BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 1 2009, pages 98–104
doi:10.1093/bioinformatics/btn593

Systems biology

Align human interactome with phenome to identify causative genes and networks underlying disease families

Xuebing Wu, Qifang Liu and Rui Jiang*

MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation,
Tsinghua University, Beijing 100084, China

Received on July 16, 2008; revised on September 24, 2008; accepted on November 11, 2008

Advance Access publication November 13, 2008

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Understanding the complexity in gene–phenotype relationship is vital for revealing the genetic basis of common diseases. Recent studies on the basis of human interactome and phenome not only uncovers prevalent phenotypic overlap and genetic overlap between diseases, but also reveals a modular organization of the genetic landscape of human diseases, providing new opportunities to reduce the complexity in dissecting the gene–phenotype association.

Results: We provide systematic and quantitative evidence that phenotypic overlap implies genetic overlap. With these results, we perform the first heterogeneous alignment of human interactome and phenome via a network alignment technique and identify 39 disease families with corresponding causative gene networks. Finally, we propose AlignPI, an alignment-based framework to predict disease genes, and identify plausible candidates for 70 diseases. Our method scales well to the whole genome, as demonstrated by prioritizing 6154 genes across 37 chromosome regions for Crohn's disease (CD). Results are consistent with a recent meta-analysis of genome-wide association studies for CD.

Availability: Bi-modules and disease gene predictions are freely available at the URL http://bioinfo.au.tsinghua.edu.cn/alignpi/

Contact: ruijiang@tsinghua.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Recently, several large-scale studies have systematically evaluated the complex relationship between human genetic diseases and genes, revealing prevalent phenotypic overlap (van Driel *et al*., 2006) and genetic overlap (Rzhetsky *et al*., 2007) between human diseases. Our previous effort in the genome-wide inference of disease genes for 5080 human diseases reveals a modular organization of the genetic landscape of human diseases (Wu *et al*., 2008). These endeavors further spur the transition from the Mendelian 'one gene – one phenotype' rule to a 'muti-gene – multi-phenotype' paradigm. It is now well recognized that phenotypes are the outward manifestation of network effects among products of multiple genes. For example, a macrophage-enriched network has been shown to be responsible for a group of metabolic traits (Chen *et al*., 2008). As genes and diseases are highly intra- and inter-connected, the new paradigm requires new network-based framework to reduce the complexity and to facilitate the discovery of novel disease genes (Pujana *et al*., 2007).

We have shown that a simple linear regression model efficiently captures the underlying architecture of the human interactome and phenome networks (Wu *et al*., 2008). The human disease phenome is depicted by a network of disease phenotypes, with edges weighted by phenotypic overlap scores. Similarly, the interactome is a network of genes linked by physical interactions between their protein products. The two networks are further linked by gene–phenotype associations. We have shown that the proximity between disease genes in the gene network could explain the phenotypic overlap of diseases, and the success of this model suggests a global concordance of the topology between the phenotype network and the gene network. It remains interesting to see whether a direct comparison of the network topology can identify consistent or 'conserved' parts between the human interactome and phenome networks. For example, it may be possible that we could find a group of phenotypically overlapped diseases (a disease module), with a corresponding group of causative genes (a gene module). In such a scenario, the causative gene network may suggest a common pathway for the disease family and explain the overlap between the diseases. In addition, the alignment could also provide an effective way to peel modular sub-structures (or *bi-module* here) from the modular genetic landscape of human diseases, hence greatly reducing the complexity for further analysis.

As a proof-of-concept, we compare human interactome and phenome networks with the network alignment technique, which is originally proposed for comparing protein networks (Sharan and Ideker, 2006). Typically, network alignment works on networks from two species and seeks to identify pairs of sub-networks, one from each species, with sequence similarity between nodes (proteins) from different species. The identified pairs of sub-networks are thought to be conserved protein complexes or pathways. The alignment takes three inputs: two protein networks from different species and some inter-network links (similarity in sequence). We call this a homogenous alignment, because the aligned networks are of the same type (protein–protein interaction network). However, technically, network alignment can also be applied to heterogeneous networks, as far as there are inter-network links defining the correspondence between nodes from two networks. In this study, we perform the first heterogeneous alignment of human interactome and

*To whom correspondence should be addressed.

phenome networks, with inter-network links defined as the causal relationships between genes and diseases.

