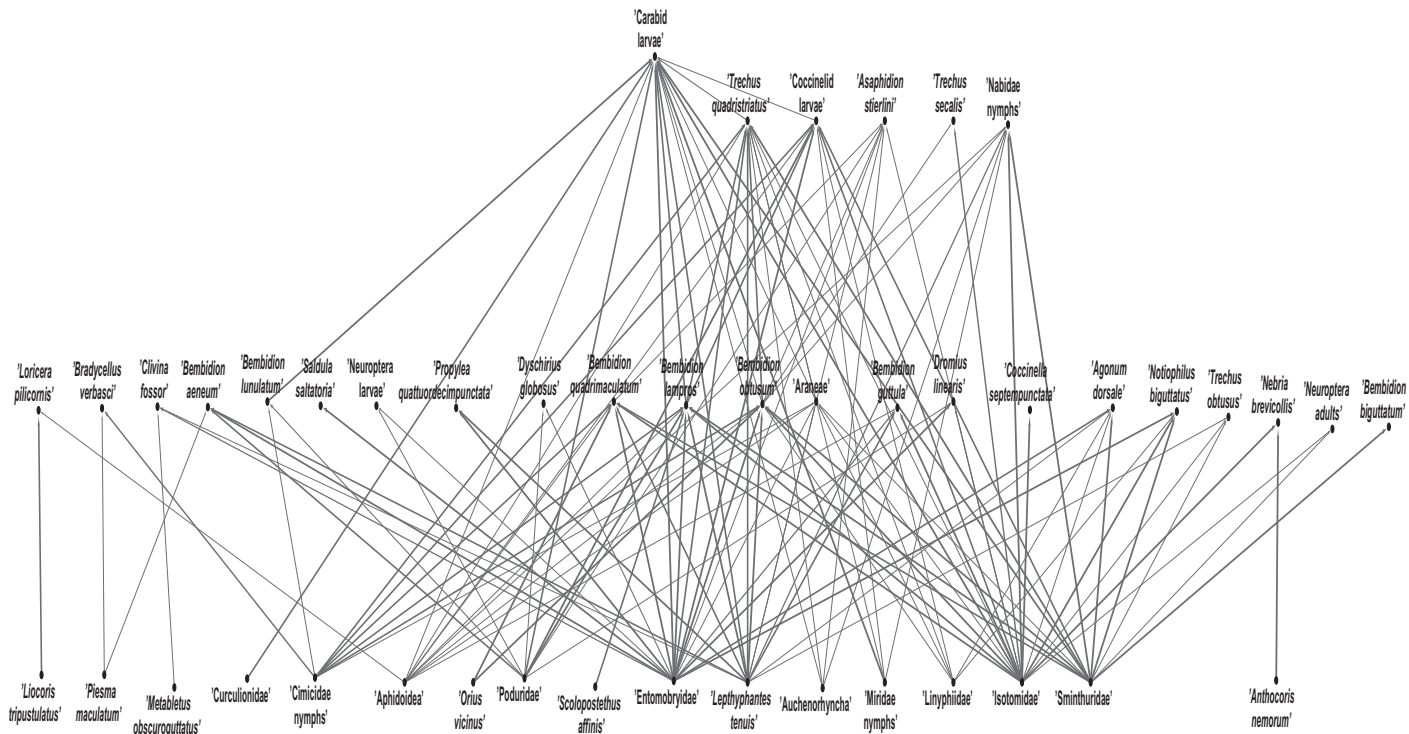
Figure 4.5 Hypothetical trophic network (food web) constructed by machine learning from Vortis sampled invertebrates in the FSE data. Thickness of trophic links represents probabilities which are estimated using HFE (Tamaddoni-Nezhad et al., 2012a).

of the training data. This probability estimation method is called hypothesis frequency estimation (HFE).

2.2.2 Hypothesis frequency estimation

We use hypothesis frequency estimation (Tamaddoni–Nezhad et al., 2012a) to estimate probabilities for hypothetical trophic links based on their frequency of occurrence when randomly sampling the hypothesis space. HFE is a Probabilistic ILP (PILP) technique that is based on direct sampling from the hypothesis space. In some ILP systems, including Progol 5.0, training examples also act as seeds to define the hypothesis space (e.g., a most specific clause is built from the next positive example). Hence, different permutations of the training examples define different parts of the hypothesis space. We use this property to sample from the hypothesis space by random permutations of the training data. Probability of ground hypotheses can be estimated based on the frequency of occurrence when random permutations of the training data (and hence different seeds for defining the hypothesis space) are considered. Using this technique, the thickness of trophic links in a probabilistic food web represents probabilities that are estimated based on the frequency of occurrence from random permutations of the training data (e.g., see Fig. 4.5). A probabilistic trophic network can be also represented using standard PILP representations such as Stochastic Logic Programs (SLPs) (Muggleton, 1996) or ProbLog (De Raedt et al., 2007). For this we can use relative frequencies in the same way probabilities are used in PILP. We can then use the probabilistic inferences based on these representations to estimate probabilities. For example, the probability $p(abundance(a, s, up))$ can be estimated by relative frequency of hypotheses that imply a at site s is up. Similarly, $p(abundance(a, s, down))$ can be estimated and by comparing these probabilities we can decide to predict whether the abundance is up or down.

This method has been used in the leave-one-out cross-validation experiments in (Tamaddoni–Nezhad et al., 2012a) to measure the predictive accuracies of probabilistic tropic networks. In this chapter, we use HFE together with a new cross-validation approach that measures predictive accuracies across different crops.

2.2.3 Cross-validation of food webs across different crops

In this chapter, we use a new cross-validation test strategy to measure the predictive accuracies of probabilistic trophic networks. This test strategy is similar to the leave-one-out test described in Tamaddoni–Nezhad et al. (2012a), however, in the new approach we cluster the data across different

crops. In this cross-validation method, we evaluate a food web constructed from a set of crops on unseen data from a different crop. The new cross-validation method can be summarised as follows:

    For each crop c = 'B', 'M', 'S' and 'W' do

        For each (predator) species s

            1) $Test_{cs}$ = the abundance of species s in sites for crop c (leave-out test data)

            2) Generate food web N_{cs} using the data from all crops excluding $Test_{cs}$

            3) A_{cs} = Predictive Accuracy of N_{cs} on $Test_{cs}$

        End

    End

The average of predictive accuracies A_{cs} for a particular crop 'c' show how well the food webs generated from all crops except crop 'c' can predict the data for crop 'c'. By plotting the average predictive accuracies A_{cs} for all species and crops we can generate the learning curves. Predictive accuracy is defined as the proportion of correctly predicted left-out test examples. This cross-validation method is used in Section 3.3.

2.2.4 Machine learning of functional food webs

The limitations of the species-based approaches and the need for considering functions were discussed in Section 1.2. In particular, the trophic interactions between functional groups of species are important in functional ecology for predicting changes in agro-ecosystem diversity and productivity (Caron-Lormier et al., 2009; Duffy, 2008). Here, we extend the initial study in Bohan et al. (2011) to also learn and evaluate functional food webs directly from the same data used in the initial study to learn species-based food webs, by assuming that the functional types of species are given as part of background knowledge. In this new setting, the abducible predicate *eats(A, B)* is defined between functional groups rather than (individual) species. The observable predicate is the same as before: *abundance(X, S, up)* (or *abundance(X, S, down)*) indicates that the abundance of X at site S is *up* (or *down*). Given this information and the background information on functional type of each species, trophic networks for functional groups can be learned using a similar approach to the one used for learning species-based food webs, as described in the previous sections. We also need a rule that describes the observable predicate in terms of eats relation between functional groups as shown below:

    *abundance(X, S, up) **if***

        predator(X) and

        co_occurs(S, X, Y) and

        bigger_than(X, Y) and

> *abundance(Y, S, up) and*
> *ft(X, XFunc_ID) and*
> *ft(Y, YFunc_ID) and*
> *eats(XFunc_ID, YFunc_ID).*

This rule is similar to that used for learning species-based food webs, except that instead of defining the abducible predicate *eats* on the species *X* and *Y*, it is defined on the functional types for *X* and *Y*, that is, *XFunc_ID* and *YFunc_ID*. Functional types of species are defined by the predicate *ft* that is given as part of the background knowledge. As discussed in Section 2.1, species can be classified into 'trophic-functional types' using general traits that reflect their functional, primarily resource acquisition and attributes.

As in the previous setting for learning species-based food webs, an A/ILP system such as Progol 5 can generate abductive hypotheses in the form of *eats* relation between some functional types, such that a positive compression over data is provided by the suggested hypotheses.

2.3. Automatic corroboration of trophic links using text mining

2.3.1 Text mining and literature networks

The hypothetical Vortis food web proposed by machine learning (see Fig. 4.5) was examined manually and comparison with the literature showed that many of the links are corroborated (Bohan et al., 2011). Figure 4.6 shows manual corroboration for some prey (columns) and predator (rows) species combination in Fig. 4.5. Each pairwise hypothesised link has a strength (i.e., frequency between 1 and 10) followed by references (in square brackets) in the literature (listed in Appendix A) supporting the link. This table is a subset of the manual corroboration presented in Bohan et al. (2011) and only prey/predators are shown that have at least one link with strength more than or equal to 7. This table shows that many of the links, suggested by the model, are corroborated by the literature. In particular, links in the model ascribed with high frequency correspond well with those having multiple references in the literature. For example, there are 15 links with more than 2 references and 8 of these are with frequency 10 and from these all the 3 links with 3 references (marked by green circles) have frequency 10. In addition, there are also highly frequent links with no references in the literature, and these could potentially be novel hypotheses for future testing with targeted empirical data.

A manual corroboration is therefore useful to confirm some of hypotheses and also to identify potential novel hypotheses. However, manual

	Anthocoris nemorum	Bembidion lampros	Bembidion lunulatum	Bembidion obtusum	Cimicidae nymphs	Curculionidae	Entomobryidae	Isotomidae	Lepthyphantes tenuis	Liocoris tripustulatus	Miridae nymphs	Orius vicinus	Poduridae	Scolopostethus affinis	Sminthuridae
Agonum dorsale							9 [13]	4	5 [6]						10 [13]
Bembidion aeneum							10 [11]		9 [6]			9 [11]			
Bembidion biguttatum															10 [11]
Bembidion guttula							7 [11]	10 [11]	9 [6]						
Bembidion lampros					9		10 [11]	10 [11]	10 [6]				10 [11]	9	10 [11]
Bembidion obtusum					9		10 [11]	10 [11]	10 [6]			10			10 [11]
Bembidion quadrimaculatum					9		10 [11]	10 [11]	9 [6]			10			9 [11]
Bradycellus verbasci					8										
Carabid larvae		9	10	3		9	10 [2]	10 [2]	9		10		9 [2]		10 [2]
Clivina fossor							7 [12,5]	2							
Coccinelid larvae		9		9			10 [14,16]	9 [14,16]			1		10 [14,16]		9 [14,16]
Coccinella septempunctata							10 [14,16]								
Dromius linearis							10 [1]	7					3		
Loricera pilicornis										9					
Nebidae nymphs				3			10	7				2 [10,7]			10 [9]
Nabria brevicollis	10							10 [15]							
Notiophilus biguttatus							10 [4,11,3]	10 [4,11,3]							10 [4,11,3]
Propylea quattuordecimpunctata							10 [14,16]		10 [16]						
Saldula saltatoria							10 [8]								
Trechus quadristriatus		9		9	9		9 [15,3]	4 [15,3]	9 [6]			2			10 [15,3]
Trechus secalis				2				8 [15,3]							

Figure 4.6 Manual corroboration of trophic links for some prey (columns) and predator (rows) species combination from Fig. 4.5. Each pairwise hypothesised link has a strength (i.e., frequency between 1 and 10) followed by references (in square brackets) in the literature (listed in Appendix A) supporting the link. Multiple references are indicated by yellow and green circles and potential novel hypotheses by dashed red circles.

corroboration of hypothetical trophic links is difficult and requires significant amounts of time. Here, we demonstrate how a text-mining technique can be adopted for automatic corroboration of hypothetical food webs from ecological resources. This is particularly useful for larger food webs such as the pitfall food web and merged Vortis and pitfall food web that we consider in this chapter (see Figs. 4.8 and 4.11).

Many existing text-mining techniques, as used in different domains, work by relating given terms based on their co-occurrence in the literature. For example, PubGene (Jenssen et al., 2001) is a public web service that can generate a network of possible relationships between biomedical terms based on their co-occurrence in medical texts from MEDLINE database. In a *literature network* like this the co-occurrence of a pair of entities (e.g., genes) is shown by an edge annotated by the number of papers where the co-occurrences have been found.

2.3.2 Generating literature network for the species in a food web

The text mining of trophic links used in our project was inspired by PubGene and other similar text-mining tools that are based on co-occurrences: we search for co-occurrences of pairs of ecological species that are related by a trophic interaction lexicons, for example, eat, feed, prey and consume.

Figure 4.7 shows manual corroboration versus automatic corroboration using text mining, illustrating how a literature network can be generated automatically based on the co-occurrences of predators/prey in the relevant context. The pairs of predators/prey species (from a given food web) and the interaction lexicons (from a dictionary file) are used to generate *queries*. Then the text-mining module searches through the text of available publications to match each query. The publications can be in a local database or accessed via a search engine (e.g., Google Scholar). The output of the text mining for each query is the number of publications that matched that query. The output for a whole food web can be represented by a literature network in which the number associated with each edge is related to the number of papers where the co-occurrences of the predator/prey species have been found with at least one trophic interaction lexicon (eat, feed, prey, or consume).

