Detection of Multi-clustered Genes and Community Structure for the Plant Pathogenic Fungus *Fusarium graminearum*

Laura Bennett[1], Artem Lysenko[2], Lazaros G. Papageorgiou[3], Martin Urban[4], Kim Hammond-Kosack[4], Chris Rawlings[2], Mansoor Saqi[2,*], and Sophia Tsoka[1,*]

[1] Department of Informatics, School of Natural and Mathematical Sciences, King's College London, Strand, London, WC2R 2LS, UK
{laura.bennett,sophia.tsoka}@kcl.ac.uk
[2] Department of Computational and Systems Biology, Rothamsted Research, Harpenden, Herts, AL5 2JQ, UK
{chris.rawlings,mansoor.saqi,artem.lysenko}@rothamsted.ac.uk
[3] Centre for Process Systems Engineering, Department of Chemical Engineering, University College London, Torrington Place, London, WC1E 7JE, UK
l.papageorgiou@ucl.ac.uk
[4] Department of Plant Biology and Crop Sciences, Rothamsted Research, Harpenden, Herts, AL5 2JQ, UK
{martin.urban,kim.hammond-kosack}@rothamsted.ac.uk

Abstract. Exploring the community structure of biological networks can reveal the roles of individual genes in the context of the entire biological system, so as to understand the underlying mechanism of interaction. In this study we explore the disjoint and overlapping community structure of an integrated network for a major fungal pathogen of many cereal crops, *Fusarium graminearum*. The network was generated by combining sequence, protein interaction and co-expression data. We examine the functional characteristics of communities, the connectivity and multi-functionality of genes and explore the contribution of known virulence genes in community structure. Disjoint community structure is detected using a greedy agglomerative method based on modularity optimisation. The disjoint partition is then converted to a set of overlapping communities, where genes are allowed to belong to more than one community, through the application of a mathematical programming method. We show that genes that lie at the intersection of communities tend to be highly connected and multifunctional. Overall, we consider the topological and functional properties of proteins in the context of the community structure and try to make a connection between virulence genes and features of community structure. Such studies may have the potential to identify functionally important nodes and help to gain a better understanding of phenotypic features of a system.

Keywords: Community structure, overlapping communities, integrated networks, multi-functional genes, phytopathogenic fungi.

* Corresponding authors.

D. Gilbert and M. Heiner (Eds.): CMSB 2012, LNCS 7605, pp. 69–86, 2012.

1 Introduction

In addition to genome sequence data, a large amount of multiple complimentary types of biological data is available for many organisms, such as gene expression, protein interactions and phenotypic information. Integration of data from various sources gives rise to complex networks where nodes are proteins (or other gene products) and edges capture intricate associations between them [1, 2]. The analysis of topological features in these networks can not only uncover information about the underlying functional properties of individual nodes and relevant gene products, but it can also reveal the principles of how genes group to assemble entire cellular systems.

Network analysis is therefore a popular means of investigating the link between topological and functional features in biological systems. Community structure detection in biological networks is widely employed to derive an understanding of molecular interactions. The standard community structure detection problem involves the identification of a partition of a complex network into disjoint communities (also sometimes known as modules or clusters) such that interactions within a community are maximised and interactions between communities are minimised. Many approaches exist, including divisive [3], agglomerative [4], spectral [5] and mathematical programming methods [6-8]**.** The communities detected represent semi-independent functional units of an entire system, where members are likely to share some common characteristic. The identification of such communities allows information on members with unknown functional properties to be inferred.

However, the constraint of disjoint communities, where a node can only belong to one community, may not offer the most realistic abstraction of a system. For example, some proteins can carry out more than one task or belong to more than one protein complex [9, 10]. The identification of overlapping communities can reveal the multi-functionality of nodes and determines which nodes act as bridges between different functional groups or co-ordinate multiple tasks so as to hold the system together. For example, such roles are important in social networks modelling the spread of disease, where potential immunisation targets are individuals that bridge communities [11]. Moreover, in a biological system, a multi-clustered gene may act as a communicator, transferring biological information between functional units [12]. It is currently not well explored whether genes/proteins that have multiple community membership also possess particular functional and topological properties. Here we test such hypotheses in the case study of the plant pathogenic fungus *Fusarium graminearum.*

Although a less well-covered area of research than standard community structure detection, several methods for the detection of overlapping communities have been proposed. One of the first methods was the Clique Percolation method [13]. Subsequently, a wide range of approaches followed including spectral methods [14], non-negative matrix factorisation of various feature matrices [15, 16], local optimisation of a fitness function [17, 18], greedy agglomerative algorithms [19] and mathematical programming approaches [8].

Integrated functional networks can provide a framework to begin to explore genotype-phenotype relationships. For example if a gene disruption experiment of a given gene leads to a certain outcome (a given phenotype) the network may provide

clues to suggest the underlying mechanisms that are affected and may aid thereby hypothesis generation. Clustering such integrated networks into disjoint and overlapping communities can identify functional communities and give insight into higher levels of biological organisation [20, 21]. As proteins take part in multiple processes a better description of the underlying biological themes may be provided by consideration of the overlaps between communities.

The Ascomycete fungus *Fusarium graminearum* is a major pathogen of wheat and other cereal crops. The complete genome sequence (with about 13,718 protein coding genes) of *Fusarium graminearum* has been determined [22] and additional data on gene expression [23] and predicted protein interactions [24] also exist. Floral infections by *Fusarium* can have a significant impact on grain yield and quality. In addition, infection by the fungus leads to contamination of the grain by various mycotoxins including deoxynivalenol (DON), which makes the grain harmful for human consumption and also for animal feed. In this study we explore the modular properties of an integrated network for *Fusarium graminearum*. Detection of disjoint and overlapping community structure is employed as a means of elucidating topological-functional relationships in the pathogen. Additionally, we relate the modular organisation of the network to virulence genes known to be required for pathogen infection and disease formation. Such an analysis may lead to better understanding of phenotypic features of the system, for example, potential insights into infection-related pathways.

