A network-based maximum link approach towards MS identifies potentially important roles for undetected ARRB1/2 and ACTB in liver cancer progression

Wilson Wen Bin Goh*

Department of Computing,
Imperial College London, UK
E-mail: wen.goh07@imperial.ac.uk
*Corresponding author

Yie Hou Lee

Singapore-MIT Alliance for Research and Technology,
Singapore
E-mail: yie.hou@smart.mit.edu

Zubaidah M. Ramdzan

Rosalind and Morris Goodman Cancer Centre,
McGill University, Canada
E-mail: zubaidah.mohamedramdzan@mcgill.ca

Maxey C.M. Chung

Department of Biological Sciences
and Department of Biochemistry,
National University of Singapore, Singapore
E-mail: maxey_chung@nuhs.edu.sg

Limsoon Wong

Department of Computer Science and Department of Pathology,
National University of Singapore, Singapore
E-mail: wongls@comp.nus.edu.sg

Marek J. Sergot

Department of Computing,
Imperial College London, UK
E-mail: m.sergot@imperial.ac.uk

Abstract: Hepatocellular Carcinoma (HCC) ranks among the deadliest of cancers and has a complex etiology. Proteomics analysis using iTRAQ provides a direct way to analyse perturbations in protein expression during HCC progression from early- to late-stage but suffers from consistency and coverage issues. Appropriate use of network-based analytical methods can help to overcome these issues. We built an integrated and comprehensive Protein-Protein Interaction Network (PPIN) by merging several major databases. Additionally, the network was filtered for GO coherent edges. Significantly differential genes (seeds) were selected from iTRAQ data and mapped onto this network. Undetected proteins linked to seeds (linked proteins) were identified and functionally characterised. The process of network cleaning provides a list of higher quality linked proteins, which are highly enriched for similar biological process Gene Ontology terms. Linked proteins are also enriched for known cancer genes and are linked to many well-established cancer processes such as apoptosis and immune response. We found that there is an increased propensity for known cancer genes to be found in highly linked proteins. Three highly-linked proteins were identified that may play an important role in driving HCC progression – the G-protein coupled receptor signaling proteins, ARRB1/2 and the structural protein beta-actin, ACTB. Interestingly, both ARRB proteins evaded detection in the iTRAQ screen. ACTB was not detected in the original dataset derived from Mascot but was found to be strongly supported when we re-ran analysis using another protein detection database (Paragon). Identification of linked proteins helps to partially overcome the coverage issue in shotgun proteomics analysis. The set of linked proteins are found to be enriched for cancer-specific processes, and more likely so if they are more highly linked. Additionally, a higher quality linked set is derived if network-cleaning is performed prior. This form of network-based analysis complements the cluster-based approach, and can provide a larger list of proteins on which to perform functional analysis, as well as for biomarker identification.

Keywords: biological networks; PPINs; MaxLink; liver cancer; HCC; hepatitis B; proteomics expansion pipeline.

Reference to this paper should be made as follows: Lee, Y.H., Ramdzan, Z.M., Chung, M.C.M., Wong, L. and Sergot, M.J. (xxxx) 'A network-based maximum link approach towards MS identifies potentially important roles for undetected ARRB1/2 and ACTB in liver cancer progression', *Int. J. Bioinformatics Research and Applications*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes:

Wilson Wen Bin Goh is currently pursuing his PhD in Imperial College London. His research interests include cancer proteomics and miRNA-networks.

Lee Yie Hou is currently a research scientist at Singapore-MIT Alliance for Research and Technology. His research interest spans the use of proteomics, lipidomics and their integration with clinical and experimental data to understand the biology of diseases including gynaecological disorders, tropical infectious diseases, drug-induced liver

injuries and cancer. He has strong interests in pursuing translation biomedical research in collaborative environments and aims to apply his work onto the unbiased discovery of candidate biomarkers and novel therapeutics.

Zubaidah M Ramdzan is currently a postdoctoral fellow in the Goodman Cancer Centre in McGill University, Montreal, Canada. Her research interests include the integration of proteomics and genomics to study regulation of transcription factors during tumorigenesis.

Maxey C.M. Chung, PhD is an Associate Professor in the Department of Biochemistry, Yong Loo Lin School of Medicine, and Department of Biological Sciences, Faculty of Science, National University of Singapore. His laboratory has been focusing on the applications of proteomics in biomarker discovery on gastrointestinal cancers, and also on the identification and elucidation of the proteins and the pathways involved in cancer metastasis. He is currently serving as Senior Editor for *Proteomics* and *Proteomics - Clinical Applications* and as a member of the Editorial Boards of *Journal of Proteome Research* and *Journal of Proteomics*.

Limsoon Wong is a provost's chair professor of computer science and a professor of pathology at the National University of Singapore (NUS). He is currently head of the computer science department at NUS. Before joining NUS, he was the Deputy Executive Director for Research at A*STAR's Institute for Infocomm Research. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He has/had served on the editorial boards of Information Systems, Journal of Bioinformatics and Computational Biology, Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Drug Discovery Today, and Journal of Biomedical Semantics.

Marek Sergot is Professor of Computational Logic in the Department of† Computing, Imperial College London, and Head of the Logic and Artificial† Intelligence section. He graduated in Mathematics at the University of† Cambridge† and then worked in mathematical modelling before joining the† Logic Programming Group at Imperial College in 1979. His research is in† logic for knowledge representation and reasoning, with particular† interests in temporal reasoning and the logic of action.