We refer to the numbers of references generated by this approach as Literature Hits (LH). LH is not as accurate as manual corroboration; however, it would be a useful estimate for comparison between different links. In order to normalise the number of co-occurrences for a pair of species we
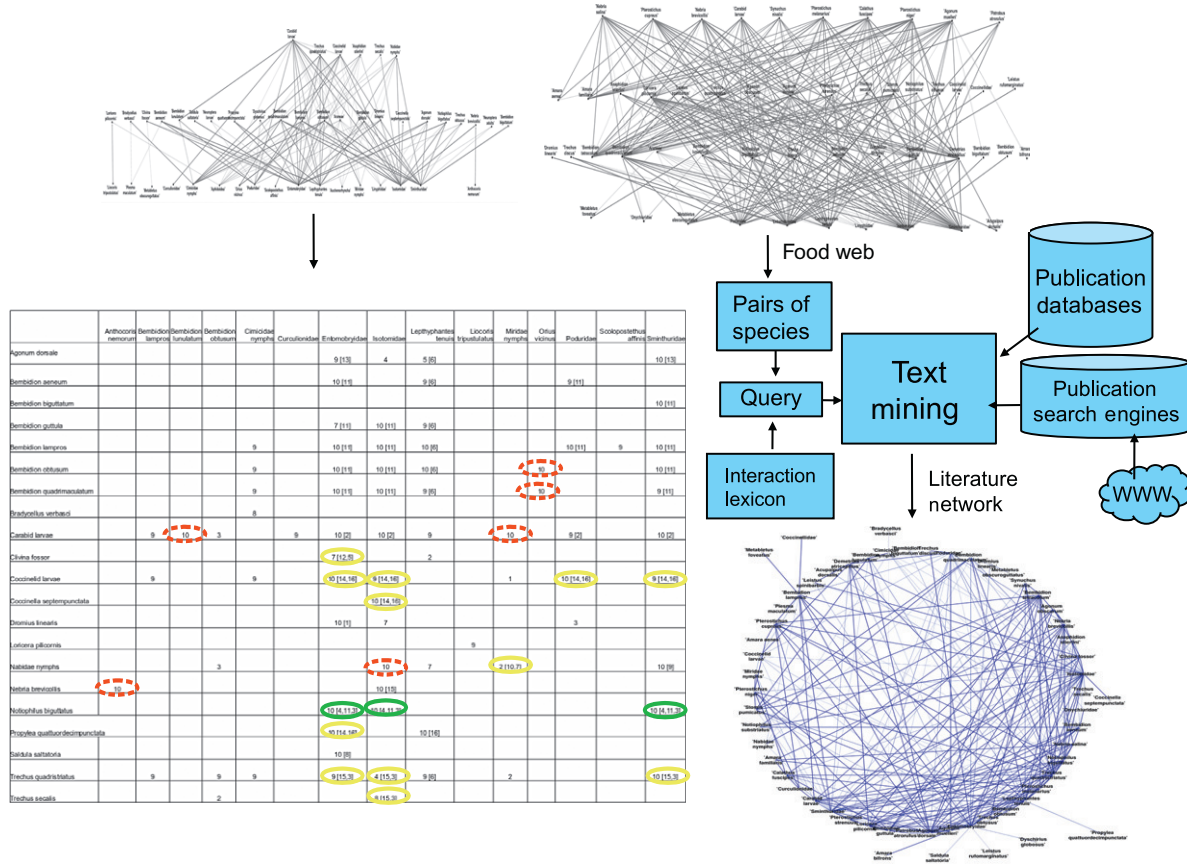
Figure 4.7 Manual corroboration versus automatic corroboration. In automatic corroboration, a literature network is automatically generated for a food web using text mining from publications.

divide it by the sum of the occurrences of each individual species extracted from the same sources. We refer to this as text-mining ratio (TMR). The text-mining module is implemented as a program that reads queries from a file and returns number of references, that is, the LH.

There are many publication databases for gene and protein interactions, which has been used for text-mining purposes. However, the databases that can be used for trophic interactions are very limited. We examined several general publications databases including Web of Knowledge,[1] Web of Science[2] and Cab Direct.[3] However, to gain access to a larger and more current set of publications, the search engine used in this study is Google Scholar.[4] Unlike most of other available publication databases, this method allows us to have access to full text as well as abstract and keywords. The numbers of hits is therefore based on the co-occurrences in the abstract and full text (whenever available).

3. RESULTS

3.1. Pitfall versus Vortis food webs

The pitfall dataset is larger than that for the Vortis samples and therefore the newly learned food webs are significantly larger than the initial webs. Figure 4.8 shows the food web resulted from pitfall data. Thickness of trophic links represent probabilities that are estimated based on the frequency of occurrence from 10 random permutations of the training data using the HFE approach described in Section 2. The food web constructed from pitfall data is more complex than the Vortis food web. While the number of nodes is very close to the number of nodes in the Vortis food web (51 vs. 48), the number of trophic links in pitfall food web is more than two times the number of links in Vortis food web (318 vs. 137). In other words, more trophic interactions can be learned from pitfall data.

This increase in link density could be partly explained by the different sample coverage of the two methodologies. The Vortis only samples those invertebrate species present on the surface of plants or the soil at the time of sampling. The protocol therefore only samples day-active species that are small or weak enough to be sucked up. As open traps, the pitfall traps can

[1] http://www.webofknowledge.com
[2] http://scientific.thomson.com/products/wos/
[3] http://www.cabdirect.org/
[4] http://scholar.google.com
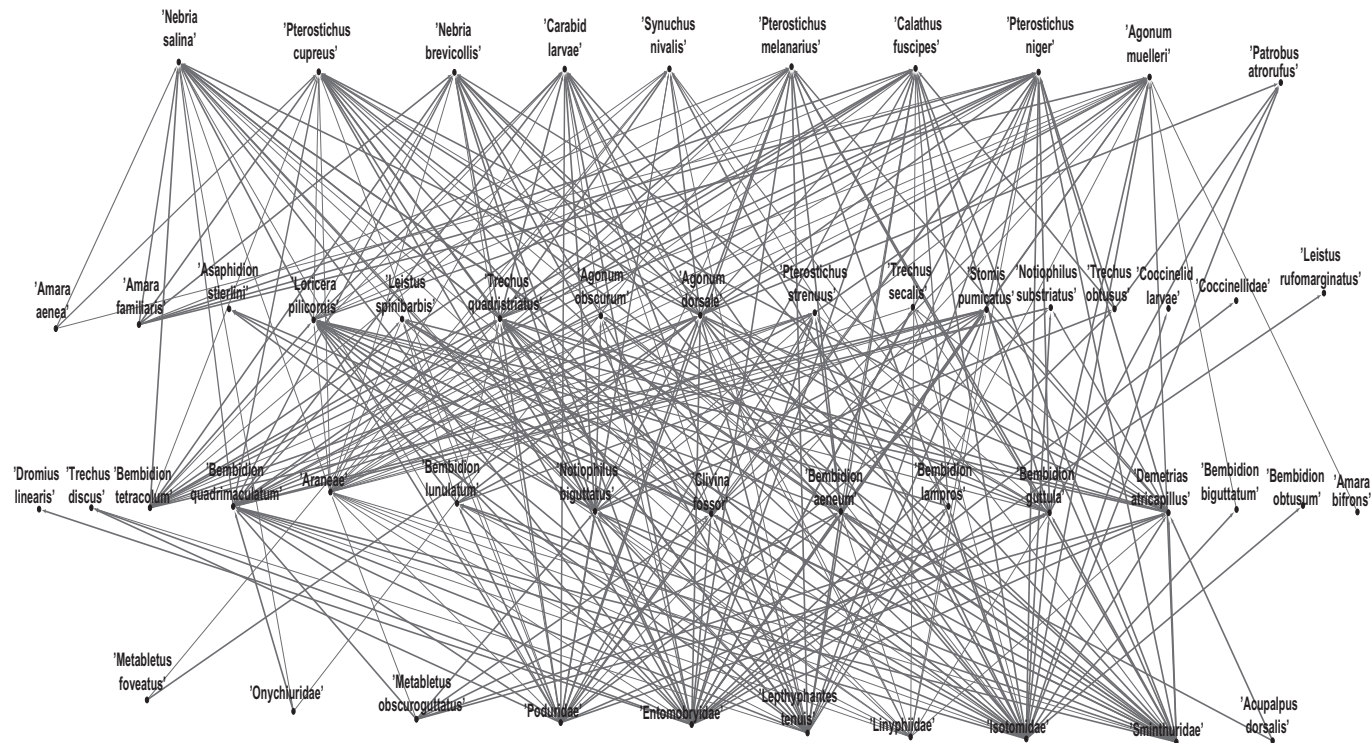
Figure 4.8 New hypothetical food web constructed by machine learning from pitfall data.

sample much larger species and this sampling takes place over a much longer period, while the traps are open, collecting both day- and night–active species and increasing overall sample diversity.

Figure 4.9 shows the frequencies of common links from Vortis versus pitfall in a scatter diagram. Figure 4.10 shows a food web constructed by common links from Vortis and pitfall. The Pearson correlation between Vortis and pitfall frequencies is 0.773: thus the frequencies of common links from Vortis and pitfall data are correlated, which suggest that the method used to generate the trophic links produces compatible results using data from two different sampling methods and is therefore robust. It also suggests that the Vortis and pitfall data are sufficiently compatible to be merged together for learning a larger food web.
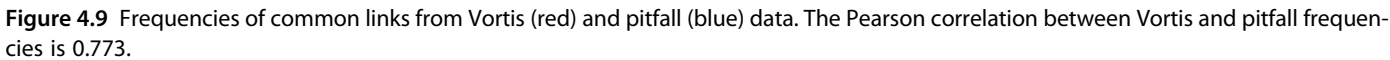
The food web learned from merged Vortis and pitfall data is shown in Fig. 4.11 and has 72 nodes and 407 edges. Note that this food web is not necessarily the same as that constructed by merging trophic links from Vortis and pitfall food webs. For example, machine learning can suggest trophic links between species present in one data as predator and species sampled in the other as prey: these trophic relations cannot be captured by merging trophic links from Vortis and pitfall food webs. From a machine learning point of view, one would also expect some sort of sharpening and convergence of the previous hypotheses as the observational data increase.

3.2. Functional food webs

Figure 4.12 shows trophic network constructed by learning trophic interactions between functional groups from Vortis data. In this food web, each functional group is represented by a species that can be viewed as an archetype for the functional group as in Caron-Lormier et al. (2009).

The functional food web appears to bear all the hallmarks we would expect of the species-based food web: the detritivores and many small herbivores sit at the base of the structure while predators of various sizes sit at higher trophic levels, capped by the generalist functional group to which the large carabids belong. The spiders appear to reside at the base of the web and while these animals are small, they are predators and at least at face value they would appear to be in a rather unexpected position. Recent work on predation on spiders, however, has suggested that this hypothesised position in an animal–animal network is correct (Davey et al., 2013), at least for part of the agricultural season.

Figure 4.9 Frequencies of common links from Vortis (red) and pitfall (blue) data. The Pearson correlation between Vortis and pitfall frequencies is 0.773.
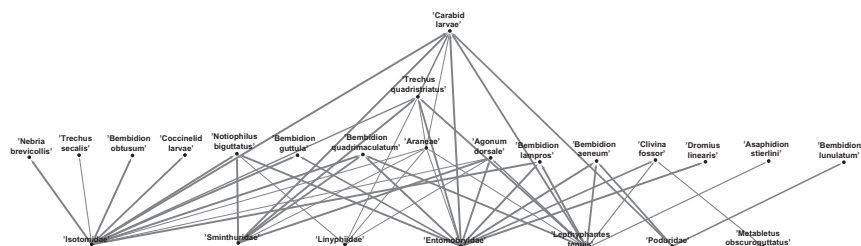
Figure 4.10 Common trophic network from Vortis and pitfall data.

## 3.3. Cross-validation of species and functional food webs across different crops

Evaluating food webs constructed from a set of crops on unseen data from a different crop was done by repeatedly constructing food webs from all crops data, excluding test data from a particular crop 'c' and measuring the predictive accuracy on these test data. Figure 4.13A and B shows predictive accuracies of Vortis species-based and functional food webs on different crops. The average predictive accuracies (the proportions of correctly predicted left-out test examples) are reported with standard errors associated with each point. Due to different default accuracies for different crops, the learning curves start from different levels of accuracies. However, the overall increase in the average accuracies (when 0–100% of the training examples are provided) is around 16 for crops M and S and around 18 for crops W and B in the species-based food webs and around 16 for crops S and W and around 20 for crops M and B in the functional food webs. By averaging the predictive accuracies over different crops, we get overall predictive accuracies for species and functional food webs, as shown in Fig. 4.13C. According to this figure, in all cases the predictive accuracies were significantly higher than the default accuracy of the majority class (55.6% for Vortis data).

Predictive accuracies for the functional food webs were the same or higher than their species-based counterpart, particularly at low to medium size of training examples. This suggests that the functional food webs are at least as accurate as their species-based counterpart, despite being much more compact. The apparent decline in the predictive accuracy for the functional food webs after the 50% point could be related to overfitting the training data.