2 Methods

An integrated network for *Fusarium graminearum* was constructed using information from sequence similarity, co-expression and predicted protein interactions (PPI). The sequence similarity network was constructed from all-versus-all sequence matching of the proteins in version 3.2 of the *Fusarium graminearum* annotation (at ftp://ftpmips.gsf.de/FGDB/v32) implemented on a TimeLogic® Tera-BLAST™ (Active Motif Inc., Carlsbad, CA) system with a threshold E-value for bidirectional best hits of 10^{-6}. Co-expression information was obtained from the publicly available set of Fusarium expression studies from PLEXdb [23] that used Fusarium Affymetrix GeneChip arrays. The similarity of expression profiles was measured using weighted Pearson correlation coefficient, according to the method in [25]. The PPI information was taken from the predicted core PPI of [26]. Two nodes are linked if any of the following properties is satisfied: (i) a bi-directional sequence similarity BLAST hit comprised of unidirectional hits with an expected value of less than 10^{-6}, (ii) correlation of gene expression with an absolute value of Pearson correction greater than 0.88, or (iii) PPI link from the dataset from [26]. Integration of the various data sources was carried out using the Ondex data integration platform [27, 28].

The community structure of this network was detected using the greedy agglomerative method known as Louvain [4], where nodes are allocated communities based on the maximum increase in the Newman modularity measure [29]. This results in a *hard partition* of the network into disjoint communities. The hard partition

can then be converted into a *soft partition* where communities are allowed to overlap, using the method described in [8]. This mathematical programming approach fixes the community membership of all nodes that only interact with nodes in their own community in the hard partition (*isolated nodes*), whereas nodes that form interactions across communities (*border nodes*) can belong to more than one community. A mixed integer non linear programming (MINLP) model, known as OverWeiMod, is formulated to optimise the sum of the *community strength* (CS) [18] across all communities according to node-community assignments.

The inclusion of an overlapping parameter *r* allows the user to control the extent of overlapping, where the larger the value of *r*, the smaller the overlap**.** The choice of *r* is user-defined and we show here that certain topological and functional criteria can indicate a range of values. The output of OverWeiMod is a partition of the network nodes belonging to more than one community, which we define as *multi-clustered nodes* (as opposed to *mono-clustered* nodes). The *belonging coefficient* (BC) gives a measure of strength of membership of a node to a community according to the community's gain in CS with the presence of the node**.** For example, a node belonging to two communities with BC equal to 0.5 in both cases belongs equally to the two communities. However, if the node belongs to one community with a BC equal to 0.7 and to the other with a BC equal to 0.3, this indicates a stronger attachment to the first community over the second.

3 Results

3.1 Disjoint Community Structure Detection

The integrated *Fusarium graminearum* network comprises 9521 nodes (proteins), 80997 links and is made up of 439 disconnected components. Table 1 shows the distribution of sizes of the connected components. This analysis focuses on the largest connected component of 8364 nodes and 79931 links, as community structure of smaller components is of limited scope.

Table 1. Connected components in the integrated network

No. of nodes	2	3	4	5	6	7	9	10	11	16	8364
No. of components	288	101	23	10	5	3	3	2	2	1	1

The main component of the network is partitioned by Louvain [4], which detects a partition of 91 disjoint communities with modularity equal to 0.7973. The resultant community structure has an 'uneven' community size distribution, with 89 communities of size <500 and 2 large communities with 1007 and 1951 nodes. The output of the Louvain method was compared with another well-known community structure method, QCUT [5], based on the spectral properties of the network Laplacian. QCUT finds a partition with 53 communities (modularity equal to 0.7665), 51 of which have <500 nodes and two larger communities with 1198 and 2968 nodes.

This is in agreement with the 'uneven' community structure found by the Louvain method. Based on Louvain finding the slightly larger value of modularity, we use this hard partition in the following analysis.

The above disjoint community structure is illustrated by a 'meta-view' of the partition in Figure 1, with (i) the size of communities, (ii) the number of shared nodes across communities in the overlapping community structure (discussed in section 3.2) and (iii) their functional content. The functional coherence of a community was described by the Average Information Content of the Most Informative Common Ancestor set (AIC-MICA) a metric defined in [28], which can be used to gauge the degree of commonality of gene annotations in a particular set. This method works by identifying a set of representative Most Informative Common Ancestor (MICA) terms, where the information content (IC) is calculated based on how frequently a particular annotation is found in an annotation set for a given species. The MICA term is defined as a term in a hierarchically-organised ontology graph, which has the highest possible IC value whilst also acting as a subsumer for all terms in a particular set. The AIC-MICA approach takes as input a set of annotated entities and returns a non-redundant set of MICA terms that are applicable to at least a certain fraction of these entities, as specified by the user. The AIC-MICA statistic itself is an average of their IC values, which can serve as an indicator of annotation commonality within a set of entities. A higher value would indicate that most of the MICAs for the set are found lower in the ontology and therefore commonality in annotation is at a level with higher specificity. Here we have used the Gene Ontology (GO) [30] in which the functional role of a gene product is described at three levels: biological process (BP), molecular function (MF) and cellular component (CC).

We looked at the annotation for all three aspects of GO for the communities in the Louvain partition with at least 5 annotated nodes and used the AIC-MICA approach to find the most specific terms applicable to at least 60% of the nodes. We find that 43 communities are assigned a term from the BP aspect of the Gene Ontology, 52 are assigned a term from the MF aspect of GO and 35 are assigned a term from the CC aspect of GO. Figure 1 shows the corresponding MICA BP terms and their percentage of coverage for the largest communities. Some highly functionally coherent communities detected were "transport", "blood vessel morphogenesis" and "carbohydrate metabolic process" (communities 3, 31 and 88 respectively, 100% coverage) and "oxidation-reduction process", "transport" and "regulation of transcription, DNA-dependent" (communities 28, 60 and 76 respectively, coverage >90%). Other communities with a strong functional coherence point to 'vitamin transport' (community 78), 'nucleotide biosynthetic process' and 'serine family amino acid metabolic process'. Expectedly, larger communities show less homogeneous functional content and therefore a broader GO term is assigned, e.g. community 79, the largest community is assigned "cellular process". Overall, the hard partition detected by the Louvain method appears to find some biologically coherent communities.