1 Introduction

Hepatocellular Carcinoma (HCC) ranks among the deadliest cancers (El-Serag, 2004). Its risk factors are varied – and include viral infection, germline mutations and alcohol-induction (Villanueva et al., 2010). Additionally, this cancer type can be histologically classified into poor, moderate and well-differentiated stages. Generally, the more poorly differentiated, the more advanced the cancer. However, histological characterisation is limited, and may have poor accuracy in properly staging the cancer patient. It also provides limited insight into the molecular interactions underlying the disease.

On one hand, high-throughput methods such as microarrays and RNA sequencing have been very useful in enhancing our molecular understanding of HCC. However, they only measure RNA levels, not protein level. Thus, the evidences provided are indirect. On the other hand, there are many difficulties associated with high-throughput protein analyses or proteomics. Recent improvements in mass spectrometry (MS)-based technologies, however, have greatly increased coverage and detection range. In particular, isobaric Tag Relative and Absolute Quantitation (iTRAQ)-based technologies have recently gained widespread popularity for their higher detection limits and ability to multiplex up to 8 samples simultaneously (Tan et al., 2008; Choe et al., 2007). Despite these improvements, proteomics still suffer from coverage and consistency issues. The coverage issue – that is, the ability to cover the entire proteome – arises in part due to the limited detection range of MS instruments, as well as due to inherent sample complexity. The consistency issue – that is, whether the same results are produced in repeated runs – arises in the seemingly random data acquisition and mechanics of current MS instruments, caused by dominance of the proteome by a few highly abundant proteins which leads to the oversampling of high abundance peptide ions (Liu et al., 2004).

These two problems make it difficult to analyse MS data in a comprehensive way. However, it is possible to partially overcome these issues by taking advantage of the fact that proteins tend to work in groups rather than as singular entities. In our previous work, we proposed a powerful complex prediction algorithm termed the Proteomics Expansion Pipeline (PEP) (Goh et al., 2011). PEP first identifies the group of high-confidence proteins or "seeds" from the proteomic screen – i.e., proteins that are consistently found in patients with significant over- or under-expression. These seeds are then mapped to nodes in a large integrated Protein-Protein Interaction Network (PPIN). An expanded subnetwork is then extracted from the PPIN by taking the immediate neighbours of the seeds in the PPIN. The subnetwork is then clustered using CFinder (Adamcsek et al., 2006). Each cluster is then ranked based on the average expression value of the proteins it contains. This includes the expression values of non-seeds as well. Proteins (in high-ranking clusters) not found in the proteomics screen are then screened against the original mass spectra for evidence of existence.

PEP uses a very comprehensive PPIN comprising data from HPRD (Keshava Prasad et al., 2009), BioGRID (Stark et al., 2006), IntAct (Aranda et al., 2010), and DIP (Xenarios et al., 2002), as well as data from literature (Stelzl et al., 2005; Rual et al., 2005). Although combining PPINs improves coverage of the protein interactome, it also compounds the noise present in them (von Mering et al., 2002). So, PEP uses the iterated Czekanowski-Dice distance (CD-distance) technique from CMC (Liu et al., 2009) to identify and eliminate potential noise edges from the integrated PPIN. The CD-distance technique is very effective – it produced a cleaned integrated PPIN having a significantly higher level of functional and localisation coherence, after eliminating about 50% of the edges from the original integrated PPIN.

We applied PEP to a group of 12 hepatocellular carcinoma (HCC) patients, of whom 5 were clinically diagnosed to be in the moderate (mod) and 7 in the poor stage. In our analysis, we found that most of the detected mod-stage proteins were also found in poor-stage patients. In terms of pathway enrichments, mod-stage

patients appeared to exhibit signs of immune response not observed in poor-stage patients, while poor-stage patients exhibited widespread metabolic deregulations. From the network-based PEP analysis, we uncovered several interesting clusters which might be crucial in driving mod-stage cancer to poor stage. Of these, the cluster comprising of PRKDC, WRN, XRCC5/6 and PCNA appeared most interesting.

The PEP approach is largely focused on cluster discovery and analysis, as well as recovery of low abundance and low confidence proteins. However, there are other network-based approaches which can be used on the cleaned PPIN. This may produce results that can augment our existing findings. More interestingly, it may reveal insights that have been missed. One useful approach may be Maxlink, introduced by Ostlund et al. (2010). It is a method for identifying novel cancer genes based on a given set of identified oncogenes. Maxlink first requires a set of oncogenes (seeds) to be identified based on literature search and the Cancer Gene Census (Futreal et al., 2004). It then produces a ranked list of new candidate genes based on the number of links they have in the FunCoup PPIN database (Alexeyenko and Sonnhammer, 2009) to the seed set. The higher the number of connections to seeds, and the lower the number of connections to non-seeds, the higher the rank. This approach relies on two reasonable hypotheses. The first hypothesis is that a protein should participate in the same biological processes, biological functions, or protein complexes that are over-represented among its interaction partners (Schwikowski et al., 2000; Hishigaki et al., 2001). The second hypothesis is that proteins in the same complex should have more interactions between themselves than with proteins outside the complex (Chen and Yuan, 2006).

Maxlink has not yet been explicitly tested on proteomics data. In this work, we apply a Maxlink-type approach on our HCC proteomics data.