The underlying rationale for aligning human interactome and phenome networks is the consistency between phenotypic overlap and genetic overlap. That is, phenotypic overlap between two disease phenotypes implies their shared pathogenesis. This consistency assumption has not yet been verified systematically and quantitatively. However, a similar hypothesis, that similar diseases (or mutant phenotypes) are caused by functionally related genes (Oti and Brunner, 2007), has been supported by more and more evidences from not only model organisms (Fraser and Plotkin, 2007; Lee *et al.*, 2008; McGary *et al.*, 2007) but also human (Goh *et al.*, 2007; Lage *et al.*, 2007; Lim *et al.*, 2006; van Driel *et al.*, 2006; Wood *et al.*, 2007), and has led to remarkable success in screening candidate disease genes (Lage *et al.*, 2007; Wu *et al.*, 2008). Recently, van Driel *et al.* (2006) quantified the pairwise phenotypic similarity/overlap among 5080 human disease phenotypes by examining the overlap of medical terms that describe the phenotypes. Later, Rzhetsky *et al.* (2007) estimated the genetic overlap between 161 disorders based on their frequency of co-occurrence in 1.5 million patient records. With these quantitative data, we are able to verify the correlation between phenotypic overlap and genetic overlap.

2 METHODS

2.1 Data source

The gene network contains 34 364 manually curated protein–protein interactions of 8919 human genes, and is obtained from HPRD (Mishra *et al.*, 2006). The phenotype network consists of 5080 human phenotypes defined in the OMIM database (McKusick, 2007) and the pairwise similarity scores are calculated by text mining, reported by van Driel *et al.* (2006). The gene–phenotype links are defined in the morbidmap of OMIM and 1428 can be mapped to our dataset. The genetic overlap estimation between 161 disorders is published by Rzhetsky *et al.* (2007). Disease category information is from a manual classification concerning the physiological system affected (Goh *et al.*, 2007). Linkage loci with unknown molecular basis are extracted from the OMIM database (entries with prefix %). Gene position information is obtained from NCBI.

2.2 Network alignment and bi-module analysis

We use the network comparison toolkit developed by Ideker lab for network alignment (http://chianti.ucsd.edu/nct/index.php), which implements the model proposed by Sharan *et al.* (2005). Here, we briefly describe the framework applied to our problem. First, the input networks are assembled into a network alignment graph, and then a log likelihood ratio model is used to score the sub-networks on the weighted alignment graph. The scoring model compares the fit of a sub-network to the desired structure (linear path or clique) versus its likelihood given that each network is randomly constructed. Finally, an algorithm searches exhaustively over the alignment graph to identify high-scoring sub-networks. We have tried most of the tunable parameters in this algorithm, and found that they actually have quite limited impact. Therefore, we use their default settings. We call the identified pairs of sub-networks *bi-modules*, each comprising a disease module (the disease sub-network) and a gene module (the gene sub-network), together with gene–disease links between them. We perform enrichment analysis to find over-represented gene functions and disease categories for each bi-module. Gene functions (Gene Ontology terms) analysis for the gene module is carried out by DAVID (Dennis *et al.*, 2003): http://david.abcc.ncifcrf.gov/. The *P*-value of enriched disease category is calculated using Fisher's exact test, which has been widely used for enrichment analysis (Al-Shahrour *et al.*, 2007; Beissbarth and Speed, 2004).