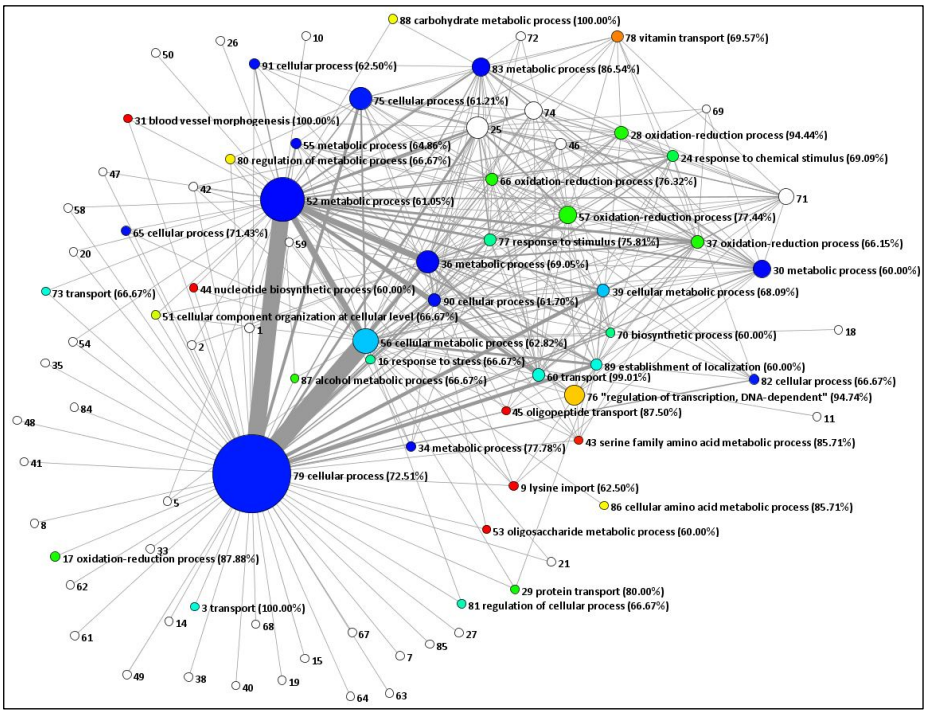

Fig. 1. The meta-view of the hard partition of the main component detected by the Louvain method, where nodes represent communities. The thickness of the links between the communities corresponds to the number of genes that are shared between communities in the overlapping community structure discussed in Section 3.2. For the larger communities, the MICA (BP) term is shown next to the corresponding community and the corresponding percentage of coverage (visualisation generated in Ondex [27, 28]).

3.2 Overlapping Community Structure

The hard partition of the main connected component is converted to a soft partition with overlapping communities using the mathematical programming method, OverWeiMod [8]. The hard partition results in 3877 border nodes, which are the potential multi-clustered nodes. As mentioned earlier, the community membership of 4487 isolated nodes that are only associated to intra-community edges, are fixed and do not change in the course of the conversion procedure. In other words, the MINLP is solved with only border nodes allowed to be assigned to multiple communities. Figure 2 shows the results for r ranging from 0.4 to 1.1. The range of values of r is chosen to some extent arbitrarily and we discuss the suitability of the range in forthcoming sections. Table 2 shows how the number of communities that a multi-clustered node belongs to changes with r. We find that for $0.4 \le r \le 0.5$ the multi-clustered nodes belong to up to 6 communities, but as r increases, and the extent of overlap decreases, this range also decreases. When $r = 1.1$ multi-clustered nodes only belong to two communities maximum.

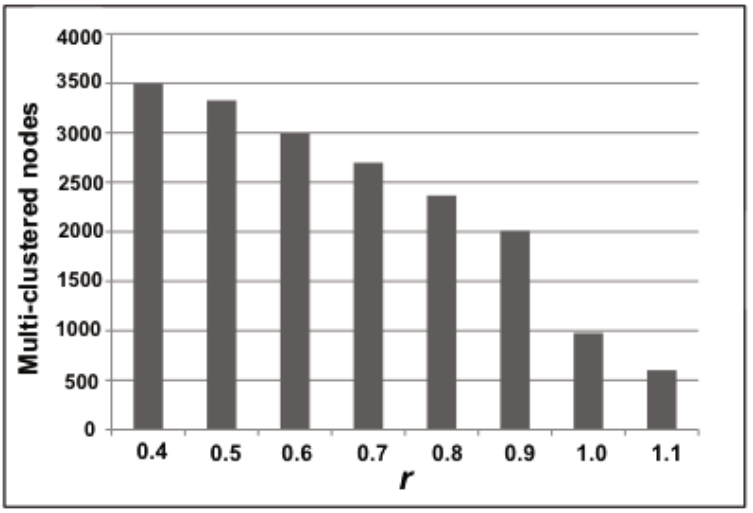

Fig. 2. The number of multi-clustered nodes detected by OverWeiMod for $0.4 \le r \le 1.1$

As previously mentioned, the overlapping community structure problem may be subject to multiple interpretations according to the underlying problem statement and user requirements. Our approach is to consider a hard partition of a network and examine the nodes that form interactions across community borders, to assess their associations with communities other than their own according to the hard partition employed. As described in Section 1, many different approaches to the overlapping community structure detection problem exist and the variation in methodology can affect results considerably [8]. Consequently a direct comparison between methods may not be a fair evaluation of performance. In any case, for real life networks the 'real' cover is not known and so therefore the aim is to show that according to the user's interpretation of the problem, the chosen method finds biologically relevant solutions.