2 Methods

2.1 Experimental setup

The experimental setup is described briefly here; details are given in supplementary methods. Liver tissues were obtained from 12 male patients diagnosed with HCC and suffered from cirrhosis with chronic Hepatitis B Virus (HBV) infection. There was no metastasis at the point of surgery. Tissues collected were grouped according to histology report; 5 had moderately differentiated HCC (mod) and 7 had poorly differentiated HCC (poor). Paired tissues were obtained from each patient, one from the adjacent non-tumour region (normal) and the other from the tumour region of the resected liver. Mixed protein lysate from each patient was put through an initial phase of iTRAQ followed by 2D liquid chromatography. Finally, the resultant spectrum was resolved by peptide database search using Mascot.

2.2 Selection of seed proteins

A seed is defined as meeting the following requirements: Support by at least 4 poor patients, and with a combined differential score ≥ 1.2. The combined differential score is calculated as the average score of the protein ratios (tumour over self non-tumour). If the ratio is below 1 (under-expressed), the reciprocal is used.

2.3 Network integration and cleaning

An integrated PPIN was built comprising of data from HPRD (Keshava Prasad et al., 2009), BioGRID (Stark et al., 2006), IntAct (Aranda et al., 2010), and DIP (Xenarios et al., 2002), as well as data from literature (Stelzl et al., 2005; Rual et al., 2005). The various IDs were mapped using BioMart to gene names. This network was then filtered using the iterated CD-distance method from CMC (Liu et al., 2009), and the top 90% of the highest non-zero scoring edges are kept. The resultant combined network displayed the properties of a typical PPIN such as a power-law distribution of the degrees, disassortativity (hubs less likely to be linked to each other) and small-world (small diameter) (data not shown).

2.4 Identification of linked proteins

The code was written in PERL. Let the network G be comprised of nodes V and edges E. From the set of seeds $X \subseteq V$, the set of non-seeds Y is derived ($Y = V - X$). The set of linked proteins L are those proteins in Y that have at least 1 connection to proteins in X. That is, $L = \{y \in Y \mid 1 \leq |\{x \in X \mid (x, y) \in E\}|\}$.

2.5 Gene-Ontology (GO)-based characterisation and coherence measurement

Annotation and the GO tree (ver 1.2 OBO) files for Homo Sapiens were downloaded from geneontology.org (dated 23 April 2011). UniProtKB accessions were mapped to Ensembl Gene IDs and gene names via Biomart. Informative biological process terms were extracted from the GO OBO file; as in Zhou et al. (2002), a term is considered informative if it is annotated to at least 30 genes and no direct descendent of the term is annotated to at least 30 genes. Significance testing for each cluster was performed using the hypergeometric test with Bonferroni correction ($p \leq 0.05$).

To evaluate the quality of linked proteins derived from the cleaned and uncleaned integrated network, we measured Gene Ontology Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) term coherence for every edge – i.e., a seed protein connected to a linked protein – derived from the cleaned and uncleaned network. Edge coherence is calculated by counting the number of shared GO terms in each category for every GO-annotated edge divided by the total number of considered edges.

3 Results

3.1 Identification of linked proteins and the important effects of network cleaning

235 seeds were returned from the dataset. From the cleaned dataset, 288 linked proteins were found to share at least one other connection with a seed. From the uncleaned dataset, 902 linked proteins were returned.

We then built two sets of derived networks from cleaned and uncleaned networks, obtaining edges formed between seed and linked proteins (seed + linked)

from the reference integrated PPIN, and checked the extent of GO term sharing. It is observed that the cleaned network boasts much higher quality edges where the joined nodes tend to have deep sharing of GO terms. Hence, the linked proteins derived from the cleaned network is likely to be more biologically relevant. The improvement in quality as a result of the cleaning step is observed to be at least two folds; see Table 1.

Table 1 GO term coherence of linked proteins derived from cleaned and uncleaned networks

Network	*BP*	*MF*	*CC*
Cleaned	0.180	0.376	0.755
Uncleaned	0.035	0.121	0.300

It can be observed that the cleaned network boasts higher quality edges where the joined nodes tend to have deep sharing of GO terms. Hence, the linked proteins derived from the cleaned network is likely to be more biologically relevant (BP: biological process, MF: molecular function and CC: cellular localisation).

To see whether the improvement in the 3 GO categories (biological process – BP, Molecular Function – MF and Cellular Localisation – CC) is greater than the network generally, we calculated the log odds ratio. That is, seed+linked from cleaned network/ seed+linked from uncleaned network divided over total cleaned/total uncleaned network. Interestingly, there was a 1.5X enhancement for BP terms whereas there were no improvements for MF (1.02X) and CC terms (1.02X). This indicates that the cleaned seed+linked protein network is highly enriched for proteins in shared biological processes. The significant enhancement of GO term coherence in the cleaned network indicates that the cleaning step is important. It also improves analytical results in combination with the Maxlink approach.

There appears to be a strong linear correlation between the ranks of the linked proteins (sorted in descending order by the number of connections to seeds) in the cleaned and uncleaned networks; Figure 1. This is evident from the string of points forming a near perfect diagonal and is not particularly surprising. However, it can also be seen that a large number of points are ranked below the diagonal. This means that they are ranked relatively higher than they would actually be after the cleaning step. It also means that the cleaned linked proteins is enriched for linked proteins with high ranks from the poor network. Although this is less direct evidence than measuring GO coherence as above, it does demonstrate the efficacy and relevance of the cleaning procedure.

We then turned to informative GO term enrichment for the linked proteins in the cleaned network. This produced 87 significant GO BP terms (for 288 proteins). To find out whether these terms are closely associated in the GO tree, we calculated the shortest-path lengths between all terms, and returned the average path length. A null distribution is then generated by picking a number of proteins equal to the linked proteins from the reference network, calculating the significant informative GO terms, and calculating the average GO term path length. This is repeated 1000 times. For linked proteins, we find that the GO terms are significantly more closely associated (Z-score $= -3.98$, $p = 0.000034$).