Figure 4.14A–C shows similar learning curves for pitfall food webs and compare predictive accuracies of species and functional food webs when different percentage of the training data is provided. As shown in the graphs, the differences between the default accuracies for different crops are less than those from Vortis data, which makes the comparison easier. As for the Vortis data, the food web tested on crop B appear to have the highest overall
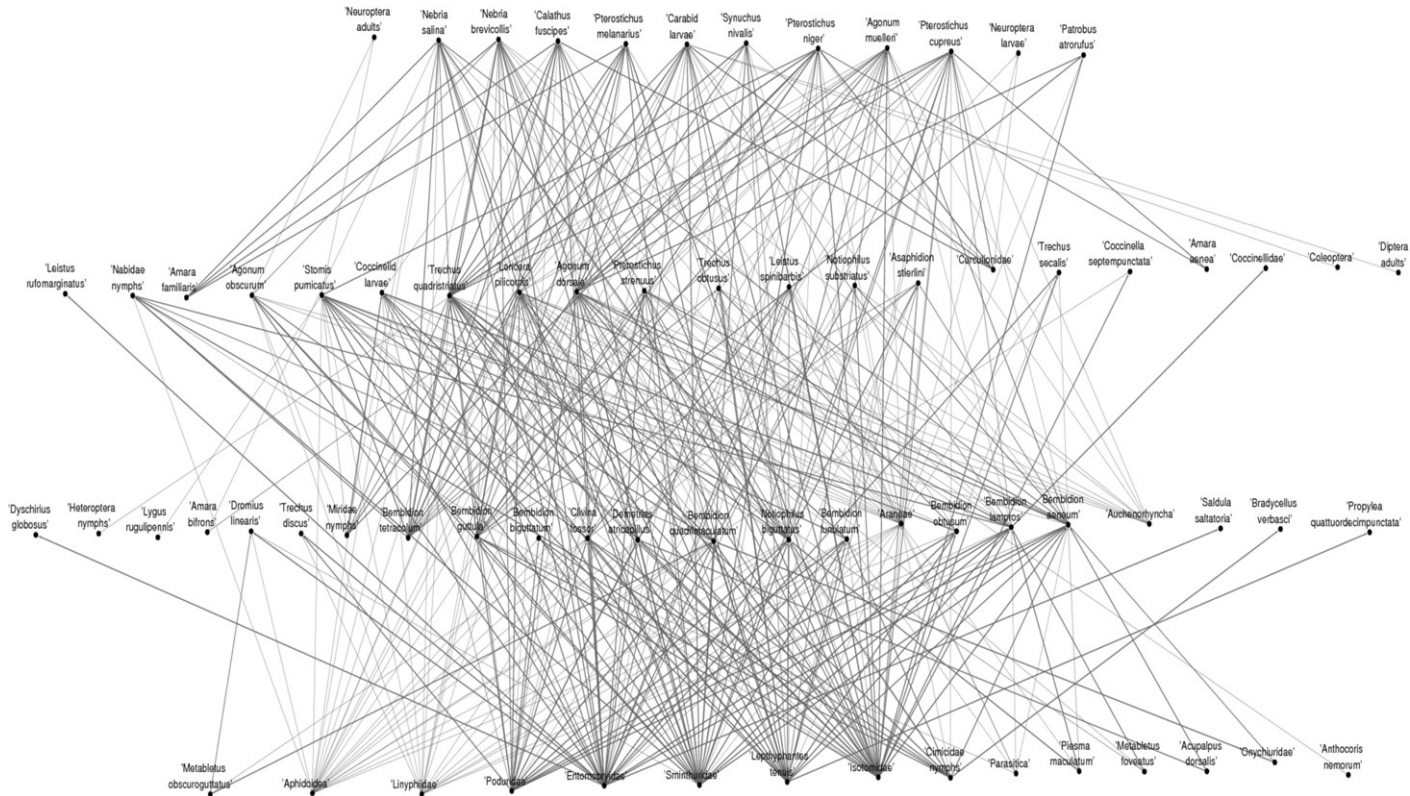
Figure 4.11 Trophic network from merged Vortis and pitfall data. The information in this probabilistic food web is also given in tabular format in Appendix A.
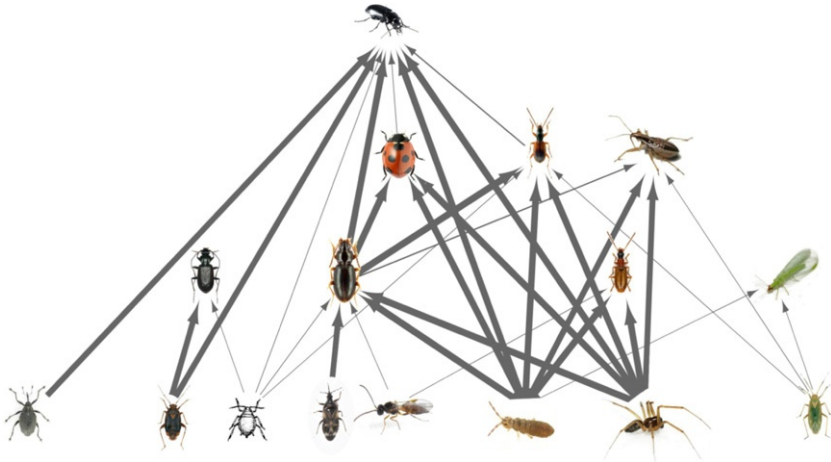
Figure 4.12 Functional food webs learned from Vortis data. Thickness of trophic links represents probabilities estimated using HFE.

increase in the predictive accuracies, both in the species–based and functional food webs (Fig. 4.14A and B). Figure 4.14C suggests that the pitfall functional food webs closely follow their species–based counterpart, confirming what was observed for Vortis food webs, that is, functional food webs can accurately estimate species–based food webs.

A comparison between Figs. 4.13A and 4.14A shows that the pitfall food webs have a higher overall increase in predictive accuracy (19 vs. 16), which suggest a relatively higher accuracy for pitfall food webs. This difference probably stems from the differences between the two protocols: the pitfall dataset is larger than the Vortis dataset, and it contains both day- and night–active species, resulting in a food web with more edges.

It would be interesting to test how sensitive the functional food webs are with respect to the particular choice of functional groups: we are currently working on a new approach aimed at learning trophic links as well as functional groups from data, using predicate invention in a meta–interpretive learning setting (Muggleton et al., 2013) to learn the best functional grouping from data and does not require a pre–specified grouping.

3.4. Automatic corroboration of learned trophic links
3.4.1 Common trophic network
Here we use the automatic corroboration approach described in Section 2.3 and provide the pairs of trophic interactions from the common trophic interaction as input to generate a literature network. Figure 4.15 shows a
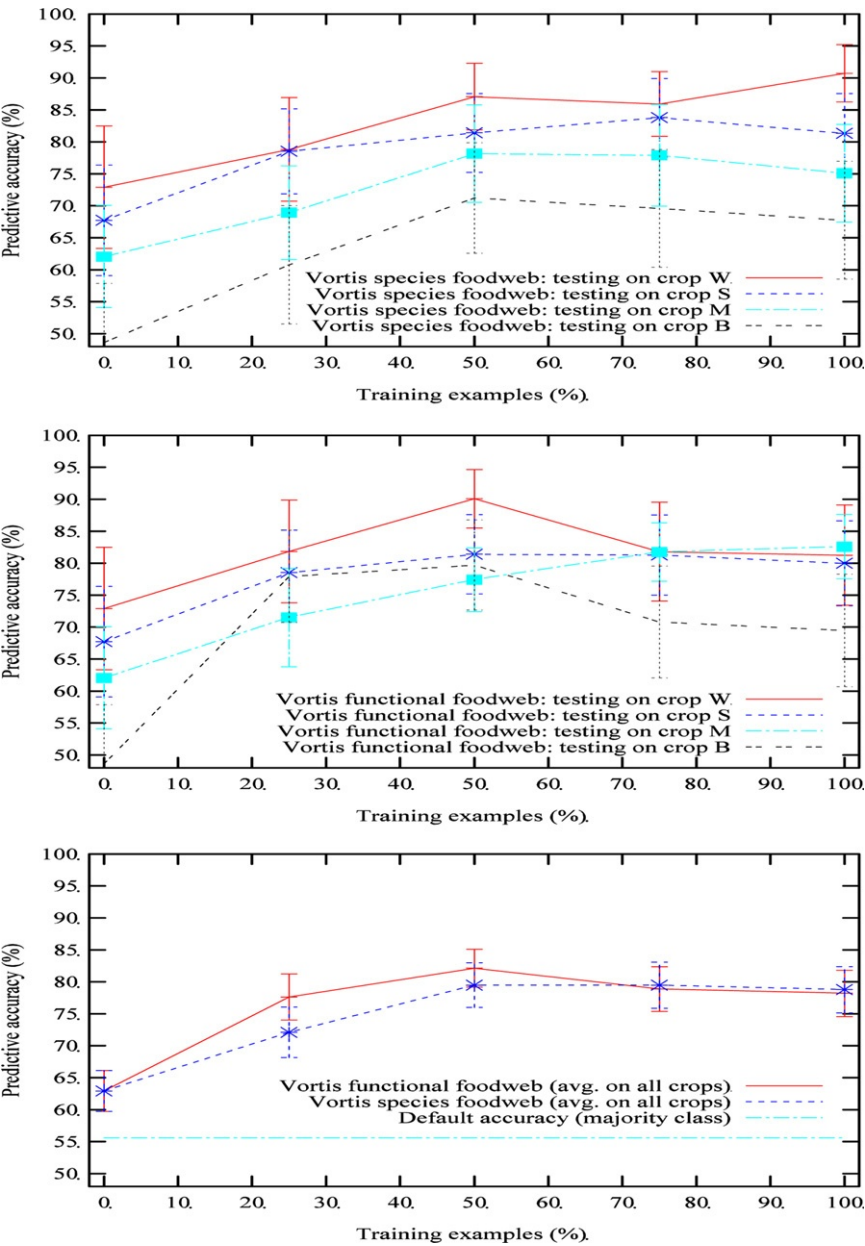
Figure 4.13 Learning curves for Vortis probabilistic food webs: (A) species food web and (B) functional food web on different crops; (C) overall species and functional food webs.
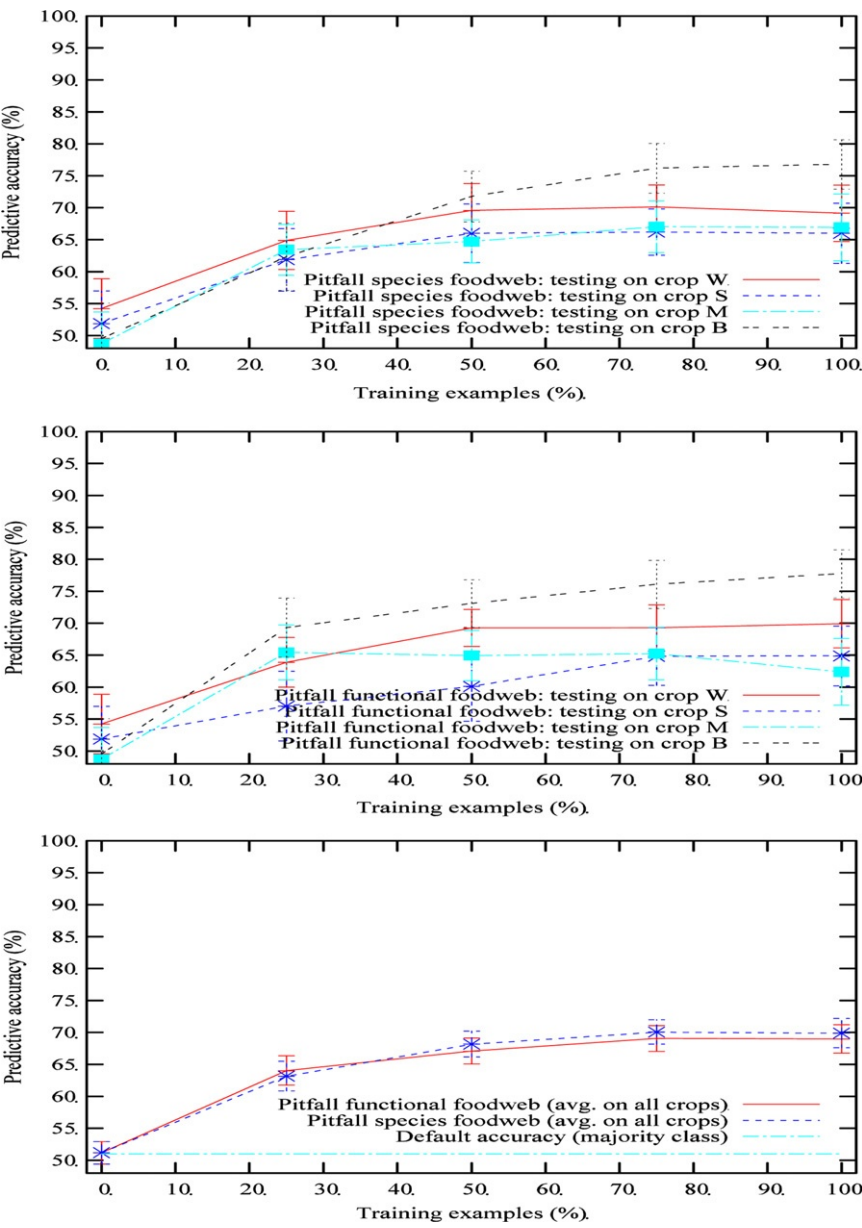
Figure 4.14 Learning curves for pitfall probabilistic food webs: (A) species food web and (B) functional food web on different crops; (C) overall species and functional food webs.
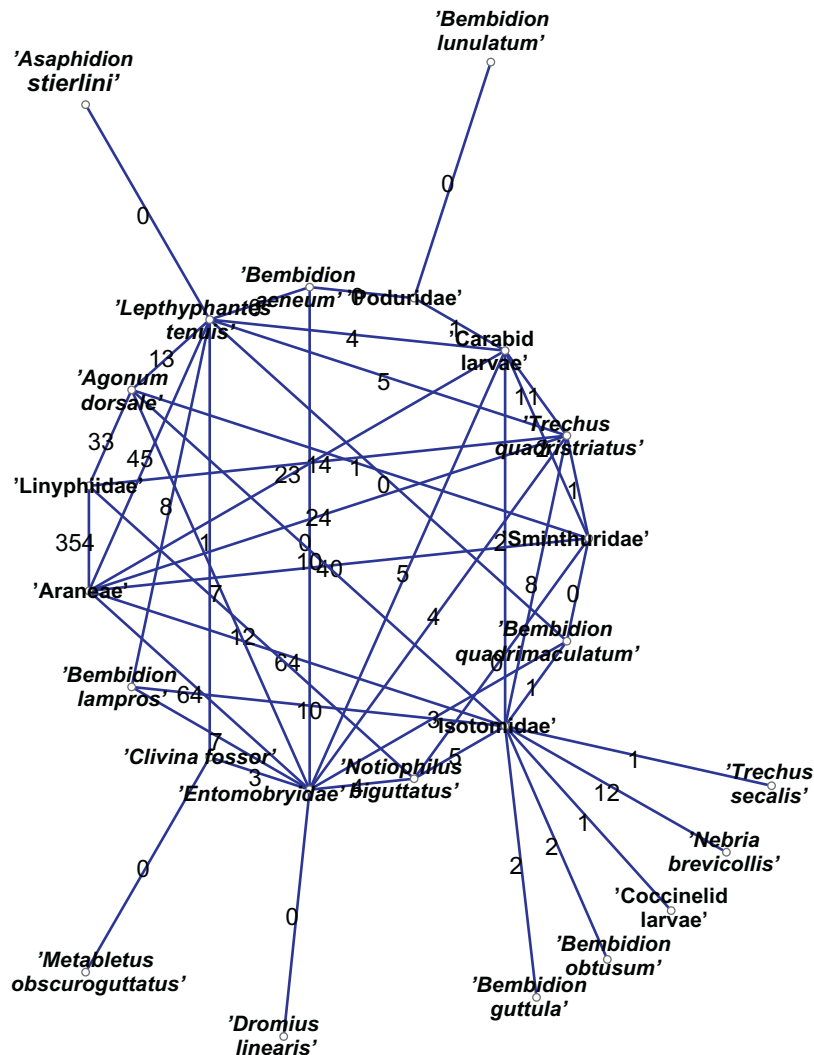
Figure 4.15 A literature network for common trophic links from Vortis and pitfall data. The number on each edge represents the number of papers where the co-occurrences of species have been found with a trophic interaction lexicon.

literature network generated for the common trophic links from Vortis and pitfall data. The number on each edge represents the number of papers where the co-occurrences of species have been found with at least one trophic interaction lexicon (eat, feed, prey or consume). These numbers are based on the co-occurrences in the abstract and full text (whenever available) using Google Scholar as the search engine for the text mining. In order to

normalise the number of co-occurrences, we divide them by the sum of the occurrences of each individual species in the pair to compute the TMR.