In order to explore the robustness of the approach implemented in OverWeiMod we consider the greedy agglomerative method, known as the Overlapping Cluster Generator (OCG) method which has been shown to identify multifunctional proteins in PPI networks [19]. OCG is based on an adapted modularity measure applicable to overlapping communities and bears a similar methodological framework to OverWeiMod. An initial partition of 'centred cliques' is generated. Then, the elements are joined together iteratively in order of increasing average modularity gain, where modularity in this case is an overlapping equivalent of the Newman modularity [29].

OCG detects a soft partition of the main component of the integrated network with 808 communities with 3877 multi-clustered nodes. Of the 808 modules, 201 comprise only 2 nodes, 47 have 3 nodes and 33 have 4 nodes and the remaining modules range from between 5 and 211 nodes. Of the 3877 nodes, 1628 nodes belong to 2 communities, 692 belong to 3, 440 belong to 4 and the remaining multi-clustered nodes belong to between 5 and 58 communities. Figure 3 shows the breakdown of multi-clustered nodes according to the method they were detected by. In each case a considerable level of agreement can be seen between the two methods.

Table 2. Number of communities the multi-clustered nodes belong to ($0.4 \le r \le 1.1$)

	Number of communities				
r	**2**	**3**	**4**	**5**	**6**
0.4	2297	804	263	96	29
0.5	2330	718	208	66	4
0.6	2354	512	106	22	0
0.7	2305	319	59	14	0
0.8	2135	217	14	0	0
0.9	1868	138	1	0	0
1	960	16	0	0	0
1.1	601	0	0	0	0

Variation in results is due to fundamental differences in methodology. In particular, OCG starts its agglomerative procedure with an initial cover of the network comprising a large number of modules, which are subsequently fused until one of three stopping criteria are achieved. Consequently, if the stopping criteria are met after relatively few iterations, the resulting number of modules is high. The number of overlapping modules in the final cover of the network detected by OverWeiMod on the other hand depends on the method used to find the hard partition. In this study we use a method based on modularity optimisation, a well-recognised approach to community structure detection, employed by many methods and that has been shown to find relevant solutions in bioinformatics applications [20, 31]. However, the debate about which is the most realistic partition of the network, is beyond the scope of this study. Our main aim being to show that our method assigns structurally and functionally important nodes with biological significance to multiple modules. Here we put less importance on directly comparing methods in terms of the nodes they find to be multi-clustered and more on what 'type' of node is multi-clustered. Such properties are discussed in the next section.

Overall the results of OverWeiMod and OCG vary greatly in terms of (i) number of modules in each of the partitions of the network and (ii) the number of communities that the multi-clustered nodes can belong to. Despite these differences, there are still a considerable number of proteins that are multi-clustered by both methods.

3.3 Evaluation of Multi-clustered Nodes

If we consider genes that belong to more than one community as bridges, connectors between functional units, or communicators that spread information in a system, one would imagine them to exhibit properties that reflect such capabilities. In this section, we consider features that distinguish multi-clustered from mono-clustered genes and additionally show how these features can indicate an appropriate range of values for the overlapping parameter, *r*.

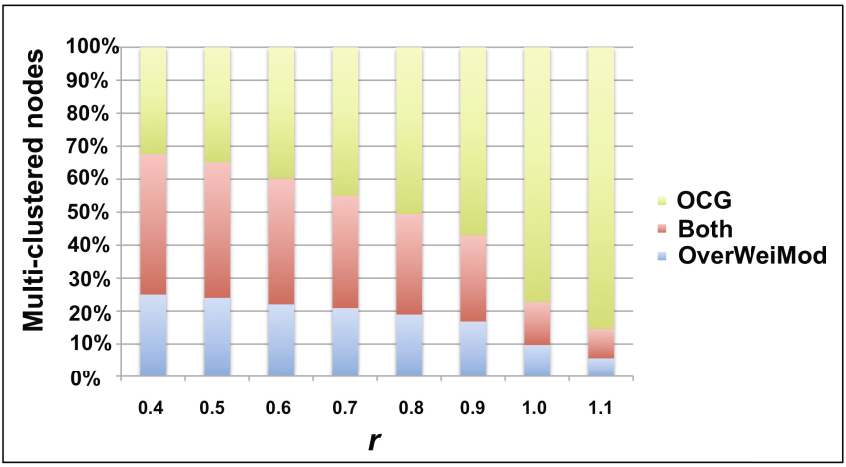

Fig. 3. Breakdown of multi-clustered proteins according to the methods they were found by. For each value of parameter *r* in the figure, the set of multi-clustered nodes detected by OCG remains constant.

3.3.1 Node Degree

We compare the average degree of the nodes with multiple community membership with the equivalent values for nodes that belong to only one community. For $0.4 \le r \le 1.1$, the range of values of parameter *r* tested in section 3.2, multi-clustered nodes have a higher average degree than the mono-clustered nodes (Table 3). We determine if the population means are statistically significantly different using the Mann–Whitney–Wilcoxon U test as implemented in the R statistical computing environment [32], where a p-value < 0.01 is significant. For all values of *r* tested, the average node degree of the multi-clustered nodes is significantly larger than the average degree of the mono-clustered nodes (Table 3). This result indicates that multi-clustered genes tend to have a higher number of interactions than those that belong to only one community. This result is intuitive, since if proteins lying in the overlapping sections play a connector role in the system, interacting with two or more communities, then it reasonable that they are more likely to interact with more partners compared to isolated nodes. This result indicates that multi-clustered nodes are topologically significant, however more important perhaps is to establish the likely functional roles of such connector nodes, as described in section 3.3.2.

It is also worth mentioning that, although the inclusion of parameter *r* is advantageous as it offers greater flexibility to the user, it is also necessary to determine a reasonable range of values for each network. Node degree can be used as an indicator of such a range if we assume that nodes belonging to more than one community should have more interactions than those that do not. Therefore, in terms of node degree, these results indicate our range of values for *r* is reasonable for this network. In the next section we show that by looking at the functionality of the multi-clustered proteins can be reduced the range of values.