Table 2 List of most highly connected proteins to the seed set (sorted in descending order)

Protein	*Links*
ARRB1	13
ACTB	12
ARRB2	12
TRAF6	9
PPP2R2B	8
MCC	7
TBK1	6
TNFRSF1B	6
TP53	6
CFTR	5
IKBKG	5
NFKB2	5
RIPK1	5
FN1	4
PTMA	4
REL	4
SMAD2	4
SMAD3	4
VHL	4
ARF6	3
CASP3	3
CBL	3
CTSB	3
DYNLL1	3
EIF1B	3
EIF6	3
LARP1	3
MAP3K7IP1	3
PLG	3
PRKCD	3
RAF1	3
RELB	3
S100A1	3
SUMO4	3

3.2 Properties of the most highly linked proteins: ACTB and the ARRB1/2

In the cleaned dataset, 34 proteins share at least 3 connections to the seeds; Supplementary Table 1. This set includes known oncogenes such as NFKB2, RAF1, REL, TP53 and VHL.

Of these 34 highly linked proteins, ARRB1/2 and ACTB were found to be most connected to the seeds. Interestingly, these 3 proteins were not found in the set of detected non-seed proteins either. Hence, it is possible that these proteins were not picked up by MS.

To verify this, we turned to another MS-protein identification algorithm, Paragon. In Goh et al. (2011), we found that there was good correlation between

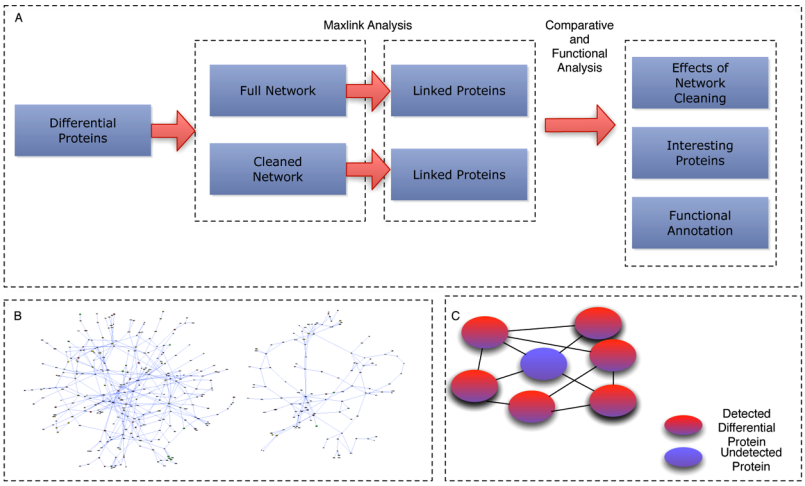

Figure 1 (A) Overview of analytical pipeline. Two sets of networks are used; a cleaned and uncleaned network, to discover linked proteins undetected by the MS screen. The results are then compared using GO coherence measurements and ranks correlation analysis. The set of interesting linked proteins are then functionally annotated using GO terms. (B) A cleaned (left) and uncleaned network (right). Note that this is for illustration purposes. The actual networks are too complex to visualise. (C) An example of a Maxlinked protein (blue). A Maxlinked protein is one that is highly connected to detected differentially expressed proteins (see online version for colours)

the reported ranks and ratios of Mascot and Paragon. However, Paragon reported many more proteins than Mascot even although these extra proteins were found to originate from lower quality MS/MS spectra. We generated a Paragon excess list comprising the read outs from all 12 patients not found in Mascot. Here, ACTB was found to be supported in all 12 patients. It was also found to be very confidently predicted in Paragon with a normalised average rank of 0.028 (out of 1). The omission of ACTB in the set of detected proteins could be due to a variety of factors – e.g., Mascot's filtering parameters, incomplete coverage in its database or differences in peptide matching algorithms (Shilov et al., 2007). ARBB1 and 2 however, were not found to be predicted in Paragon.

The inter-connections of linked proteins to the seeds are shown in Figure 2. This network appears to be quite sparse, and is probably not suitable for performing cluster analysis. ARRB1 and ARRB2 share many seeds (Figure 2 inset). This includes HSPA8, HNRNPM, HSPA5, TUBA1C, HNRNPH1, FLNA, CLTC, S100A9, NCL and ANXA2. This set of proteins appear to be important in vesicle-mediated transport ($p = 0.0016$), as well as actin cytoskeleton reorganisation ($p = 0.00885$).

The subnet formed by ARRB1/2 and ACTB (Figure 2 inset) shows that ACTB is less strongly connected to ARRB1/2. Here, proteins linked only to ARRB1 and 2 are colored yellow; ARRB1 and ACTB in light blue; all 3 in purple; proteins not shared are in pink. GO term analysis of the 20 connected seeds does not reveal any term typically associated with cancers. Instead, many of the terms are more