Here, we test whether the TMRs from the literature network in Fig. 4.15 are correlated with: (1) the frequencies of learned trophic links (HFE); and (2) the number of references from the manual corroboration used by Bohan et al. (2011).

First, we test if the frequencies of common trophic links and the TMR for these links are correlated using the probabilistic food webs generated from Vortis and pitfall data and considering the frequencies of links common to both methods. As in Tamaddoni-Nezhad et al. (2012a), we removed species that only have weak links (Araneae and Linyphiidae). We also use the literature network generated for the common trophic links from Vortis and pitfall data as shown in Fig. 4.15. This includes the number of co-occurrences of a pair of species in the common trophic network. Figure 4.16 shows the correlation between the hypothesis frequencies for common Vortis links and the TMR. Spearman's correlation ρ value is 0.699 with p-value 0.01, so the frequencies are strongly correlated.

This figure also suggests that the points which are uncorrelated correspond to cases with high values of HFE but low values for TMR: we will return to these cases later.

In the second experiment, we test if the TMR and the number of references from manual corroboration are correlated, using the number of references in the literature (from manual corroboration) for common Vortis links. Spearman's correlation ρ value is 0.731 with p-value 0.01, suggesting
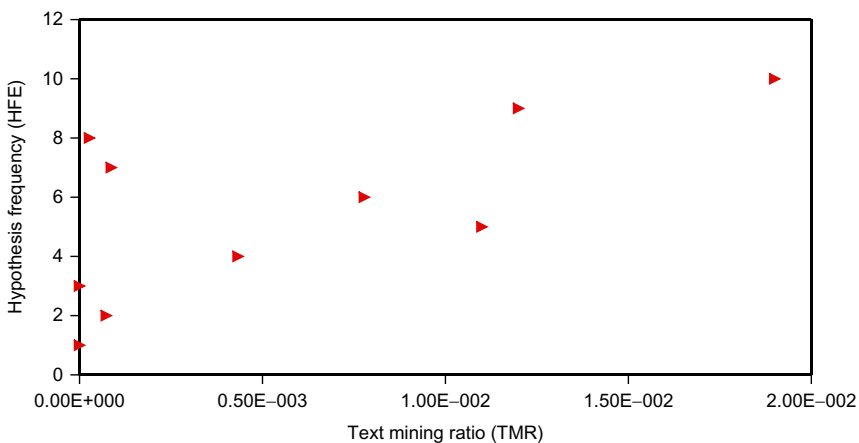
Figure 4.16 Correlation between the hypothesis frequencies for common links and the text-mining ratio (TMR). Spearman's correlation value, $\rho = 0.699$ with $p \leq 0.01$.

that the automatic corroboration using text mining is consistent with the manual corroboration by the domain experts (Fig. 4.17).

3.4.2 Merged trophic network

We use the automatic corroboration approach described in Section 2.3 and provide the pairs of trophic interactions from the merged network as input to generate a literature network. Figure 4.18 shows a literature network that is generated for the merged trophic network from Vortis and Pitfall data (Section 3.1). As for the common literature network, each edge represents co-occurrences of species, however, instead of using numbers, the thickness of each edge now represents a degree of strength based on the number of LH of species that have been found with at least one trophic interaction lexicon (eat, feed, prey or consume). These numbers are based on the co-occurrences in the abstract and full text (whenever available) using Google Scholar as the search engine for the text mining.

   The frequencies of the learned trophic links (HFE), the number of hits from the literature (Lit. Hits) and the TMR for the trophic network learned from Vortis and pitfall are given in Appendix B. As in Tamaddoni–Nezhad et al. (2012a), we removed weak links with frequencies less than 3 (out of 10) from the analysis. We computed Spearman's correlation between the frequencies (HFE) and the total number of hits from the literature (Lit. Hits). Unlike the common literature network, a manual corroboration for merged trophic links is not available and so we can only study the correlation between the frequencies of the learned trophic links (HFE) and the automatic corroboration.
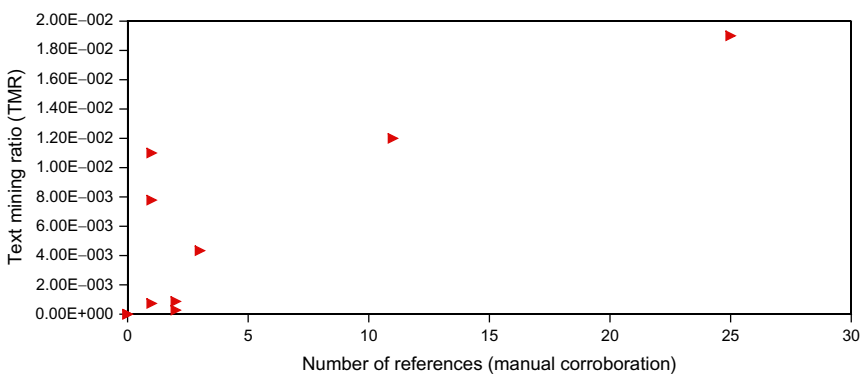
Figure 4.17 Correlation between the TMR and the number of references (from manual corroboration). Spearman's correlation ρ value is 0.731 with *p*-value 0.01.

Figure 4.18 Literature network for merged Vortis and pitfall data.

Figure 4.19 shows the correlation between the hypothesis frequencies (HFE) for merged trophic links and the total number of hits for these links (Lit. Hits). Spearman's correlation ρ value is 0.821 with p-value 0.01: frequencies of merged trophic links (using HFE) were consistent with the total text mining hits for those links, especially for links with high HFE values.

This figure also suggests that the points which are not well correlated correspond to cases with high values of HFE but low values for total number of hits: that is, there are trophic links which have been learned with high frequencies but with little or no references in the literature. These out-of-correlation cases are comparable with potential novel hypotheses marked by dashed red circles in Fig. 4.6 (hypotheses with high values of HFE and no references in the literature), and merit further study to test whether they are realised in nature.

In general, the numbers of references found by automated corroborations are higher and also include those found by manual corroboration.

Figure 4.19 Correlation between the hypothesis frequencies (HFE) for merged trophic links and the total number of hits for these links. Spearman's correlation ρ value is 0.821with p-value 0.01.

However, the automated corroboration is not as accurate as manual corroboration and the results also include the co-occurrences that are not directly relevant to a trophic interaction. Nevertheless, the numbers of references generated by automated corroboration are useful as a relative measure for comparison between different trophic links.

4. DISCUSSION AND CONCLUSIONS

We find that, as with the much smaller Vortis dataset, the machine learning methodology developed by Bohan et al. (2011) produced convincing food webs from a pitfall sampling dataset from UK agricultural fields. Comparison of the trophic links between species present in both the learned Vortis and pitfall webs revealed a significant correlation, consistent with two non-independent *a priori* expectations: firstly, that for a learning methodology to be working correctly it should produce similar links from independent datasets, such as the Vortis and pitfall data; and secondly, that we would expect that similar network structure would be learned amongst the same suite of species sampled using different protocols. The A/ILP learning methodology does indeed appear to be working and we are learning the same

information from distinct protocols, and we take this to mean that we can merge these two datasets and learn larger networks of trophic interactions.

Comparison of our automatic validation methodology with both a manual literature review, for the Vortis-learned food web, and the web learned from the merged Vortis and pitfall data also produced significant rank correlations, suggesting that we can validate, automatically, food webs using large online databases and that the food web learned from the merging of the datasets has a valid footprint in the literature. In essence, the data produced by combining distinct protocols can be learned to produce a valid food web. Finally, we have shown that functional approaches to trophic networks appear to have similar explanatory power to species-based food webs, but with apparently greater parsimony.

This chapter brings together a series of important developments that we believe could, potentially, greatly facilitate and speed up the construction of food webs and advances in network ecology in general, and not just for agro-ecosystems. Prior to developing the methodologies we describe here, construction of empirical food webs, or any ecological interaction network, entails observation of the interactions between species. This has typically been done through direct observation in the field (e.g., Woodward et al., 2012) and laboratory and/or through more indirect methods such as searching the literature for published evidence of interactions (e.g., Layer et al., 2010). In a system with 300 species, such as might exist in a typical arable agricultural system in northern Europe, this could entail researching evidence of around 90,000 interactions (300^2 interactions in a system of 300 species, with cannibalism) at a huge cost in terms of time and money. While some ecological knowledge of interactions between species would greatly reduce the number of interactions to be searched, the amount of work involved is still considerable.

Our methodologies move things on to a rather different position. Now it should be possible, with appropriate baseline data, to: (i) automatically learn networks of interactions, quickly and efficiently; (ii) validate those networks using automated validation methodologies; and (iii) in place of searching for/observing interactions between all the species in the system, only then attempt to directly observe those species interactions that strike us as unexpected, such as the 'spiders as prey' links (Bohan et al., 2011; Davey et al., 2013), saving considerable research time and money. Moreover, our work suggests that data from different protocols can be merged, following satisfaction of the tests we have developed, to produce much larger networks. In principal, we can also show that functional descriptions, as exemplified by the functional

description developed by Caron-Lormier et al. (2009) and Hawes et al. (2009), have considerable power of explanation for agricultural food webs.

As with the work of Bohan et al. (2011), the results we present are individual, hypothetical 'eats' trophic relationships that have been assembled into candidate heterotrophic, arable food webs that are relevant to the national scale. The learned webs are for the soil-surface active (pitfall–trap protocol) and epigeal (Vortis protocol) invertebrates present within the arable system and the learning allows us to consider, in turn, each hypothesised trophic link. The food web learned from merged data is richer than either the individual pitfall or Vortis food webs alone or merged together. For example, machine learning can suggest trophic links between species present in one dataset as predator and species sampled in the other as prey. These trophic relations cannot be captured by simply merging trophic links from Vortis and pitfall food webs.

As with the findings of Bohan et al. (2011), detritivore Collembola were hypothesised to be the predominant prey items within the putative network, as expected from direct observation. The learnt food web suggests that these large generalist or omnivorous carabid beetles were indeed the important predators within the food web; an expectation also supported by their high abundance and large biomass in the two sampling protocols. As previously, members of the *Pterostichus*, *Nebria*, *Bembidion* and *Trechus* genera were hypothesised to prey upon a variety of species and taxa, including one another, suggesting that predation within the guild of predators could play an important role in the structuring and dynamics of agricultural ecosystems.

Trophic links discovered through learning might be tested formally using molecular diagnostics, or the more traditional gut dissection and observational methodologies (Traugott et al., 2013). However, we feel that beyond an acceptable period of testing to show that learning methods produce valuable trophic hypotheses in different situations and for different species combinations, repeated testing of whole networks would miss the value of the approach, and resources could be better directed elsewhere. Automated discovery will have most value when it is used for rapid network learning without the burden of observation that is currently required for food web construction (Ings et al., 2009). After the method has 'proved its mettle,' however, such network learning and generation will still require some level of testing and verification. This should probably be limited to testing links that were not expected, or cannot be accounted for, rather than extensive retesting of well-established trophic interactions.

The example of the spiders would appear as a model for this statement: in Bohan et al. (2011), and again here, spiders were found to be at the base of the food web, as prey. Yet the very great majority of spiders, and all those in the dataset, are obligate predators (Bell et al., 2005). Why were these at the base of the food web? Our initial concern was that this positioning was an artefact of the small size of the spiders, identifying them by default as prey items for the learning. Recently, however, Davey et al. (2013) tested this learnt trophic link explicitly using molecular methods and found that this apparently illogical hypothetical link was indeed supported by the molecular data: carabids, such as *P. melanarius*, prey extensively upon spiders; a finding corroborated by Eitzinger and Traugott (2011).

Bohan et al. (2013) raised the question of generality in ecological data and 'growing' larger datasets for learning wider networks. For them, one measure of generality was whether the links learnt for a suite of species in two independent datasets, say the data from the Vortis and pitfall protocols, were similar; put simply, the Vortis '*eats*' predicates should apply with high predictive power to the pitfall, and vice versa. By this criterion, the learning methods are general because the link structures learnt from the two protocols were significantly correlated. It further suggests that the two protocols contain the same information; at least as far as shared species. Thus, we argue that the sets of treatment ratios calculated from the Vortis and pitfall sample data can be merged and we could 'grow' the food web by learning hypothetical trophic links across a dataset larger than that derived from any one sampling protocol. This is potentially important as many ecosystems have physical attributes, such as the water column and benthos in aquatic systems, that necessitate the use of different methods to sample different habitats and parts of the food web.