Table 3. The average degree of multi- and mono-clustered nodes detected by OverWeiMod

r	Multi-clustered	Mono-clustered	p-value
0.4	26.98	13.46	< 2.2e-16
0.5	27.37	13.66	< 2.2e-16
0.6	28.13	14.09	< 2.2e-16
0.7	28.33	14.73	< 2.2e-16
0.8	29.34	15.08	< 2.2e-16
0.9	28.91	16.02	< 2.2e-16
1	26.67	18.11	< 2.2e-16
1.1	31.95	18.12	< 2.2e-16

3.3.2 Gene Ontology Term Analysis

Where node degree offers a topological measure for distinguishing between multi-clustered and mono-clustered genes, GO annotations can offer a distinction based on functional features. As seen in [19], and in line with our interpretation of multi-clustered nodes as bridges between multiple functions, one would expect multi-clustered genes to be associated with a higher number of GO annotations than those belonging to only one community.

To test this hypothesis, we compare the number of Gene Ontology terms annotated to multi-clustered and mono-clustered genes to determine which group has a significantly higher number of GO annotations. The annotations are taken from the MIPS *Fusarium graminearum* database [33]. The *F. graminearum* genome has 4915 genes annotated with 13,883 GO terms from all three aspects of GO (molecular function (MF), biological process (BP) and cellular component (CC)). As the complete genome sequence comprises 13,718 protein coding genes, only roughly a third of the genome is annotated. In the integrated network, 4251 proteins have no annotations, 4311 are annotated with at least one GO term and when considering each GO category, there are more proteins unannotated than annotated in the network.

Due to the high number of genes without GO annotations, we first consider all three aspects of GO terms together (ALL GO). Each gene in the main component of the *Fusarium* network is mapped to its GO terms where possible, and the average number of GO terms for all four categories (ALL GO, MF, BP and CC) is calculated. For ALL GO, BP and CC the multi-clustered proteins have a statistically significant higher average number of GO terms than the mono-clustered proteins, for $0.4 \leq r \leq 0.9$. For MF, the average number of GO terms for the multi-clustered proteins is not significantly higher than mono-clustered nodes for all values of *r*. Overall, it seems that multi-functionality of multi-clustered nodes is better endorsed at the BP level (i.e. in terms of participation in multiple biochemical pathways), rather than the MF (i.e. the individual biochemical tasks) of the corresponding gene product. Future work to consolidate such observations by accounting for lack of comprehensive annotations for this genome would be recommended.

Similar to the node degree analysis, we can use the number of GO terms assigned to genes to suggest a potential rage values for *r* where OverWeiMod detects multi-clustered genes with desirable properties. If we assume that multi-clustered nodes are multifunctional, we can use the ALL GO count as an indicator that considering values of *r* between 0.4 and 0.9 inclusive is reasonable for this network.

Table 4. The significance value of the difference between average number of GO annotations for multi- and mono-clustered nodes. Significant p-values are shown in bold (<0.01).

r	ALL GO	MF	BP	CC
0.4	**1.62E-04**	2.31E-01	**4.40E-06**	**1.06E-03**
0.5	**1.95E-03**	1.23E-01	**2.09E-05**	**7.37E-04**
0.6	**7.23E-03**	5.35E-01	**1.29E-03**	**5.33E-04**
0.7	**7.74E-04**	9.28E-01	**5.96E-04**	**9.05E-04**
0.8	**3.48E-04**	3.95E-01	**5.72E-05**	**1.84E-03**
0.9	**4.92E-03**	4.94E-01	**4.91E-07**	**1.03E-04**

3.3.3 Functional Cartography of Multi-clustered Genes

Throughout this study we hypothesise that multi-clustered nodes play an important role topologically and functionally in the network, considering them as bridges or communicators between functional units, helping to maintain the structure of the system. This idea has been reinforced by showing that (i) they are more connected than mono-clustered nodes and (ii) for some aspects of the Gene Ontology, they have more functional annotations that mono-clustered nodes. We relate the overlapping community structure to a node role classification scheme proposed in [20]. Each node is assigned a role based on its position in the hard partition of the network. A node's role is characterised according to two measures: within-community degree z-score and participation coefficient (see [20] for details). The within-community degree z-score measures how well a node is connected with nodes in its own community and the participation coefficient measures how uniformly the nodes' links are distributed among the other communities in the partition.

The node classification scheme in [20] can be summarised as follows. Based on the within-community degree z-score, nodes are classified as hubs and non-hubs, where hubs have a higher number of links with nodes in their own communities. Non-hubs are then classified into 4 roles: R1, ultra-peripheral nodes, R2, peripheral nodes, R3, non-hub connector nodes and R4, non-hub kinless nodes. Hubs are also classified into 3 roles: R5, provincial hubs, R6, connector hubs and R7, global kinless hubs. Both R3 and R6 nodes are labelled 'connector' nodes according to the classification scheme as they have by definition a large participation coefficient, indicating a high distribution of links with communities other than their own. Consequently, the removal of these nodes may result in poorly connected communities or even the disconnection of communities and therefore having an impact on the global structure. On the application of the classification scheme to metabolic networks it is found that R3 and R6 nodes are the most preserved across the species tested, suggesting that their role is more structurally relevant and similar results are predicted for other systems, including protein interaction and gene regulation networks [20].

We assign node roles to the *Fusarium graminearum* network. The distribution of node role types is shown in Table 5. We determine whether the proportion of R3 and R6 nodes is significantly higher in multi-clustered than mono-clustered nodes indicating that the multi-clustered nodes do indeed have a bridge/connector role in the system. For $r = 0.4$, all 165 R3 nodes and all 50 R6 nodes belong to the set of

multi-clustered nodes. For $0.4 \le r \le 1$, there is a higher proportion of R3 nodes in the multi-clustered set than the mono-clustered set, and for R6 nodes, the range is $0.4 \le r \le 1.1$. We use the Fisher's exact test to decide if any difference in proportions is significant. We find that there is a significantly higher proportion of R3 nodes in the multi-clustered nodes for $0.4 \le r \le 0.8$ and similarly for R6 nodes for $0.6 \le r \le 0.9$ (results in Table 6).