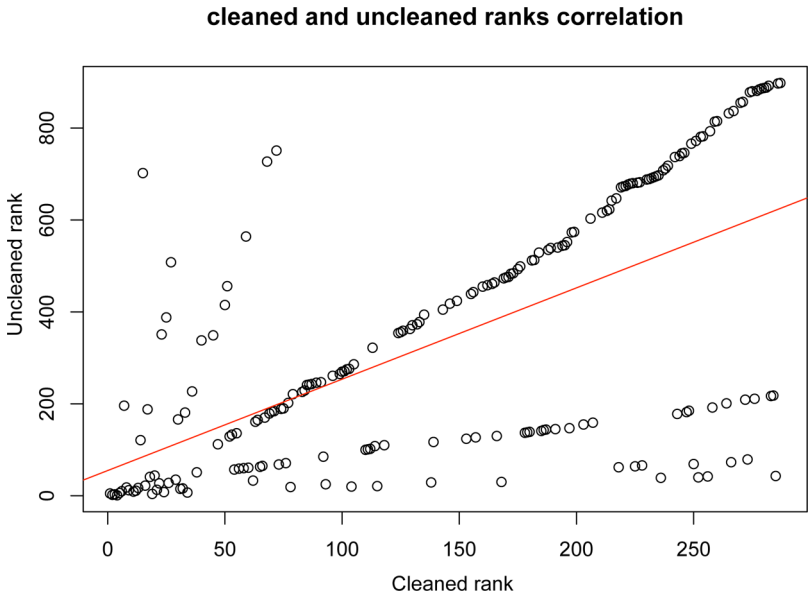

Figure 2 Ranks correlations between uncleaned and cleaned networks. There is a strong linear ranks correlation between the linked proteins found in the cleaned and uncleaned networks. The trendline is approximated with a gradient of 2 and y-intercept of 55 (adjusted R-squared = 0.39, $p \leq 2.2e-16$) (see online version for colours)

akin to functionalities associated with the liver, e.g., vesicle-mediated transport and transport. However, stress responses and wound healing is represented by half of the proteins, and it does agree with our previous observations where many of the significant clusters were also enriched for stress responses.

ARRB1/2 are signaling proteins of G protein-coupled receptors (GPCRs). They are known to play an important role in tumour tissue invasion and metastasis. Rosano et al. (2009) showed that in ovarian cancer, silencing of both ARRB1 and 2 inhibited endothelin-A (ET(A)R) receptor-driven silencing, resulting in SRC suppression, mitogen-activated protein kinase (MAPK), AKT activation, EGFR transactivation and, most importantly, complete inhibition of ET-1-induced beta-catenin/TCF transcriptional activity and cell invasion. In colorectal cancer, it was reported that the association of ARRB1 with SRC is critical for carcinoma cell migration as well as metastatic spread of cancer to liver in vivo (Buchanan et al., 2006). This association is stimulated by the expression of prostagladin E, and may act by activation of the EGFR controlled pathways. Like Rosano et al. (2009), this study also implied a functional role for ARRB1 as an important mediator of tumour invasion and metastasis. Interestingly, to our best knowledge, ARRB1/2 have not been reported as a crucial factor in driving oncogenic progression in HCC from mod to poor. However, the fact that it is linked to the most number of our MS-detected dysregulated proteins, coupled to its enrichment in other metastatic tumours suggests a potentially important role in driving HCC progression.

Actins are highly conserved proteins that are involved in cell motility, structure, and integrity. ACTB or beta-action is a major constituent of the

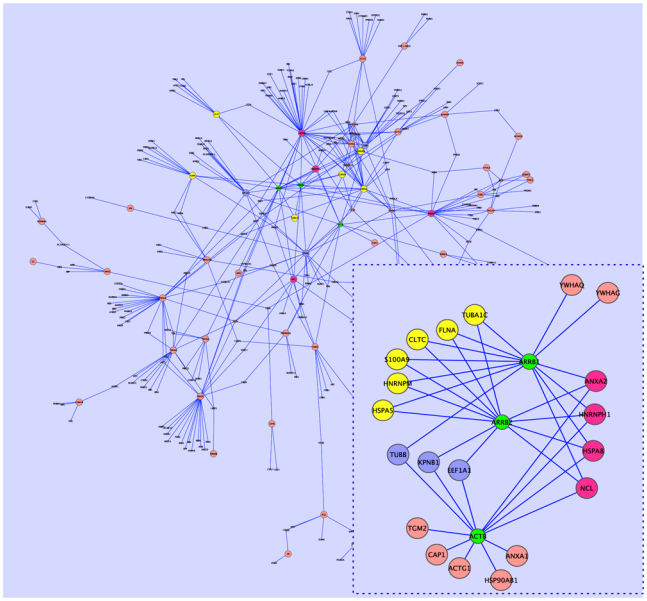

Figure 3 Background: the inter-connections of linked proteins to the seeds; Inset: the connections between ARBB1,2 and ACTB. Background: The network comprising seeds and linked proteins is topologically sparse. Inset: Here, proteins linked only to ARRB1 and 2 are labelled yellow; ARRB1 and ACTB in light blue; connections to all 3 in purple; proteins not shared are in pink (see online version for colours)

contractile apparatus and one of two nonmuscle cytoskeletal actins. Because it is a housekeeping protein, it is commonly used for normalisation in gene expression studies. However, here, we found that ACTB is connected to a disproportionate number of dysregulated proteins in the cleaned network (as well as in uncleaned), and could possibly be involved in driving HCC progression. Indeed, several studies have shown that ACTB is differentially expressed in cancer. This includes differential expression of ACTB in N1S1 rat hepatoma (Chang et al., 1998), colon carcinoma/colorectal cancer (CRC) (Sagynaliev et al., 2005), and blood cancers such as Chronic Myelogenous Leukemia (CML), Chronic Lymphocytic Leukemia (CLL), Acute Myelogenous Leukemia (AML) (Lupberger et al., 2002). In human colon adenocarcinoma (Nowak et al., 2005), hepatoma (Popow et al., 2006) and melanoma (Goidin et al., 2001), there is a tendency for ACTB to be dysregulated in cells with greater metastatic capacity.