The text–mining approach, for validating learnt food webs, presented in this chapter was inspired by PubGene and other similar text–mining tools that are based on the co-occurrence of given terms. For our case, however, co-occurrences of pairs of ecological species are considered that are related by trophic interaction lexicons, that is, *eat, feed, prey, consume*. Inspection of the results suggests that the numbers of references generated by automated corroboration are useful as a relative measure for comparison between different trophic links. In general, the numbers of references found by automated corroborations are higher and also include those found by manual corroboration. However, the automated corroboration is not as accurate as manual corroboration and the results also include the co-occurrences that are not directly relevant to a trophic interaction. The results presented in this report indicate the feasibility of automatic corroboration of food webs that are too large for manual corroboration. The results also suggest that the

frequencies of merged trophic links (using HFE) are consistent with the total number of hits for these links. As future work, we would like to improve the accuracy of automated corroboration by extending the approach to include more advanced techniques, for example, natural language processing. However, the results presented in this chapter indicate the feasibility of automatic corroboration of food webs. Moreover, the current approach was able to successfully identify hypothetical trophic relations for which there are little or no information in the literature (potential novel hypotheses).

This is to our knowledge the first time that text mining has been used to automate corroboration of food web particularly by full text searches through ecological papers. The text-mining work presented here can be viewed as a first step towards automatic corroboration of machine learned food webs and we believe that the initial results are encouraging.

Using functional approaches for understanding and predicting system change is currently a topic of much debate amongst ecologists in both terrestrial and aquatic systems (e.g., Friberg et al., 2011). Functional explanations potentially have value as a synthetic taxonomy for the organisms present in agriculture. Whereas species presence changes between crops and before/after management, and therefore species-based webs of these systems might be expected to have low generality, functionally based webs could allow relatively straightforward analysis and prediction of ecological differences between crops or managements, while preserving some of the details of diversity that humans prize. Machine learning approaches might be used to provide a test of ecological functionality. Using the ecosystem-wide description of the arable food web, it might be possible to ask whether species- or functionally based descriptions yield food webs that have greater parsimony and might, therefore, be more robust predictors of the effects of environmental change on agro–ecosystem diversity and productivity.

We demonstrated that a functional network, constructed by learning trophic links between functional groups is at least as accurate as the trophic network for individual species, despite being less complex (i.e., having fewer nodes and edges). This could offer insights into the role of trophic uniqueness and functional redundancy in natural food webs. We believe that this chapter represents a breakthrough for ecosystem and food web research, which should lead to important advances in ecological theory and improved management of ecosystems under environmental change. While we envisage immediate opportunities to apply this knowledge to optimise the delivery of food and other ecosystem services from agricultural land, the next step is to demonstrate whether functional explanations learned in one cropping situation also have predictive value for novel situations.

ACKNOWLEDGEMENTS

APPENDIX A. REFERENCES USED FOR MANUAL CORROBORATION (FIG. 4.6)

1. Alexander, K.N.A., 2002. The invertebrates of living and decaying timber in Britain and Ireland—a provisional annotated checklist. Engl. Nat. Res. Rep. 467, 1–142.
2. Bauer, T., 1982. Prey-capture in a ground-beetle larva. Animal Behav. 30 (1), 203–208.
3. Bell, J.R., Andrew King, R., Bohan, D.A., Symondson, W.O.C., 2010. Spatial cooccurrence networks predict the feeding histories of polyphagous arthropod predators at field scales. Ecography 33 (1), 64–72.
4. Berg, K., 2007. The role of detrital subsidies for biological control by generalist predators evaluated by molecular gut content analysis. PhD Thesis, Universiẗats–und Landesbibliothek Darmstadt.
5. Desender, K., Pollet, M., 1985. Ecological data on *Clivina fossor* (coleoptera, carabidae) from a pasture ecosystem. ii. reproduction, biometry, biomass, wing polymorphism and feeding ecology. Rev. Ecol. Biol. Sol. 22 (2), 233–246.
6. Dinter, A., 1998. Intraguild predation between erigonid spiders, lacewing larvae and carabids. J. Appl. Entomol. 122 (1–5), 163–167.
7. Lattin, J.D., 1989. Bionomics of the nabidae. Annu. Rev. Entomol. 34 (1), 383–400.
8. Lindsey, J., 2006. Ecology of Commanster, http://www.commanster.eu/commanster/insects/bugs/spbugs/saldula.saltatoria.html.
9. Pons, X., Lumbierres, B., Albajes, R., 2009. Heteropterans as aphid predators in intermountain alfalfa. Eur. J. Entomol. 106 (3), 369–378.
10. Schaefer, C.W., Panizzi, A.R., 2000. Heteroptera of Economic Importance. CRC.

11. Sunderland, K.D., 1975. The diet of some predatory arthropods in cereal crops. J. Appl. Ecol. 12 (2), 507–515.
12. Sunderland, K.D., Lovei, G.L., Fenlon, J., 1995. Diets and reproductive phenologies of the introduced ground beetles harpalus–affinis and clivina–australasiae (coleoptera, carabidae) in New-Zealand. Aust. J. Zool. 43 (1), 39–50.
13. Toft, S., 2005. The quality of aphids as food for generalist predators: implications for natural control of aphids. Eur. J. Entomol. 102 (3), 371.
14. Turner, B.D., 1984. Predation pressure on the arboreal epiphytic herbivores of larch trees in southern England. Ecol. Entomol. 9 (1), 91–100.
15. Warner, D.J., Allen-Williams, L.J., Warrington, S., Ferguson, A.W., Williams, I.H., 2003. Mapping, characterisation, and comparison of the spatio-temporal distributions of cabbage stem ea beetle (psylliodes chrysocephala), carabids, and collembolan in a crop of winter oilseed rape (brassica napus). Entomol. Exp. Appl. 109 (3), 225–234.
16. Weber, D.C., Lundgren, J.G., 2009. Assessing the trophic ecology of the coccinellidae: their roles as predators and as prey. Biol. Control 51 (2), 199–214.

APPENDIX B

Hypotheses frequency estimation (HFE), literature hits and text-mining ratio (TMR) for pair of predator/prey for food web and literature network generated from merged Vortis and pitfall data.

Predator	Prey	HFE	Lit. hits	TMR
Bembidion lampros	*Lepthyphantes tenuis*	10	48	0.026
Trechus quadristriatus	*Bembidion tetracolum*	10	65	0.054
Pterostichus cupreus	*Agonum dorsale*	9	97	0.081
Loricera pilicornis	*Bembidion aeneum*	9	41	0.037
Bembidion guttula	Isotomidae	10	3	8.049e−04
Leistus spinibarbis	Isotomidae	8	1	2.773e−04
Agonum dorsale	*Bembidion lampros*	9	202	0.105
Asaphidion stierlini	Isotomidae	7	1	2.826e−04
Carabid larvae	*Trechus quadristriatus*	8	45	0.031

Continued

Predator	Prey	HFE	Lit. hits	TMR
Nebria brevicollis	*Lepthyphantes tenuis*	5	32	0.016
Pterostichus melanarius	Sminthuridae	10	19	5.000e−03
Calathus fuscipes	Isotomidae	10	3	7.145e−04
Pterostichus cupreus	*Notiophilus biguttatus*	7	52	0.033
Calathus fuscipes	*Bembidion guttula*	9	37	0.038
Pterostichus niger	Isotomidae	10	3	6.749e−04
Carabid larvae	*Bembidion lampros*	5	58	0.032
Agonum obscurum	*Clivina fossor*	8	13	0.014
Bembidion lampros	Poduridae	10	4	1.825e−03
Pterostichus niger	*Trechus quadristriatus*	10	84	0.045
Trechus quadristriatus	*Bembidion quadrimaculatum*	10	78	0.054
Pterostichus niger	*Stomis pumicatus*	9	58	0.042
Stomis pumicatus	*Bembidion tetracolum*	10	30	0.043
Coccinelid larvae	*Demetrias atricapillus*	4	0	0.000e+00
Bembidion biguttatum	*Lepthyphantes tenuis*	9	2	2.915e−03
Nebria brevicollis	*Agonum obscurum*	5	13	8.398e−03
Pterostichus strenuus	Entomobryidae	10	2	6.341e−04
Bembidion quadrimaculatum	Sminthuridae	10	1	4.558e−04
Bembidion lunulatum	*Lepthyphantes tenuis*	10	4	5.908e−03
Carabid larvae	*Notiophilus biguttatus*	7	54	0.036
Asaphidion stierlini	*Lepthyphantes tenuis*	5	2	3.350e−03
Coccinelid larvae	Poduridae	4	2	2.186e−03
Agonum dorsale	Entomobryidae	10	15	4.583e−03
Pterostichus strenuus	*Bembidion aeneum*	9	27	0.044
Synuchus nivalis	*Notiophilus biguttatus*	5	20	0.018
Bembidion obtusum	Sminthuridae	10	4	2.096e−03
Nebria brevicollis	Poduridae	8	5	2.126e−03

Predator	Prey	HFE	Lit. hits	TMR
Stomis pumicatus	Entomobryidae	10	2	6.513e−04
Nebria brevicollis	*Notiophilus biguttatus*	6	196	0.080
Pterostichus melanarius	*Amara familiaris*	9	99	0.038
Calathus fuscipes	*Lepthyphantes tenuis*	7	13	0.010
Coccinella septempunctata	Isotomidae	9	31	2.860e−03
Pterostichus niger	Curculionidae	10	55	8.140e−04
Stomis pumicatus	*Bembidion quadrimaculatum*	10	36	0.038
Nebria brevicollis	*Bembidion guttula*	4	55	0.032
Loricera pilicornis	Entomobryidae	10	13	3.569e−03
Synuchus nivalis	Sminthuridae	9	2	1.110e−03
Synuchus nivalis	*Trechus quadristriatus*	10	19	0.018
Loricera pilicornis	*Notiophilus biguttatus*	7	150	0.077
Agonum obscurum	Sminthuridae	8	0	0.000e+00
Synuchus nivalis	*Bembidion tetracolum*	4	12	0.028
Notiophilus biguttatus	Sminthuridae	10	7	2.657e−03
Pterostichus cupreus	Curculionidae	9	49	7.294e−04
Bembidion quadrimaculatum	Poduridae	10	1	7.062e−04
Trechus quadristriatus	Isotomidae	9	23	5.230e−03
Nebria brevicollis	*Agonum dorsale*	8	156	0.075
Clivina fossor	Entomobryidae	10	7	1.995e−03
Notiophilus biguttatus	*Lepthyphantes tenuis*	10	20	0.013
Stomis pumicatus	*Bembidion obtusum*	9	30	0.046
Bembidion aeneum	Cimicidae nymphs	9	1	7.874e−03
Demetrias atricapillus	Isotomidae	10	8	2.099e−03
Trechus discus	Poduridae	9	0	0.000e+00
Stomis pumicatus	*Lepthyphantes tenuis*	10	5	5.269e−03
Trechus obtusus	*Bembidion tetracolum*	6	26	0.036

Predator	Prey	HFE	Lit. hits	TMR
Trechus quadristriatus	*Demetrias atricapillus*	10	65	0.052
Bembidion guttula	Sminthuridae	10	1	5.244e−04
Coccinelid larvae	Sminthuridae	9	1	5.907e−04
Trechus quadristriatus	*Lepthyphantes tenuis*	10	37	0.025
Notiophilus substriatus	Isotomidae	9	0	0.000e+00
Pterostichus cupreus	*Trechus quadristriatus*	9	66	0.044
Pterostichus cupreus	*Bembidion tetracolum*	5	28	0.032
Bembidion aeneum	Poduridae	10	1	9.930e−04
Trechus quadristriatus	*Bembidion guttula*	10	44	0.038
Pterostichus cupreus	Isotomidae	8	18	4.430e−03
Agonum dorsale	*Bembidion lunulatum*	10	12	0.016
Notiophilus substriatus	*Bembidion aeneum*	8	12	0.047
Notiophilus biguttatus	Entomobryidae	10	18	4.952e−03
Pterostichus melanarius	Miridae nymphs	6	1	4.507e−04
Dyschirius globosus	Entomobryidae	10	3	9.836e−04
Agonum muelleri	Poduridae	4	0	0.000e+00
Bembidion lampros	Isotomidae	10	28	5.846e−03
Bembidion lunulatum	Entomobryidae	8	1	3.573e−04
Bembidion aeneum	*Lepthyphantes tenuis*	9	2	3.017e−03
Trechus obtusus	*Bembidion obtusum*	8	26	0.039
Bembidion guttula	Entomobryidae	10	2	6.880e−04
Carabid larvae	Isotomidae	10	10	2.499e−03
Leistus spinibarbis	Entomobryidae	8	2	7.179e−04
Bembidion aeneum	Sminthuridae	10	1	5.602e−04
Calathus fuscipes	Curculionidae	9	60	8.913e−04
Bembidion biguttatum	Poduridae	9	1	9.709e−04
Coccinelid larvae	Isotomidae	9	2	5.693e−04
Pterostichus niger	Amara aenea	7	66	0.036