Table 5. Node role type distribution

Role Types	R1	R2	R3	R4	R5	R6	R7
No. of nodes	4669	3323	165	0	157	50	0

The node classification scheme is employed to help describe the type of node that lies within the intersections of communities, offering a more sophisticated topological description than node degree alone. We find that nodes described as connectors are significantly enriched in the multi-clustered nodes. These are node roles that have previously been shown to be more structurally relevant than other node roles in biological networks [20]. This goes some way to supporting our claims that multi-clustered nodes play an important part in anchoring the communities of the network and thus contributing to the global functioning of the system. Furthermore, OverWeiMod is successfully detecting such nodes.

Table 6. The FDR-adjusted p-values for difference in proportion of R3 and R6 nodes in multi- and mono-clustered sets (significant values shown in bold)

r	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	**1.1**
R3	**1.31E-07**	**1.61E-09**	**1.11E-07**	**1.08E-05**	**6.90E-05**	3.96E-02	5.22E-01	1.38E-02
R6	2.21E-02	1.19E-02	**9.36E-05**	**1.08E-05**	**9.28E-07**	**3.74E-07**	1.65E-01	1.41E-01

4 Verified Virulence Genes

The analysis described next is to try to link the modular structure of the integrated network to known virulence genes. The experimentally verified *Fusarium graminearum* virulence genes known to be required for different aspects of the infection and disease formation process, were extracted from the Pathogen-Host Interaction database [30-32] and others were manually obtained from the scientific literature. We found that 79 out of 98 of the verified virulence proteins map to the integrated network of which 75 are in the main component. We also refer to the set of 75 genes as the verified virulence (VV) nodes.

First, the distribution of the VV nodes in the communities of the integrated network is considered. These nodes appear in only 15 of the 91 communities (Table 7), with the largest community (community 79, Figure 1) containing over half of the VV nodes (39 out of 75). For each of the 15 communities, we test if the community has a statistically significant higher proportion of VV nodes than the rest of the network using Fisher's exact test. Only the largest community and another of size 48

(community 80 in Figure 1), with 7 VV nodes, encompass a statistically significant high proportion of the proteins (FDR adjusted p-values 8.38E-07 and 1.35E-06 respectively). Although community 79 does contain a significant number of VV nodes, the corresponding BP MICA term, "cellular process", has an AIC of only 1.29, indicating that this is highly functionally diverse. Community 80 however is more coherent with an AIC of 4.88, and its BP MICA term is "regulation of metabolic process". The 7 VV genes in community 80 are all predicted to be transcription factors of the Zinc finger (Cys_2His_2) type [34].

We observe here that the verified virulence nodes are concentrated in two communities, suggesting that pathogenicity processes may be linked to specific parts of the network. Even at this preliminary stage, such effects indicate that there is a clear link between functional features and related interactions at network level. Similar analyses of community structure can therefore support further reverse molecular genetics and biochemistry experiments and thereby determine how these communities relate to underlying biochemical pathways.

We further partition community 79 and look at the distribution of the VV nodes in this new hard partition. The Louvain method detects a partition with 19 communities. The 39 VV nodes that are in community 79 in the hard partition of the main component are found in 8 of the communities of the re-partitioned community 79 (Table 8). Again, checking for overrepresentation of VV nodes in the individual communities shows that no community is significantly enriched. This may reflect the fact that proteins involved in a wide range of processes have a role in virulence because of the overall complexity of the infection process. In PHI-base [35, 36], there are many proteins with a "general" functional role such in basic metabolism, signal transduction and transcription and far fewer with a specific function such as toxin biosynthesis or infection structure formation.

Table 7. Distribution of the verified virulence (VV) nodes among communities and corresponding biological process average information content (BP AIC)

Comm. no.	7	16	28	39	51	52	56	57	64	71	75	76	79	80	82
No. VVs	1	1	5	1	1	3	2	1	1	2	4	6	39	7	1
BP AIC	-	27.5	3.62	2.07	4.48	1.2	2.07	3.62	-	-	1.29	5.76	1.29	4.88	1.29

Table 8. Distribution of verified virulence (VV) proteins in the hard partition of community 79

Community number	**1**	**5**	**7**	**9**	**14**	**15**	**16**	**17**
No. of VV proteins	9	3	2	2	5	12	5	1

We next consider the connection between multiple community membership and the VV genes. For $0.4 \le r \le 1.1$, we show the number of VV genes that are found by OverWeiMod to belong to more than one community in Table 9. The VV genes were not seen to be significantly overrepresented in either the multi-clustered genes or the mono-clustered genes for all values of *r*. However we do note that nearly half (49.3%) of the VV genes belong to more than one community for $r = 0.4$, and these may still play an important role in the system.

As we have shown previously, the number of multi-clustered nodes detected by OverWeiMod decreases as *r* increases. Nodes that remain multi-clustered for higher values of *r* may indicate cases that are more inclined to belong to multiple communities. We find 4 VV genes that are multi-clustered for *r* equal to 1 and 1.1. This may indicate that these proteins are more robustly multi-clustered than other pathogenicity-associated genes and therefore are more strongly inclined to belong to multiple communities. These four genes are: FGSG_04104 (probable guanine nucleotide-binding protein beta subunit), FGSG_08028 (conserved hypothetical protein), FGSG_10142 (related to transcription factor atf1+) and FGSG_03747 (NPS6 related to AM-toxin synthetase (AMT)). Probable guanine nucleotide-binding protein beta subunits are known to be part of a signalling process upstream of a range of biochemical pathways and therefore likely to be part of several communities. FGSG_03747 is a large protein with 7 predicted InterPro domains. It codes for a non-ribosomal peptide synthetase (NRPS). In this case the product and its function is already known. It is an extracellular siderophore that is used by Fusarium to bring the essential nutrient iron into the fungal hyphae and to protect against the cellular damage caused by various reactive oxygen species [37]. This link to a transfer process from outside the hyphae cell to inside may indicate likely bridging functional roles, therefore justifying why this gene may belong to more than one community. Such results suggest that OverWeiMod has the potential to detect appropriate multi-clustered nodes and therefore shows promise in predicting candidate multi-functional genes.