Of the 34 most highly linked proteins, there is an enrichment for significant GO terms commonly associated with cancer. For example, apoptosis and programmed cell death ($n = 16$, $p = 5.19e - 09$), activation of immune responses ($n = 7$, $p = 6.15e - 08$), response to stress ($n = 18, p = 1.40e - 06$), positive regulation of NF-kappaB transcription factor activity ($n = 5, p = 4.30e - 05$), and negative regulation of cell proliferation ($n = 8, p = 7.33e - 05$). Comparing the highly linked proteins (at least 3 connections to seeds) and lower linked proteins against the Cancer Gene Census (Futreal et al., 2004), we found 1.86X more enrichment for known cancer genes among the highly linked proteins. This reinforces the notion

that increased connectivity to seed proteins is likely to imply potential oncogenic function.

4 Discussions

4.1 How the Maxlink approach complements PEP

Both Maxlink (Ostlund et al., 2010) and PEP (Goh et al., 2011) addresses to an extent incomplete coverage issues in proteomics. However, they do this differently. Maxlink identifies additional proteins based on the number of connections to seed proteins whereas PEP identifies significantly differentially expressed submodules formed by the neighbours of the seeds. Functional analysis reveals a common enrichment of terms such as apoptosis and stress responses. However, Maxlink picked up immune responses which we did not observe in PEP. One likely possibility is that these proteins are poorly connected in the reference PPIN and therefore did not qualify as clusters. Since Maxlink only considers proteins connected to seeds regardless of their inter-connectivity, it provides an additional dimension to the results from cluster analysis. The Maxlink approach is also dependent on the quality of the reference network. We show here that the process of network cleaning greatly reduces the number of linked proteins from 902 to 288 but the latter set is enriched for coherent terms as well as highly ranked linked proteins in the former.

However, both methods are not able to deal adequately with the consistency issue in proteomics – viz., the unrepeatability of results from the same samples. Furthermore, the dependence on identifying seeds from the iTRAQ screen filters off a large amount of the limited available information, because only proteins supported by the majority of samples and clearly differentially expressed are primarily considered.

Hence, there is still avenue for further development of methods that can deal with these shortfalls.

4.2 The role of ARRB1/2 proteins and ACTB in driving HCC progression

It is interesting that the most linked proteins in HCC to seeds turned out to be non-classical oncogenes. This reinforces the notion of how complex cancer is, and how limited current knowledge is. There is limited literature evidence on the roles of ARRB1/2 and even beta-actin in driving metastasis. Although many of the reported literature documents other cancer types, especially more aggressive cancers, it is possible that dysregulation of these proteins can also have similar effects in driving HCC progression.

Furthermore, although found in more aggressive tumours, GO term analysis of the shared neighbours by these 3 proteins revealed no significant cancer associated term, aside from wound healing and stress response. This could be also be due to the limited annotations in GO, as well as due to the limited scope in analysing only PPIN information. ARRB1/2 proteins are GPCR signaling proteins and may drive invasiveness and metastasis via several different pathways ranging from ET(A)R, SRC and EGFR, AKT and MAPK (Rosano et al., 2009; Buchanan et al., 2006).

The role of beta-actin is more interesting given that it is a well-known housekeeping protein with widespread expression. It is typically used, alongside GADPH, as a marker for normalisation of gene expression experiments. Its functional role in cancer is not particularly well-characterised despite literature evidence indicating its dysregulation in more aggressive cancer types (Ruan and Lai, 2007; Popow et al., 2006; Nowak et al., 2005).

In our derived network, we noted that ARRB1/2's shared neighbours were enriched for the GO term, actin cytoskeleton reorganisation ($p = 0.00885$). This is effected by FLNA and S100A9, which are shared between ARRB1/2. FLNA or filamin-A, an actin-binding protein, that is widely expressed and regulates re-organisation of the actin cytoskeleton by interacting with integrins, transmembrane receptor complexes and second messengers. S100A9 is a calcium binding protein, and may be implicated in leukemia (Cheok et al., 2003). Furthermore, we found significant crosstalk with actin via shared neighbours. Indeed, by looking at the shared neighbours they seemed to converge on mRNA fate. NCL or nucleolin forms part of mRNP complex which decides on mRNA localisation, translation and turnover (Moore et al., 2005). EEF1A1, the translation elongation factor on the other hand is required for binding of aminoacyl-tRNA to the ribosome during translation. The perturbed mRNA dynamics could be the result of

- interaction of host cell with HBV and/or
- dysregulated protein synthesis in malignant neoplastic transformation to poorly-differentiated HCC to promote tumour growth.

In the process of dedifferentiation from small, well differentiated to moderately differentiated and finally poorly differentiated HCC tumours, the vasculature remodels substantially and abnormally (Sonoda et al., 1989). This vasculogenic and angiogenic switch is critical for tumour growth. Endothelial cell motility drives the formation and maintenance of blood vessels and to do so, actin dynamics is required. Of note, HBV upregulates and stabilises HIF-α, and subsequently stimulate the cascade of signalling events that lead to angiogenesis (Moon et al., 2004). Similarly, HBVx activates RhoA, a small GTPase that regulates actin (Fukui et al., 2006). Together, our results support active angiogenesis and vasculogenesis as important molecular events that occur in the progression of HBV-induced HCC which requires the participation of actin. At the same time, tubulins A and B and CAP1 are connected to ARRB and actin seeds, suggesting the importance of the regulation of cytoskeletal events in multiple cellular responses during oncogenesis.