Predator	Prey	HFE	Lit. hits	TMR
Asaphidion stierlini	*Bembidion guttula*	4	4	0.013
Nebria brevicollis	Sminthuridae	9	8	2.556e−03
Calathus fuscipes	Entomobryidae	10	4	1.184e−03
Leistus spinibarbis	*Demetrias atricapillus*	7	16	0.035
Synuchus nivalis	Isotomidae	10	2	5.522e−04
Nabidae nymphs	Sminthuridae	10	1	5.959e−04
Leistus rufomarginatus	Entomobryidae	10	1	3.505e−04
Carabid larvae	Miridae nymphs	9	2	3.333e−03
Nebria brevicollis	Isotomidae	8	24	4.848e−03
Pterostichus melanarius	*Bembidion tetracolum*	10	90	0.037
Clivina fossor	Sminthuridae	10	7	2.791e−03
Nebria salina	Entomobryidae	10	0	0.000e+00
Bembidion aeneum	Acupalpus dorsalis	9	3	0.017
Pterostichus niger	Miridae nymphs	9	0	0.000e+00
Saldula saltatoria	Entomobryidae	10	2	7.080e−04
Bembidion quadrimaculatum	Isotomidae	10	5	1.246e−03
Agonum dorsale	*Bembidion tetracolum*	9	48	0.053
Nebria salina	*Trechus quadristriatus*	7	28	0.025
Leistus spinibarbis	*Clivina fossor*	9	20	0.021
Trechus obtusus	*Lepthyphantes tenuis*	9	6	6.231e−03
Demetrias atricapillus	Sminthuridae	10	3	1.506e−03
Nebria salina	*Bembidion tetracolum*	10	20	0.040
Coccinelid larvae	*Bembidion lampros*	10	2	1.489e−03
Carabid larvae	*Lepthyphantes tenuis*	9	9	8.499e−03
Bradycellus verbasci	Cimicidae nymphs	10	1	6.711e−03
Nebria brevicollis	*Trechus quadristriatus*	10	181	0.076
Notiophilus biguttatus	*Piesma maculatum*	10	1	9.643e−04
Carabid larvae	*Bembidion tetracolum*	4	9	0.011

Predator	Prey	HFE	Lit. hits	TMR
Coccinelid larvae	*Lepthyphantes tenuis*	9	2	3.503e−03
Nabidae nymphs	*Bembidion lampros*	9	1	7.530e−04
Carabid larvae	Curculionidae	9	63	9.386e−04
Notiophilus substriatus	*Bembidion quadrimaculatum*	9	14	0.021
Propylea quattuordecimpunctata	*Lepthyphantes tenuis*	10	4	6.601e−03
Pterostichus niger	Sminthuridae	10	1	3.810e−04
Trechus quadristriatus	Entomobryidae	10	14	3.913e−03
Bembidion aeneum	Isotomidae	10	1	2.774e−04
Pterostichus strenuus	*Bembidion guttula*	6	46	0.062
Bembidion obtusum	Poduridae	10	1	8.850e−04
Trechus secalis	Isotomidae	9	4	1.079e−03
Pterostichus strenuus	Poduridae	4	2	1.453e−03
Demetrias atricapillus	Cimicidae nymphs	9	0	0.000e+00
Agonum muelleri	*Bembidion guttula*	9	37	0.059
Pterostichus cupreus	Sminthuridae	8	8	3.567e−03
Demetrias atricapillus	Entomobryidae	10	3	1.003e−03
Loricera pilicornis	*Bembidion tetracolum*	6	68	0.053
Trechus obtusus	Entomobryidae	8	3	9.724e−04
Agonum dorsale	Poduridae	10	3	2.007e−03
Trechus obtusus	*Notiophilus biguttatus*	4	59	0.042
Pterostichus cupreus	Amara aenea	8	43	0.030
Agonum dorsale	*Notiophilus biguttatus*	10	92	0.058
Asaphidion stierlini	*Bembidion aeneum*	6	2	0.011
Notiophilus biguttatus	Cimicidae nymphs	9	1	1.024e−03
Agonum muelleri	Isotomidae	10	4	1.038e−03
Nabidae nymphs	Cimicidae nymphs	10	1	0.050
Loricera pilicornis	Poduridae	10	4	2.146e−03

Predator	Prey	HFE	Lit. hits	TMR
Bembidion quadrimaculatum	*Metabletus obscuroguttatus*	9	2	3.591e−03
Bembidion guttula	Cimicidae nymphs	9	1	4.016e−03
Bembidion biguttatum	Isotomidae	9	2	5.513e−04
Nebria salina	*Amara familiaris*	9	28	0.040
Agonum muelleri	*Lepthyphantes tenuis*	7	8	8.762e−03
Nebria salina	*Loricera pilicornis*	9	37	0.031
Clivina fossor	Poduridae	10	1	5.780e−04
Pterostichus cupreus	Entomobryidae	5	8	2.467e−03
Nebria brevicollis	*Amara familiaris*	9	80	0.041
Agonum obscurum	*Bembidion guttula*	9	10	0.031
Bembidion lampros	*Metabletus foveatus*	9	16	0.012
Dromius linearis	*Metabletus obscuroguttatus*	7	3	0.015
Bembidion lampros	Entomobryidae	10	18	4.534e−03
Loricera pilicornis	*Lepthyphantes tenuis*	10	19	0.013
Agonum dorsale	*Demetrias atricapillus*	9	70	0.074
Clivina fossor	*Lepthyphantes tenuis*	10	12	8.658e−03
Carabid larvae	Entomobryidae	5	11	3.458e−03
Pterostichus melanarius	*Bembidion guttula*	9	59	0.025
Calathus fuscipes	Miridae nymphs	7	0	0.000e+00
Coccinelid larvae	Entomobryidae	7	2	7.427e−04
Pterostichus melanarius	Isotomidae	10	42	7.473e−03
Trechus quadristriatus	Cimicidae nymphs	10	2	2.174e−03
Pterostichus melanarius	*Agonum dorsale*	10	196	0.071
Agonum muelleri	Sminthuridae	9	1	4.914e−04
Bembidion quadrimaculatum	*Lepthyphantes tenuis*	10	11	0.010
Nebria brevicollis	Entomobryidae	10	13	3.148e−03

Continued

Predator	Prey	HFE	Lit. hits	TMR
Agonum muelleri	*Bembidion aeneum*	8	25	0.050
Calathus fuscipes	Sminthuridae	9	3	1.261e−03
Dromius linearis	Entomobryidae	10	4	1.410e−03
Bembidion lampros	*Piesma maculatum*	10	1	7.289e−04
Clivina fossor	Cimicidae nymphs	8	1	1.176e−03
Demetrias atricapillus	Acupalpus dorsalis	9	4	0.010
Bembidion lunulatum	Cimicidae nymphs	9	1	7.092e−03
Patrobus atrorufus	Isotomidae	10	2	5.218e−04
Bembidion lampros	Sminthuridae	10	11	3.704e−03
Agonum dorsale	Sminthuridae	9	9	3.960e−03
Bembidion lunulatum	Sminthuridae	10	1	5.559e−04
Notiophilus biguttatus	Poduridae	10	4	2.154e−03
Agonum muelleri	*Amara bifrons*	4	38	0.060
Trechus quadristriatus	Sminthuridae	10	11	4.267e−03
Agonum dorsale	*Lepthyphantes tenuis*	10	37	0.032
Bembidion lunulatum	Poduridae	10	1	9.794e−04
Coccinellidae	*Bembidion lampros*	10	172	5.991e−03
Bembidion quadrimaculatum	Entomobryidae	9	4	1.252e−03
Bembidion obtusum	Isotomidae	10	8	2.146e−03
Pterostichus strenuus	Isotomidae	7	5	1.258e−03
Calathus fuscipes	*Trechus quadristriatus*	9	109	0.067
Coccinelid larvae	Miridae nymphs	9	2	0.018
Loricera pilicornis	Sminthuridae	9	13	4.921e−03
Loricera pilicornis	*Demetrias atricapillus*	9	57	0.043
Calathus fuscipes	*Bembidion tetracolum*	10	47	0.046
Bembidion guttula	Poduridae	10	1	8.857e−04
Carabid larvae	*Agonum dorsale*	10	40	0.035

Predator	Prey	HFE	Lit. hits	TMR
Trechus discus	Entomobryidae	10	0	0.000e+00
Pterostichus niger	*Bembidion guttula*	9	47	0.039
Asaphidion stierlini	*Notiophilus biguttatus*	10	6	5.803e−03
Agonum dorsale	Isotomidae	9	31	7.574e−03
Pterostichus strenuus	*Lepthyphantes tenuis*	9	12	0.012
Stomis pumicatus	*Clivina fossor*	5	62	0.049
Stomis pumicatus	Sminthuridae	10	1	4.829e−04
Asaphidion stierlini	*Clivina fossor*	5	3	3.308e−03
Pterostichus niger	*Agonum dorsale*	10	56	0.035
Loricera pilicornis	*Bembidion guttula*	9	62	0.050
Nebria salina	Sminthuridae	9	0	0.000e+00
Agonum obscurum	Entomobryidae	10	0	0.000e+00
Carabid larvae	Sminthuridae	10	7	3.210e−03
Nebria salina	*Notiophilus biguttatus*	10	40	0.034
Loricera pilicornis	Isotomidae	10	18	4.034e−03
Bembidion quadrimaculatum	Onychiuridae	10	2	8.834e−04
Bembidion aeneum	Entomobryidae	10	1	3.591e−04
Trechus secalis	Entomobryidae	9	1	3.463e−04
Pterostichus cupreus	*Bembidion lampros*	5	150	0.079
Pterostichus melanarius	*Lepthyphantes tenuis*	10	44	0.016
Clivina fossor	Isotomidae	10	11	2.542e−03
Nabidae nymphs	*Bembidion tetracolum*	8	0	0.000e+00
Nebria brevicollis	*Bembidion tetracolum*	10	60	0.034
Agonum muelleri	Entomobryidae	7	3	9.885e−04
Patrobus atrorufus	*Bembidion guttula*	9	25	0.042
Agonum dorsale	*Bembidion aeneum*	9	23	0.031
Synuchus nivalis	*Agonum dorsale*	6	26	0.034
Nebria salina	*Bembidion aeneum*	7	15	0.045

Predator	Prey	HFE	Lit. hits	TMR
Patrobus atrorufus	*Agonum dorsale*	9	26	0.027
Synuchus nivalis	*Lepthyphantes tenuis*	5	5	7.353e−03
Bembidion aeneum	Onychiuridae	7	0	0.000e+00
Trechus quadristriatus	Poduridae	6	3	1.667e−03
Calathus fuscipes	*Amara familiaris*	9	67	0.055
Trechus quadristriatus	*Notiophilus biguttatus*	10	139	0.073
Pterostichus strenuus	*Clivina fossor*	10	82	0.061
Stomis pumicatus	*Bembidion lunulatum*	9	13	0.024
Nabidae nymphs	*Bembidion quadrimaculatum*	6	1	1.812e−03
Demetrias atricapillus	Poduridae	10	0	0.000e+00
Bembidion obtusum	*Lepthyphantes tenuis*	10	11	0.014
Pterostichus melanarius	*Trechus quadristriatus*	10	218	0.071
Agonum muelleri	*Agonum dorsale*	9	40	0.040
Notiophilus biguttatus	Isotomidae	10	22	4.938e−03
Pterostichus strenuus	Sminthuridae	5	4	1.857e−03

REFERENCES

Afroozi Milani, G., Bohan, D., Dunbar, S., Muggleton, S.H., Raybould, A., Tamaddoni-Nezhad, A., 2012. Machine learning and text mining of trophic links. In: Proceedings of the 11th International Conference on Machine Learning and Applications (ICMLA 2012). IEEE Computer Society, Washington DC, pp. 410–415.

Aukema, B., 1990. Taxonomy, life history and distribution of three closely related species of the genus *Calathus* (Coleoptera: Carabidae). Tijdschr. Entomol. 133, 121–141.

Baulcombe, D., Crute, I., Davies, B., Dunwell, J., Gale, M., Jones, J., Pretty, J., Sutherland, W., Toulmin, C., Green, N., 2009. Reaping the benefits: science and the sustainable intensification of global agriculture. The Royal Society, London, UK.

Bell, J.R., Bohan, D.A., Shaw, E.M., Weyman, G.S., 2005. Ballooning dispersal using silk: world fauna, phylogenies, genetics and models. Bull. Entomol. Res. 95, 69–114.

Bennett, E.M., Peterson, G.D., Gordon, L.J., 2009. Understanding relationships among multiple ecosystem services. Ecol. Lett. 12, 1394–1404.

Bernot, G., Comet, J.P., Richard, A., Guespin, J., 2004. Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic. J. Theor. Biol. 229, 339–347.

Blaschke, C., Andrade, M.A., Ouzounis, C., Valencia, A., 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. Proc. Int. Conf. Intell. Syst. Mol. Biol. (ISMB 99) 60–67.

Bohan, D.A., Bohan, A.C., Glen, D.M., Symondson, W.O.C., Wiltshire, C.W., Hughes, L., 2000. Spatial dynamics of predation by carabid beetles on slugs. J. Anim. Ecol. 69, 367–379.

Bohan, D.A., Boffey, C.W.H., Brooks, D.R., Clark, S.J., Dewar, A.M., Firbank, L.G., Haughton, A.J., Hawes, C., Heard, M.S., May, M.J., 2005. Effects on weed and invertebrate abundance and diversity of herbicide management in genetically modified herbicide-tolerant winter-sown oilseed rape. Proc. R. Soc. B Biol. Sci. 272, 463–474.

Bohan, D.A., Caron-Lormier, G., Muggleton, S., Raybould, A., Tamaddoni-Nezhad, A., 2011. Automated discovery of food webs from ecological data using logic-based machine learning. PLoS One 6, e29028.

Bohan, D., Raybould, A., Mulder, C., Woodward, G., Tamaddoni-Nezhad, A., Bluthgen, N., Pocock, M., Muggleton, S., Evans, D., Astegiano, J., Massol, F., Loeuille, N., Petit, S., Macfadyen, S., 2013. Chapter one - Networking agroecology: integrating the diversity of agroecosystem interactions. *Adv. in Ecol. Res.*, vol 49, 1–67.