It should be noted that there is currently a small number of experimentally verified virulence genes, and within the set of known genes there may be a bias in reflecting particular classes of proteins that have been investigated experimentally, for example intracellular signalling and transcription-associated proteins. Therefore, network analyses and systems biology strategies such as the one presented here, offer good potential to plan future experiments using a more rational basis.

Table 9. The number of verified virulence (VV) seeds that belong to more than one community for $0.4 \le r \le 1.1$

r	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	**1.1**
No. of VV proteins	37	35	32	26	25	24	3	1

5 Discussion and Conclusions

In this study we have explored the disjoint and overlapping community structure of an integrated network for the globally important plant pathogenic fungus, *Fusarium graminearum*. We have investigated the topological and functional properties of proteins that belong to more than one community compared with those that belong to only one. It is shown that proteins in the intersection between two or more communities tend to be more highly connected than mono-clustered proteins. Furthermore, topological description through node role classification scheme illustrates that multi-clustered nodes tend to be enriched with connector roles (hub

and non-hub connectors, roles R6 and R3 respectively) that are structurally important in biological networks [20]. Additionally, we consider the functional properties of the multi-clustered proteins in terms of number of GO annotations. For all three aspects of the GO combined (ALL GO) and for BP and CC multi-clustered proteins tend to have a higher number of annotations, than proteins belonging to only one community, although the same trend is not seen for MF. These results corroborate to some extent the idea that multi-clustered proteins are bridges between communities, which allow the semi-independent functional units to interact and regulate all functions required by the system.

As mentioned previously, the problem of detecting overlapping communities is not as well defined as the standard community structure detection problem, for example due to difficulties in conceptualising a uniform definition of overlapping properties and consequently methodological disparities. For this reason, methods and parameters used vary greatly and comparisons across different methods and benchmark examples are challenging. Therefore, the purpose of comparison of the multi-clustered nodes found by OverWeiMod with those found by OCG is not to assess prediction accuracy, but to demonstrate that OverWeiMod is in line with other methods in this context. In addition, the following characteristics render OverWeiMod a competitive method. First, OverWeiMod has the capability to define the strength of belonging of a node to a community, giving another level of understanding of the system. Future work includes (i) analysing how the belonging coefficients of a multi-clustered gene are distributed between communities and (ii) identifying genes that are more equally spread among functional communities than others. The authors of [20] propose that nodes with the same role should have similar topological properties, therefore an expansion of the functional cartography analysis, where we include all 7 node role types may offer insight into important properties of the nodes.

Additionally, OverWeiMod is applicable to weighted networks making it conducive to further work in assessing the suitability of data sources in the integrated network. The integrated *Fusarium* network contains information from multiple heterogeneous data sources and some of the data sources may be of better quality than others. The inclusion of weighted edges in the network might provide a more accurate and informative description of the community structure of the organism. A simple approach might be to weight the edges heuristically depending on the number of data sources that suggest an association subject to the various thresholds chosen, this can be regarded as an indication of reliability of an interaction. Another approach would be to estimate the likelihood of a functional association between two proteins given evidence from each data source. This procedure requires a benchmark such as proteins known to be functionally associated as determined from experiment or suggested by some measure (such as belonging to the same pathway) and is complicated if the different data sources are not independent (see for example [2]). As we have shown before, community structure detection of a weighted network may result in a different partition as compared to the equivalent binary network [8]. Such effects can be addressed in future work.

Finally, the flexible nature of mathematical programming framework allows for the easy implementation of additional constraints and parameters, again leading to more accurate and detailed network representations. For example, we can use prior knowledge such that nodes with similar functional annotations could be constrained to

be in the same community. Furthermore, in terms of methodology, introducing symmetry constraints to as we have done in previous models [7] may improve the efficiency of OverWeiMod.

The motivation behind this study was to gain a better understanding of the fungal pathogen, *Fusarium graminearum*. We used network analysis tools to investigate the underlying mechanisms of the fungus from a community detection perspective. In particular we looked closely at the proteins taking part in more than one functional community in an attempt to identify those that may play a role in maintaining a structurally cohesive system. As the number of verified virulence proteins increases, analytical methods featured in this study could prove promising in the detection of relationships between the topological description and functional properties, potentially leading towards a better understanding of the pathogenicity process.

Acknowledgements. LB acknowledges financial support from the School of Natural and Mathematical Sciences, King's College London. ST acknowledges support from the EU and the Leverhulme Trust (RPG-2012-686).

References

1. Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M.: A probabilistic functional network of yeast genes. Science 306(5701), 1555–1558 (2004)
2. Lee, I., Marcotte, E.M.: Integrating functional genomics data. Methods in Molecular Biology 453, 267–278 (2008)
3. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A. 99, 7821–7826 (2002)
4. Blondel, V., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)
5. Ruan, J., Zhang, W.: Identifying network communities with a high resolution. Phys. Rev. E, 77, 016104 (2008)
6. Xu, G., Bennett, L., Papageorgiou, L.G., Tsoka, S.: Module detection in complex networks using integer optimisation. Algorithms for Molecular Biology 5, 36 (2010)
7. Xu, G., Tsoka, S., Papageorgiou, L.G.: Finding community structures in complex networks using mixed integer optimisation. Eur. Phys. J. B 60, 231–239 (2007)
8. Bennett, L., Liu, S., Papageorgiou, L.G., Tsoka, S.: Detection of disjoint and overlapping modules in weighted complex networks. Advances in Complex Systems, 15, 11500 (2012)
9. Kuhner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., Castaño-Diez, D., Chen, W.-H., Devos, D., Güell, M., Norambuena, T., Racke, I., Rybin, V., Schmidt, A., Yus, E., Aebersold, R., Herrmann, R., Böttcher, B., Frangakis, A.S., Russell, R.B., Serrano, L., Bork, P., Gavin, A.-C.: Proteome Organization in a Genome-Reduced Bacterium. Science 326(5957), 1235–1240 (2009)
10. Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868), 141–147 (2002)