5 Conclusion

Identification of linked proteins helps to partially overcome the coverage issue in proteomics analysis. The set of linked proteins are found to be enriched for cancer-specific processes, and more likely so if they are highly linked. Additionally, a higher quality linked set is derived if network-cleaning is performed prior. Here, the most linked proteins (ARBB1/2 and ACTB) turned out to be non-classical cancer genes which have been evidenced to play important roles in metastasis and invasiveness although the mechanisms appear to be very complex. To the best of

our knowledge, there is not much known about the role these proteins play in HCC progression.

The Maxlink form of network-based analysis complements cluster-based approaches such as PEP, because it concentrates on seed connections rather than inter-connectivity between seeds and their neighbours. It can therefore add on to the list of proteins on which to perform functional analysis, as well as for biomarker identification. In addition, we find that cleaning the network prior to performing Maxlink provides a higher quality set of linked proteins.

Acknowledgements

WWBG is supported by a Wellcome Trust Scholarship (83701/Z/07/Z). LW is supported in part by a Singapore National Research Foundation grant NRF-G-CRP-2007-04-082(d). The iTRAQ work is supported by the Singapore Cancer Syndicate.

References

Adamcsek, B., Palla, G., Farkas, I.J., Derényi, I. and Vicsek, T. (2006) 'Cfinder: locating cliques and overlapping modules in biological networks', *Bioinformatics*, Vol. 22, No. 8, April, pp.1021–1023.

Alexeyenko, A. and Sonnhammer, E.L.L. (2009) 'Global networks of functional coupling in eukaryotes from comprehensive data integration', *Genome Res.*, Vol. 19, No. 6, June, pp.1107–1116.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S.N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K. and Hermjakob, H. (2010) 'The intact molecular interaction database in 2010', *Nucleic Acids Res.*, Vol. 38, Database issue, January, D525–D531.

Buchanan, F.G., Gorden, D.L., Matta, P., Shi, Q., Matrisian, L.M. and DuBois, R.N. (2006) 'Role of beta-arrestin 1 in the metastatic progression of colorectal cancer', *Proc. Natl. Acad. Sci. USA*, Vol. 103, No. 5, January, pp.1492–1497.

Chang, T.J., Juan, C.C., Yin, P.H., Chi, C.W. and Tsay, H.J. (1998) 'Up-regulation of betaactin, cyclophilin and gapdh in n1s1 rat hepatoma', *Oncol. Rep.*, Vol. 5, No. 2, pp.469–471.

Chen, J. and Yuan, B. (2006) 'Detecting functional modules in the yeast proteinprotein interaction network', *Bioinformatics*, Vol. 22, No. 18, September, pp.2283–2290.

Cheok, M.H., Yang, W., Pui, C-H., Downing, J.R., Cheng, C., Naeve, C.W., Relling, M.V. and Evans, W.E. (2003) 'Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells', *Nat. Genet*, Vol. 34, No. 1, May, 85–90.

Choe, L., D'Ascenzo, M., Relkin, N.R., Pappin, D., Ross, P., Williamson, B., Guertin, S., Pribil, P. and Lee, K.H. (2007) '8-plex quantitation of changes in cerebrospinal uid protein expression in subjects undergoing intravenous immunoglobulin treatment for alzheimer's disease', *Proteomics*, Vol. 7, No. 20, October, pp.3651–3660.

El-Serag, H.B. (2004) 'Hepatocellular carcinoma: recent trends in the united states', *Gastroenterology*, Vol. 127, 5 Suppl 1, November, pp.S27–S34.

Fukui, K., Tamura, S., Wada, A., Kamada, Y., Sawai, Y., Imanaka, K., Kudara, T., Shimomura, I. and Hayashi, N. (2006) 'Expression and prognostic role of rhoa gtpases in hepatocellular carcinoma', *J. Cancer Res. Clin. Oncol.*, Vol. 132, No. 10, October, pp.627–633.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) 'A. census of human cancer genes', *Nat. Rev. Cancer*, Vol. 4, No. 3, March, pp.177–183.

Goh, W.W.B., Lee, Y.H., Ramdzan, Z.M., Chung, M.C.M., Wong, L. and Sergot, M.J. (xxxx) 'A networkbased maximum link approach towards MS identifies potentially important roles for undetected ARRB1/2 and ACTB in liver cancer progression', *Int. J. Bioinformatics Research and Applications*, Vol. x, No. x, pp.xxx–xxx.

Goidin, D., Mamessier, A., Staquet, M.J., Schmitt, D. and Berthier-Vergnes, O. (2001) 'Ribosomal 18s rna prevails over glyceraldehyde-3-phosphate dehydrogenase and beta-actin genes as internal standard for quantitative comparison of mrna levels in invasive and noninvasive human melanoma cell subpopulations', *Anal. Biochem*, Vol. 295, No. 1, August, pp.17–21.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. (2001) 'Assessment of prediction accuracy of protein function from protein-protein interaction data', *Yeast*, Vol. 18, No. 6, April, 523–531.

Liu, G., Wong, L. and Chua, H.N. (2009) 'Complex discovery from weighted ppi networks', *Bioinformatics*, Vol. 25, No. 15, August, pp.1891–1897.

Liu, H., Sadygov, R.G. and Yates, J.R. 3rd. (2004) 'A model for random sampling and estimation of relative protein abundance in shotgun proteomics', *Anal. Chem*, Vol. 76, No. 14, July, pp.4193–4201.