Brooks, D.R., Bohan, D.A., Champion, G.T., Haughton, A.J., Hawes, C., Heard, M.S., Clark, S.J., Dewar, A.M., Firbank, L.G., Perry, J.N., Rothery, P., Scott, R.J., Woiwod, I.P., Birchall, C., Skellern, M.P., Walker, J.H., Baker, P., Bell, D., Browne, E.L., Dewar, A.J.G., Fairfax, C.M., Garner, B.H., Haylock, L.A., Horne, S.L., Hulmes, S.E., Mason, N.S., Norton, L.R., Nuttall, P., Randle, Z., Rossall, M.J., Sands, R.J.N., Singer, E.J., Walker, M.J., 2003. Invertebrate responses to the management of genetically modified herbicide-tolerant and conventional spring crops I. Soil-surface- active invertebrates. Philos. Trans. R. Soc. Lond. B 358, 1847–1862.

Busch, M., La Notte, A., Laporte, V., Erhard, M., 2012. Potentials of quantitative and qualitative approaches to assessing ecosystem services. Ecol. Indic. 21, 89–103.

Calzone, L., Fages, F., Soliman, S., 2006. Biocham: an environment for modeling biological systems and formalizing experimental knowledge. Bioinformatics 22 (1805–1807), 2006.

Cardinale, B.J., Harvey, C.T., Gross, K., Ives, A.R., 2003. Biodiversity and biocontrol: emergent impacts of a multi-enemy assemblage on pest suppression and crop yield in an agroecosystem. Ecol. Lett. 6, 857–865.

Caron-Lormier, G., Bohan, D.A., Hawes, C., Raybould, A., Haughton, A.J., et al., 2009. How might we model an ecosystem? Ecol. Model. 220, 1935–1949.

Carpenter, S.R., Kitchell, J.F., 1988. Consumer control of lake productivity. Bioscience 38, 764–769.

Carpenter, S.R., Kitchell, J.F., 1993. The Trophic Cascade in Lakes. Cambridge University Press, Cambridge, UK.

Champion, G.T., May, M.J., Bennett, S., Brooks, D.R., Clark, S.J., Daniels, R.E., Firbank, L.G., Haughton, A.J., Hawes, C., Heard, M.S., Perry, J.N., Randle, Z., Rossall, M.J., Rothery, P., Skellern, M.P., Scott, R.J., Squire, G.R., Thomas, M.R., 2003. Crop management and agronomic context of the Farm Scale Evaluations of genetically modified herbicide-tolerant crops. Philos. Trans. R. Soc. Lond. B 358, 1801–1818.

Chapin, F.S., Sala, O.E., Burke, I.C., Grime, J.P., Hooper, D.U., Lauenroth, W.K., Lombard, A., Mooney, H.A., Mosier, A.R., Naeem, S., 1998. Ecosystem consequences of changing biodiversity. Bioscience 48, 45–52.

Chen, Y., Xu, D., 2004. Global protein function annotation through mining genome-scale data in yeast saccharomyces cerevisiae. Nucleic Acids Res. 32, 6414–6424.

Cohen, J.E., 1991. Food webs as a focus for unifying ecological theory. Ecol. Intl. 19, 1–13.

Cohen, K.B., Hunter, L., 2004. Natural language processing and systems biology. In: Dubitzky, W., Azuaje, F. (Eds.), Artificial Intelligence Methods and Tools for Systems Biology. Springer, The Netherlands, pp. 147–173.

Cohen, K.B., Hunter, L., 2008. Getting started in text mining. PLoS Comput. Biol. 4, e20.

Cohen, J.E., Jonsson, T., Carpenter, S.R., 2003. Ecological community description using the food web, species abundance, and body size. Proc. Natl. Acad. Sci. U.S.A. 100 (4), 1781–1786.

Cohen, J.E., Schittler, D.N., Raffaelli, D.G., Reuman, D.C., 2009. Food webs are more than the sum of their tritrophic parts. Proc. Natl. Acad. Sci. U.S.A. 106, 22335–22340.

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso., M., Hannon, B., Limburg, K., Naeem, S., O'neill, R.V., Paruelo, J., 1997. The value of the world's ecosystem services and natural capital. Nature 387, 253–260.

Dale, J.M., Popescu, L., Karp, P.D., 2010. Machine learning methods for metabolic pathway prediction. BMC Bioinforma. 11, 15.

Davey, J.S., Vaughan, I.P., Andrew King, R., Bell, J.R., Bohan, D.A., Bruford, M.W., Symondson, W.O., 2013. Intraguild predation in winter wheat: prey choice by a common epigeal carabid consuming spiders. J. Appl. Ecol. 50 (1), 271–279.

De Groot, R.S., 2010. Integrating the ecological and economic dimensions in biodiversity and ecosystem service valuation. In: Kumar, P. (Ed.), The Economics of Ecosystems and Biodiversity: Ecological and Economic Foundations. UNEP/Earthprint.

De Raedt, L., Kimmig, A., Toivonen, H., LopezdeMantaras, R., 2007. Problog: a probabilistic prolog and its applications in link discovery. In: Veloso, M.M. (Ed.), Proceedings of the 20th International Joint Conference on Artificial Intelligence. (IJCAI 2007), AAAI Press, pp. 804–809.

DEFRA, 2002. The History of the Farm Scale Evaluations. http://www.defra.gov.uk/environment/fse/background/history/index.htm. Department for Environment, Food and Rural Affairs, London, UK.

Dewar, A.M., May, M.J., Woiwod, I.P., Haylock, L.A., Champion, G.T., Garner, B.H., Sands, R.J., Qi, A., Pidgeon, J.D., 2003. A novel approach to the use of geneti- cally modified herbicide tolerant crops for environmental benefit. Proc. R. Soc. Lond. B270, 335–340. http://dx.doi.org/10.1098/rspb.2003.2248.

Dickinson, G., Murphy, K., 1998. Ecosystems: A Functional Approach. Routledge, London. 190 pp.

Duffy, J.E., 2008. Why biodiversity is important to the functioning of real-world ecosystems. Front. Ecol. Environ. 7 (8), 437–444.

Dunne, J.A., Williams, R.J., Martinez, N.D., 2002. Food-web structure and network theory: the role of connectance and size. Proc. Natl. Acad. Sci. U.S.A. 99 (20), 12917–12922.

Eitzinger, B., Traugott, M., 2011. Which prey sustains cold-adapted invertebrate generalist predators in arable land? Examining prey choices by molecular gut-content analysis. J. Appl. Ecol. 48, 591–599.

Elmegaard, N., Bruus Pedersen, M., 2001. Flora and Fauna in Roundup Tolerant Fodder Beet FieldsNERI Technical Report No. 349. Ministry of Environment and Energy/ National Environmental Research Institute, Roskilde, Denmark.

Elton, C., 1927. Animal Ecology. Sidgwick and Jackson, London.

Emmerson, M.C., Raffaelli, D., 2004. Predator–prey body size, interaction strength and the stability of a real food web. J. Anim. Ecol. 73, 399–409.

Fargione, J., Tilman, D., Dybzinski, R., Lambers, J.H.R., Clark, C., Harpole, W.S., Knops, J.M.H., Reich, P.B., Loreau, M., 2007. From selection to complementarity: shifts in the causes of biodiversity–productivity relationships in a long-term biodiversity experiment. Proc. R. Soc. B 274, 871–876.

Firbank, L.G., Forcella, F., 2000. Genetically modified crops and farmland biodiversity. Science 289, 1481–1482.

Firbank, L.G., Dewar, A.M., Hill, M.O., May, M.J., Perry, J.N., Rothery, P., Squire, G.R., Woiwod, I.P., 1999. Farm-scale evaluation of GM crops explained. Nature 339, 727–728.

Firbank, L.G., Heard, M.S., Woiwod, I.P., Hawes, C., Haughton, A.J., Champion, G.T., Scott, R.J., Hill, M.O., Dewar, A.M., Squire, G.R., May, M.J., Brooks, D.R., Bohan, D.A., Daniels, R.E., Osborne, J.L., Roy, D.B., Black, H.I.J., Rothery, P., Perry, J.N., 2003. An introduction to the Farm Scale Evaluations of genetically modified herbicide-tolerant crops. J. Appl. Ecol. 40, 2–16.

Fisher, B., Turner, R.K., Morling, P., 2009. Defining and classifying ecosystem services for decision making. Ecol. Econ. 68, 643–653.

Fjellberg, A., 1980. Identification Keys to Norwegian Collembola. Norsk Entomologisk Forening, Norway.

Flach, P., Kakas, A.C., 2000. Abductive and inductive reasoning: background and issues. In: Flach, P.A., Kakas, A.C. (Eds.), Abductive and Inductive Reasoning, Pure and Applied Logic. Kluwer, Dordrecht.

Forsythe, T.G., 2000. Ground beetles, Naturalists' handbooks no. 8, second ed. Richmond Publishing, Slough, UK.

Friberg, N., Bonada, N., Bradley, D.C., Dunbar, M.J., Edwards, F.K., Grey, J., Hayes, R.B., Hildrew, A.G., Lamouroux, N., Trimmer, M., Woodward, G., 2011. Biomonitoring of human impacts in freshwater ecosystems: the good, the bad, and the ugly. Adv. Ecol. Res. 44, 2–68.

Gibson, R.H., Nelson, I.L., Hopkins, G.W., Hamlett, B.J., Memmott, J., 2006. Pollinator webs, plant communities and the conservation of rare plants: arable weeds as a case study. J. Appl. Ecol. 43 (2), 246–257.

Gressel, J., Rotteveel, T., 2000. Genetic and ecological risks from biotechnologically-derived herbicide resistant crops: decision trees for risk assessment. Plant Breed. Rev. 18, 251–303.

Green, R.E., Cornell, S.J., Scharleman, J.P.W., Balmford, A., 2005. Farming and the fate of wild nature. Science 307, 550–555.

Griffin, J.N., Méndez, V., Johnson, J.F., Jenkins, S.R., Foggo, A., 2009. Functional diversity predicts overyielding effect of species combination on primary productivity. Oikos 118, 37–44.

Hagen, M., Kissling, W.D., Rasmussen, C., de Aguiar, M.A.M., Brown, L.E., Carstensen, D.W., Alves-dos-Santos, I., Dupont, Y.L., Edwards, F.K., Genini, J., Guimaräes, P.R., Jenkins, G.B., Pedro, J., 2012. Biodiversity, species interactions and ecological networks in a fragmented world. Adv. Ecol. Res. 46, 89–210.

Haines-Young, R., Potschin, M., 2010. The links between biodiversity, ecosystem services and human well-being. In: Raffaelli, D., Frid, C. (Eds.), Ecosystem Ecology: A New Synthesis. Cambridge University Press, Cambridge.

Harmston, N., Filsell, W., Stumpf, M.P., 2010. What the papers say: text mining for genomics and systems biology. Hum. Genomics 5, 17–29.

Haughton, A., Champion, G., Hawes, C., Heard, M., Brooks, D., Bohan, D., Clark, S., Dewar, A., Firbank, L., Osborne, J., 2003. Invertebrate responses to the management of genetically modified herbicide–tolerant and conventional spring crops. II. Within-field epigeal and aerial arthropods. Philos. Trans. R. Soc. Lond. B Biol. Sci. 358, 1863–1877.

Hawes, C., Haughton, A.J., Osborne, J.L., Roy, D.B., Clark, S.J., Perry, J.N., Rothery, P., Bohan, D.A., Brooks, D.R., Champion, G.T., Dewar, A.M., Heard, M.S., Woiwod, I.P., Daniels, R.E., Young, M.W., Parish, A.M., Scott, R.J., Firbank, L.G., Squire, G.R., 2003. Responses of plants and invertebrate trophic groups to contrasting herbicide regimes in the Farm Scale Evaluations of genetically modified herbicide-tolerant crops. Philos. Trans. R. Soc. Lond. B Biol. Sci. 358, 1899–1913.

Hawes, C., Haughton, A.J., Bohan, D.A., Squire, G.R., 2009. Functional approaches for assessing plant and invertebrate abundance patterns in arable systems. Basic Appl.Ecol. 10, 34–42.

Hector, A., Bazeley-White, E., Loreau, M., Otway, S., Schmid, B., 2002. Overyielding in grassland communities: testing the sampling effect hypothesis with replicated biodiversity experiments. Ecol. Lett. 5, 502–511.

Ings, T.C., Montoya, J.M., Bascompte, J., Bluthgen, N., Brown, L., Dormann, C.F., Edwards, F., Figueroa, D., Jacob, U., Jones, J.I., Lauridsen, R.B., Ledger, M.E., et al., 2009. Ecological networks—beyond food webs. J. Anim. Ecol. 78, 253–269.

Inoue, K., 2001. Induction, abduction and consequence-finding. In: Proceedings of the International workshop on Inductive Logic Programming (ILP01). Springer-Verlag, Berlin, pp. 65–79.

Jenssen, T.K., Lægreid, A., Komorowski, J., Hovig, E., 2001. A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. 28, 21–28.

Jørgensen, S.E., Fath, B.D., 2004. Application of thermodynamic principles in ecology. Ecol. Complex. 1, 267–280.

Kakas, A.C., Riguzzi, F., 2000. Abductive concept learning. N. Gener. Comput. 18, 243–294.

Kakas, A.C., Kowalski, R.A., Toni, F., 1993. Abductive logic programming. J. Log. Comput. 2 (6), 719–770.

Korhonen, A., Seaghdha, D.O., Silins, I., Sun, L., Hogberg, J., Stenius, U., 2012. Text mining for literature review and knowledge discovery in cancer risk assessment and research. PLoS One 7, e33427.

Kraft, N.J.B., Valencia, R., Ackerly, D.D., 2008. Functional traits and niche-based tree community assembly in an Amazonian forest. Science 322, 580–582.

Lavorel, S., Grigulis, K., Lamarque, P., Colace, M.-P., Garden, D., Girel, J., Pellet, G., Douzet, R., 2010. Using plant functional traits to understand the landscape distribution of multiple ecosystem services. J. Ecol. 99, 135–147.

Layer, K., Riede, J.O., Hildrew, A.G., Woodward, G., 2010. Food web structure and stability in 20 streams across a wide pH gradient. Adv. Ecol. Res. 42, 265–299.