11. Salathé, M., Jones, J.H.: Dynamics and Control of Diseases in Networks with Community Structure. PLoS Computational Biology 6(4), e1000736 (2010)
12. Zhang, S., Wang, R.-S., Zhang, X.-S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications 374(1), 483–490 (2007)
13. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping structure of complex networks in nature and society. Nature 435, 814–818 (2005)
14. Ma, X., Gao, L., Yong, X.: Eigenspaces of networks reveal the overlapping and hierarchical community structure more precisely. J. Stat. Mech., P08012 (2010)
15. Zhang, S., Wang, R.S., Zhang, X.S.: Uncovering fuzzy community structure in complex networks. Physical Review E 76(046103) (2007)
16. Zarei, M., Izadi, D., Samani, K.A.: Detecting overlapping community structure of networks based on vertex-vertex correlations. Journal of Statistical Mechanics (P11013) (2009)
17. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics 11(033015) (2009)
18. Wang, X., Jiao, L., Wu, J.: Adjusting from disjoint to overlapping community detection of complex networks. Physica A 388, 5045–5056 (2009)
19. Becker, E., Robisson, B., Chapple, C.E., Guénoche, A., Brun, C.: Multifunctional Proteins Revealed by Overlapping Clustering in Protein Interaction Network. Bioinformatics 28(1), 84–90 (2012)
20. Guimera, R., Amaral, L.A.N.: Functional Cartography of Complex Metabolic Networks. Nature 433, 895–900 (2005)
21. Liu, G., Wong, L., Chua, H.N.: Complex discovery from weighted PPI networks. Bioinformatics 25(15), 1891–1897 (2009)
22. Cuomo, C.A., Guldener, U., Xu, J.R., Trail, F., Turgeon, B.G., Di Pietro, A., Walton, J.D., Ma, L.J., Baker, S.E., Rep, M., Adam, G., Antoniw, J., Baldwin, T., Calvo, S., Chang, Y.L., Decaprio, D., Gale, L.R., Gnerre, S., Goswami, R.S., Hammond-Kosack, K., Harris, L.J., Hilburn, K., Kennell, J.C., Kroken, S., Magnuson, J.K., Mannhaupt, G., Mauceli, E., Mewes, H.W., Mitterbauer, R., Muehlbauer, G., Munsterkotter, M., Nelson, D., O'Donnell, K., Ouellet, T., Qi, W., Quesneville, H., Roncero, M.I., Seong, K.Y., Tetko, I.V., Urban, M., Waalwijk, C., Ward, T.J., Yao, J., Birren, B.W., Kistler, H.C.: The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. Science 317(5843), 1400–1402 (2007)
23. Wise, R.P., Caldo, R.A., Hong, L., Shen, L., Cannon, E., Dickerson, J.A.: BarleyBase/PLEXdb. Methods in Molecular Biology 406, 347–363 (2007)
24. Zhao, X.M., Zhang, X.W., Tang, W.H., Chen, L.: FPPI: Fusarium graminearum protein-protein interaction database. J. Proteome Res. 8(10), 4714–4721 (2009)
25. Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K., Ohta, H.: ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Research 35(suppl. 1), D863–D869 (2007)
26. Zhao, X.-M., Zhang, X.-W., Tang, W.-H., Chen, L.: FPPI: Fusarium graminearum Protein-Protein Interaction Database. Journal of Proteome Research 8(10), 4714–4721 (2009)
27. Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P., Philippi, S.: Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics 22(11), 1383–1390 (2006)

28. Lysenko, A., Defoin-Platel, M., Hassani-Pak, K., Taubert, J., Hodgman, C., Rawlings, C., Saqi, M.: Assessing the functional coherence of modules found in multiple-evidence networks from Arabidopsis. BMC Bioinformatics 12(1), 203
29. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113 (2004)
30. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. Nat. Genet. 25(1), 25–29 (2000)
31. Chen, J., Yuan, B.: Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics 22(18), 2283–2290 (2006)
32. R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2010)
33. http://mips.helmholtz-muenchen.de/genre/proj/FGDB/
34. Son, H., Seo, Y., Min, K., Park, A., Lee, J., Jin, J., Lin, Y., Cao, P., Hong, S., Kim, E., Lee, S., Cho, A., Lee, S., Kim, M., Kim, Y., Kim, J., Kim, J., Choi, G., Yun, S., Lim, J., Kim, M., Lee, Y., Choi, Y., Lee, Y.: A phenome-based functional analysis of transcription factors in the cereal head blight fungus, Fusarium graminearum. PLoS Pathogens 7, e1002310 (2011)
35. Winnenburg, R., Baldwin, T.K., Urban, M., Rawlings, C., Kohler, J., Hammond-Kosack, K.E.: PHI-base: a new database for pathogen host interactions. Nucleic Acids Res. 64(Database issue), D459–D464 (2006)
36. Winnenburg, R., Urban, M., Beacham, A., Baldwin, T.K., Holland, S., Lindeberg, M., Hansen, H., Rawlings, C., Hammond-Kosack, K.E., Kohler, J.: PHI-base update: additions to the pathogen host interaction database. Nucleic Acids Res. 6(Database issue), D572–D576 (2008)
37. Oidea, S., Moederb, W., Krasnoffc, S., Gibsonc, D., Haasd, H., Yoshiokab, K., Turgeona, B.G.: NPS6, Encoding a Nonribosomal Peptide Synthetase Involved in Siderophore-Mediated Iron Metabolism, Is a Conserved Virulence Determinant of Plant Pathogenic Ascomycetes. The Plant Cell 18, 2836–2853 (2006)