Lupberger, J., Kreuzer, K.A., Baskaynak, G., Peters, U.R., le Coutre, P. and Schmidt, C.A. (2002) 'Quantitative analysis of beta-actin, beta-2-microglobulin and porphobilinogen deaminase mrna and their comparison as control transcripts for rt-pcr', *Mol. Cell Probes*, Vol. 16, No. 1, February, pp.25–30.

Moon, E-J., Jeong, C-H., Jeong, J-W., Kim, K.R., Yu, D-Y., Murakami, S., Kim, C.W. and Kim, K-W. (2004) 'Hepatitis b virus x protein induces angiogenesis by stabilizing hypoxia-inducible factor-1alpha', *FASEB J.*, Vol. 18, No. 2, February, pp.382–384.

Moore, M.J. (2005) 'From birth to death: the complex lives of eukaryotic mrnas', *Science*, Vol. 309, No. 5740, September, pp.1514–1518.

Nowak, D., Skwarek-Maruszewska, A., Zemanek-Zboch, M. and Laszkiewicz, M.M-B. (2005) 'Beta-actin in human colon adenocarcinoma cell lines with different metastatic potential', *Acta Biochim Pol.*, Vol. 52, No. 2, pp.461–468.

Ostlund, G., Lindskog, M. and Sonnhammer, E.L.L. (2010) 'Network-based identification of novel cancer genes', *Mol. Cell Proteomics*, Vol. 9, No. 4, April, pp.648–655.

Popow, A., Nowak, D. and Laszkiewicz, M.M-B. (2006) 'Actin cytoskeleton and betaactin expression in correlation with higher invasiveness of selected hepatoma morris 5123 cells', *J. Physiol Pharmacol*, Vol. 57, Suppl. 7, November, pp.111–123.

Keshava Prasad, T.S., Kandasamy, K. and Pandey, A. (2009) 'Human protein reference database and human proteinpedia as discovery tools for systems biology', *Methods Mol. Biol.*, Vol. 577, pp.67–79.

Rosanò, L., Cianfrocca, R., Masi, S., Spinella, F., Di Castro, V., Biroccio, A., Salvati, E., Nicotra, M.R., Natali, P.G. and Bagnato, A. (2009) 'Beta-arrestin links endothelin a receptor to beta-catenin signaling to induce ovarian cancer cell invasion and metastasis', *Proc. Natl. Acad. Sci. USA*, Vol. 106, No. 8, February, pp.2806–2811.

Rual, J-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P. and Vidal, M. (2005) 'Towards a proteome-scale map of the human protein-protein interaction network', *Nature*, Vol. 437, No. 7062, October, pp.1173–1178.

Ruan, W. and Lai, M. (2007) 'Actin, a reliable marker of internal control?', *Clin. Chim. Acta.*, Vol. 385, Nos. 1–2, October, pp.1–5.

Sagynaliev, E., Steinert, R., Nestler, G., Lippert, H., Knoch, M. and Reymond, M-A. (2005) 'Web-based data warehouse on gene expression in human colorectal cancer', *Proteomics*, Vol. 5, No. 12, August, pp.3066–3078.

Schwikowski, B., Uetz, P. and Fields, S. (2000) 'A network of protein-protein interactions in yeast', *Nat. Biotechnol.*, Vol. 18, No. 12, December, pp.1257–1261.

Shilov, I.V., Seymour, S.L., Patel, A.A., Loboda, A., Tang, W.H.., Keating, S.P., Hunter, C.L., Nuwaysir, L.M. and Schaeffer, D.A. (2007) 'The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra', *Mol. Cell Proteomics*, Vol. 6, No. 9, September, pp.1638–1655.

Sonoda, T., Shirabe, K., Takenaka, K., Kanematsu, T., Yasumori, K. and Sugimachi, K. (1989) 'Angiographically undetected small hepatocellular carcinoma: clinicopathological characteristics, follow-up and treatment', *Hepatology*, Vol. 10, No. 6, December, pp.1003–1007.

Stark, C., Breitkreutz, B-J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) 'Biogrid: a general repository for interaction datasets', *Nucleic Acids Res.*, Vol. 34, Database issue, January, D535–D539.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. and Wanker, E.E. (2005) 'A human proteinprotein interaction network: a resource for annotating the proteome', *Cell*, Vol. 122, No. 6, September, pp.957–968.

Tan, H.T., Tan, S., Lin, Q., Lim, T.K., Hew, C.L. and Chung, M.C.M. (2008) 'Quantitative and temporal proteome analysis of butyratetreated colorectal cancer cells', *Mol. Cell Proteomics*, Vol. 7, No. 6, June, pp.1174–1185.

Villanueva, A., Minguez, B., Forner, A., Reig, M. and Llovet, J.M. (2010) 'Hepatocellular carcinoma: novel molecular approaches for diagnosis, prognosis, and therapy', *Annu Rev. Med.*, Vol. 61, pp.317–328.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) 'Comparative assessment of large-scale data sets of protein-protein interactions', *Nature*, Vol. 417, No. 6887, May, pp.399–403.

Xenarios, I., Salwński, L., Duan, X.J., Higney, P., Kim, S-M. and Eisenberg, D. (2002) 'Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions', *Nucleic Acids Res.*, Vol. 30, No. 1, January, pp.303–305.

Zhou, X., Kao, M-C.J. and Wong, W.H. (2002) 'Transitive functional annotation by shortest-path analysis of gene expression data', *Proc. Natl. Acad. Sci. USA*, Vol. 99, No. 20, October, pp.12783–12788.