Layer, K., Hildrew, A.G., Jenkins, G.B., Riede, J., Rossiter, S.J., Townsend, C.R., Woodward, G., 2011. Long-term dynamics of a well-characterised food web: four decades of acidification and recovery in the Broadstone Stream model system. Adv. Ecol. Res. 44, 69–117.

Ledger, M.E., Brown, L.E., Edwards, F., Woodward, G., Milner, A.M., 2013. Drought impacts on the structure and functioning of complex food webs. Nat. Clim. Chang. 3, 223–227.

Lin, D., Chen, J., Watanabe, H., Muggleton, S.H., Jain, P., Sternberg, M., Baxter, C., Currie, R., Dunbar, S., Earll, M., Salazar, D., 2012. Does multi-clause learning help in real-world applications? In: Proceedings of the 21st International Conference on Inductive Logic Programming, LNAI 7207, Springer-Verlag, Berlin, pp. 221–237.

Lindeman, R.L., 1942. The trophic-dynamic aspect of ecology. Ecology 23, 399–418.

Lindroth, C.H., 1974. Coleoptera: Carabidae. In: Handbooks for the Identification of British InsectsIV, Royal Entomological Society of London, London. part 2.

Loeuille, N., Barot, S., Georgelin, E., Kylafis, G., Lavigne, C., 2013. Chapter six - Eco-evolutionary dynamics of agricultural networks: implications for sustainable management. Adv. Ecol. Res. 49, 339–435.

Loreau, M., Hector, A., 2001. Partitioning selection and complementarity in biodiversity experiments. Nature 412, 72–76.

Luff, M.L., 1978. Diel activity patterns of some field Carabidae. Ecol. Entomol. 3, 53–62.

Luff, M.L., 1996. Use of carabids as environmental indicators in grasslands and cereals. Ann. Zool. Fenn. 33, 185–195.

Massol, F., Petit, S., 2013. Chapter five - Interaction networks in agricultural landscape mosaics. Adv. Ecol. Res. 49, 291–338.

Mazandu, G.K., Mulder, N.J., 2012. Using the underlying biological organization of the mycobacterium tuberculosis functional network for protein function prediction. Infect. Genet. Evol. 12, 922–932.

Millennium Ecosystem Assessment, 2005. Duraiappah, A.K., Naeem, S., Agardy, T. (Eds.), Ecosystems and Human Wellbeing: Biodiversity Synthesis. World Resources Institute, Washington, DC.

Mitchell, T., 1997. Machine Learning. McGraw-Hill, New York.

Muggleton, S.H., 1991. Inductive logic programming. N. Gener. Comput. 8 (4), 295–318.

Muggleton, S.H., 1996. Stochastic Logic Programs. In: de Raedt, L. (Ed.), Advances in Inductive Logic Programming. IOS Press, Amsterdam, pp. 254–264.

Muggleton, S., Bryant, C., 2000. Theory completion using inverse entailment. In: Proceedings of the 10th International Conference on Inductive Logic Programming, Springer, Berlin, pp. 130–146.

Muggleton, S.H., De Raedt, L., 1994. Inductive logic programming: theory and methods. J. Log. Program. 19 (20), 629–679.

Muggleton, S.H., Lin, D., Pahlavi, N., Tamaddoni-Nezhad, A., 2013. Meta-interpretive learning: application to grammatical inference. Mach. Learn, http://dx.doi.org/10.1007/s10994-013-5358-3.

Mulder, C., Elser, J.J., 2009. Soil acidity, ecological stoichiometry and allometric scaling in grassland food webs. Glob. Chang. Biol. 15 (11), 2730–2738.

Mulder, C., Boit, A., Bonkowski, M., De Ruiter, P.C., Mancinelli, G., Van der Heijden, M. G., Rutgers, M., 2011. A belowground perspective on Dutch agroecosystems: how soil organisms interact to support ecosystem services. Adv. Ecol. Res. 44, 277–357.

Mulder, C., Boit, A., Mori, S., Arie Vonk, J., Dyer, S.D., Faggiano, L., Wall, D.H., 2012. Distributional (In) congruence of biodiversity-ecosystem functioning. Adv. Ecol. Res. 46, 1–88.

Mulder, C., Ahrestani, F.S., Bahn, M., Bohan, D.A., Bonkowski, M., Griffiths, B.S., Guicharnaud, R.A., Kattge, J., Krogh, P.H., Lavorel, S., Lewis, O.T., Mancinelli, G., et al., 2013. Chapter two – Connecting the green and brown worlds: allometric and stoichiometric predictability of above- and below-ground networks. Adv. Ecol. Res. 49, 69–175.

Odum, E., 1971. Fundamentals of Ecology, third ed. Saunders Press, New York.

O'Gorman, E.J., Emmerson, M.C., 2010. Manipulating interaction strengths and the consequences for trivariate patterns in a marine food web. Adv. Ecol. Res. 42, 301–419.

Pascual, M., Dunne, J.A., 2006. From small to large ecological networks in a dynamic world. In: Pascual, M., Dunne, J.A. (Eds.), Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, New York, pp. 3–24.

Perry, J.N., Rothery, P., Clark, S.J., Heard, M.S., Hawes, C., 2003. Design, analysis and statistical power of the Farm Scale Evaluations of genetically modified herbicide-tolerant crops. J. Appl. Ecol. 40, 17–31.

Petchey, O.L., Beckerman, A.P., Riede, J.O., Warren, P.H., 2008. Size, foraging, and food web structure. Proc. Natl. Acad. Sci. 105, 4191–4196.

Phipps, R.H., Park, J.R., 2002. Environmental benefits of genetically modified crops: global and European per- spectives on their ability to reduce pesticide use. J. Anim. Feed Sci. 11, 1–18.

Pocock, M.J.O., Evans, D.M., Memmott, J., 2012. The robustness and restoration of a network of ecological networks. Science 335, 973–977.

Polis, G.A., 1991. Complex trophic interactions in deserts—an empirical critique of food-web theory. Am. Nat. 138, 123–155.

Raffaelli, D.G., White, P.C.L., 2013. Ecosystems and their services in a changing world: an ecological perspective. Adv. Ecol. Res. 48, 1–70.

Raudsepp-Hearne, C., Peterson, G.D., Tengö, M., Bennett, E.M., Holland, T., Benessaiah, K., MacDonald, G.K., Pfeifer, L., 2010. Untangling the environmentalist's paradox: why is human well-being increasing as ecosystem services degrade? Bioscience 60, 576–589.

Ray, O., Broda, K., Russo, A., 2003. Hybrid Abductive Inductive Learning: a Generalisation of Progol. In: In Proceedings of 13th International Conference on Inductive Logic Programming, 311–328.

Reiss, J., Bridle, J.R., Montoya, J.M., Woodward, G., 2009. Emerging horizons in biodiversity and ecosystem functioning research. Trends Ecol. Evol. 24, 505–514.

Reuman, D.C., Cohen, J.E., 2004. Trophic links' length and slope in the Tuesday Lake food web with species' body mass and numerical abundance. J. Anim. Ecol. 73 (5), 852.

Reuman, D.C., Mulder, C., Banašek-Richter, C., Cattin Blandenier, M.-F., Breure, A.M., Den Hollander, H.A., Kneitel, J.M., Raffaelli, D., Woodward, G., Cohen, J.E., 2009. Allometry of body size and abundance in 166 food webs. Adv. Ecol. Res. 41, 1–44.

Roberts, M.J., 1993. The spiders of Great Britain and Ireland, compact edition, part 1. Harley Books, Colchester, UK.

Rott, A., Godfray, H., 2000. The structure of a leafminer-parasitoid community. J. Anim. Ecol. 69 (2), 274–289.

Roy, D.B., Bohan, D.A., Haughton, A.J., Hill, M.O., Osborne, J.L., Clark, S.J., Perry, J.N., Rothery, P., Scott, R.J., Brooks, D.R., Champion, G.T., Hawes, C., Heard, M.S., Firbank, L.G., 2003. Invertebrates and vegetation of field margins adjacent to crops subject to contrasting herbicide regimes in the Farm Scale Evaluations of genetically modified herbicide-tolerant crops. Philos. Trans. R. Soc. Lond. B Biol. Sci. 358, 1879–1898.

Santos, J.C.A., Nassif, H., Muggleton, S.H., Sternberg, M.J.E., Page, C.D., 2012. Automated identification of protein-ligand interaction features using inductive logic programming: a hexose binding case study. BMC Bioinforma. 13:162.

Smith, V., Bohan, D.A., Clark, S.J., Haughton, A.J., Bell, J.R., Heard, M.S., 2008. Weed and invertebrate community compositions in arable farmland. Arthropod Plant Interact. 2 (1), 21–30.

Southwood, T.R.E., Leston, D., 1959. Land and Water Bugs of the British Isles. Frederick Warne, London.

Speight, M.C.D., Martinez, M., Luff, M.L., 1986. The Asaphidion (Col.: Carabidae) species occurring in Great Britain and Ireland. Proc. Trans. Br. Ent. Nat. Hist. Soc. 19, 17–21.

Squire, G.R., Brooks, D.R., Bohan, D.A., Champion, G.T., Daniels, R.E., et al., 2003. On the rationale and interpretation of the farm-scale evaluations of genetically-modified herbicide-tolerant crops. Philos. Trans. R. Soc. B 358, 1779–1800.

Stachowicz, J.J., 2001. Mutualism, facilitation, and the structure of ecological communities. Bioscience 51, 235–246.

Sternberg, M.J.E., Tamaddoni-Nezhad, A., Lesk, V.I., Kay, E., Hitchen, P.G., Cootes, A., Alphen, L.B., Lamoureux, M.P., Jarrell, H.C., Rawlings, C.J., Soo, E.C., Szymanski, C.M., Dell, A., Wren, B.W., Muggleton, S.H., 2013. Gene function hypotheses for the campylobacter jejuni glycome generated by a logic-based approach. J. Mol. Biol. 425 (1), 186–197.

Swanson, D.R., 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect. Biol. Med. 30, 7–18.

Tamaddoni-Nezhad, A., Chaleil, R., Kakas, A., Muggleton, S.H., 2006. Application of abductive ILP to learning metabolic network inhibition from temporal data. Mach. Learn. 64, 209–230.

Tamaddoni-Nezhad, A., Chaleil, R., Kakas, A., Sternberg, M., Nicholson, J., Muggleton, S.H., 2007. Modeling the effects of toxins in metabolic networks. IEEE Eng. Med. Biol. 26, 37–46.

Tamaddoni-Nezhad, A., Bohan, D., Raybould, A., Muggleton, S., 2012a. Machine learning a probabilistic network of ecological interactions. In: Proceedings of the 21st International Conference on Inductive Logic Programming, LNAI 7207. Springer, Berlin, pp. 332–346.

Tamaddoni-Nezhad, A., Afroozi Milani, G., Bohan, D., Raybould, A., Muggleton, S., 2012b. Towards automatic construction and corroboration of food webs. In: Proceedings of the European Conference on Machine Learning Workshop on Learning and Discovery in Symbolic Systems Biology (ECML/LDSSB 2012) Bristol, pp. 95–102.

Thompson, R.M., Brose, U., Dunne, J.A., Hall Jr., R.O., Hladyz, S., Kitching, R.L., Martinez, N.D., Rantala, H., Romanuk, T.N., Stouffer, D.B., Tylianakis, J.M., 2012. Food webs: reconciling the structure and function of biodiversity. Trends Ecol. Evol. 27, 689–697.

Tilman, D., Lehman, C.L., Bristow, C.E., 1998. Diversity-stability relationships: statistical inevitability or ecological consequence? Am. Nat. 151, 277–282.

Tilman, D., Cassman, K.G., Matson, P.A., Naylor, R., Polasky, S., 2002. Agricultural sustainability and intensive production practices. Nature 418, 671–677.

Tixier, P., Peyrard, N., Aubertot, J.-N., Gaba, S., Radoszyck, J., Caron-Lormier, G., Vinatier, F., Mollot, G., Sabbadin, R., 2013. Chapter seven – Modelling interaction networks for enhanced ecosystem services in agroecosystems. Adv. Ecol. Res. 49, 437–480.

Traugott, M., Kamenova, S., Ruess, L., Seeber, J., Plantegenest, M., 2013. Chapter three – Empirically characterising trophic networks: what emerging DNA-based methods, stable isotope and fatty acid analyses can offer. Adv. Ecol. Res. 49, 177–224.

Tsunoyama, K., Amini, A., Sternberg, M.J.E., Muggleton, S.H., 2008. Scaffold hopping in drug discovery using inductive logic programming. J. Chem. Inf. Model. 48 (5), 949–957.

Unwin, D.M., 1988. A key to the families of British Coleoptera (and Strepsiptera). Field Stud. 6, 149–197.

Van Veen, F.J.F., Müller, C.B., Pell, J.K., Godfray, H.C.J., 2008. Food web structure of three guilds of natural enemies: predators, parasitoids and pathogens of aphids. J. Anim. Ecol. 77, 191–200.

Vellend, M., 2010. Conceptual synthesis in community ecology. Q. Rev. Biol. 8 (2), 183–206.

Woodward, G., Speirs, D.C., Hildrew, A.G., 2005a. Quantification and resolution of a complex, size-structured food web. Adv. Ecol. Res. 36, 85–135.

Woodward, G., Ebenman, B., Emmerson, M., Montoya, J.M., Olesen, J.M., Valido, A., Warren, P.H., 2005b. Body size in ecological networks. Trends Ecol. Evol. 20, 402–409.

Woodward, G., Brown, L.E., Edwards, F.K., Hudson, L.N., Milner, A.M., Reuman, D.C., Ledger, M.E., 2012. Climate change impacts in multispecies systems: drought alters food web size structure in a field experiment. Philos. Trans. R. Soc. B Biol. Sci. 367 (1605), 2990–2997.

Xiong, J., Rayner, S., Luo, K., Li, Y., Chen, S., 2006. Genome wide prediction of protein function via a generic knowledge discovery approach based on evidence integration. BMC Bioinforma. 7 (1), 268.