2.3 Benchmark test and prediction

We test the disease gene prediction framework using phenotype network with edge weight threshold of 0.50, 0.55, 0.60 and 0.65. For a threshold smaller than 0.5, the dataset is too large for the program to run, while for a threshold larger than 0.65, the gene–phenotype links are too few for a statistically reasonable validation. At each threshold, the remaining gene–disease links are used to construct the benchmark data. For each gene–disease link, we simulate a linkage locus around the true disease gene by including 108 neighboring genes as negative controls. This strategy for resembling known disease loci in the OMIM database has been widely used in previous studies (Lage *et al.*, 2007; Wu *et al.*, 2008). The 109 test genes are then treated equally by assuming links to the disease under study and go through the network alignment procedure. The genes will compete with each other in this procedure, and the one retained in the bi-module with the highest score is predicted as the causative gene. For prediction, all settings are the same as in the benchmark test, except that the genetic loci are real linkage results collected in OMIM instead of simulated loci.

3 RESULTS

3.1 Phenotypic overlap implies genetic overlap

Assuming that there are shared genetic variations underlying multiple disorders that co-occur in individual patients significantly more (or significantly less) frequently than expected, Rzhetsky *et al.* inferred the genetic overlaps between 161 disorders based on 1.5 million patient records and a sophisticated statistical model (Rzhetsky *et al.*, 2007). To investigate the phenotypic overlap between the 161 disorders, we manually map OMIM phenotypes to these disorders. Frequently, more than one OMIM phenotype will be found for one disorder. In such cases, the highest pairwise phenotypic overlap score between mapped OMIM phenotypes, one from each disorder, is used as the score for corresponding disorder pair (using score mean yields similar results). We are able to map at least one OMIM phenotype entry for 107 of the 161 disorders and assign phenotypic overlap scores to them.

We compare the average genetic overlap between disorder pairs with phenotypic overlap larger than a threshold (T) and those of smaller. At each threshold, the disorder pairs are divided into two groups: those with phenotypic overlap scores smaller than the threshold and those with phenotypic overlap scores larger than the threshold. Then, the average genetic overlap score is calculated for each group separately, and the result is plotted as bars. Results for $T = 0.4$, 0.5 and 0.6 are plotted in Figure 1a. We find indeed that disorder pairs with higher phenotypic overlap have higher genetic overlap, and this contrast becomes sharper for higher phenotypic overlap score threshold. We also calculate the Pearson's correlation coefficient (PCC) between the genetic overlap and phenotypic overlap of the same disorder pair. Similarly, we check the correlation of phenotypic overlap and genetic overlap for disorder pairs with different levels of phenotypic overlap. Given a threshold for phenotypic overlap scores, we calculate the correlation coefficient for disorder pairs whose phenotypic overlap is larger than the threshold. We first transform the genetic overlap score by a log formula $y = \ln(1 + x)$, because the score ranges from zero to several thousand. Most of the genetic overlap scores are positive, but some are negative [co-occur less frequently than expected, interpreted as

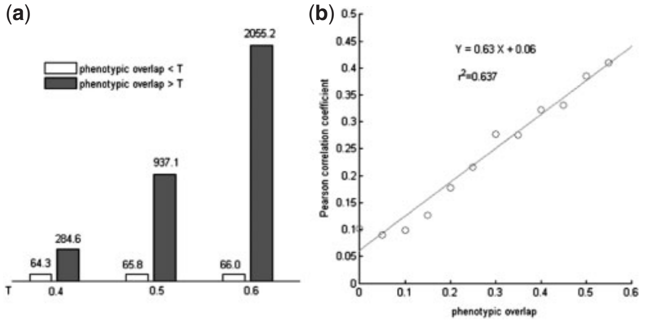

Fig. 1. Phenotypic overlap implies genetic overlap. (**a**) Average genetic overlap of disorder pairs with phenotypic overlap larger than a threshold (T) versus those of smaller. Results with $T = 0.4$, 0.5, 0.6 are given. (**b**) The correlation between phenotypic overlap and genetic overlap becomes stronger when phenotypic overlap increases. Each point (circle) represents the correlation coefficient (Y) for genetic overlap and phenotypic overlap scores when considering disorder pairs with phenotypic overlap larger than a threshold (X).

a genetic overlap via competition (Rzhetsky *et al.*, 2007)]. For the negative ones, we use their absolute value, but analysis excluding negative scores yields similar results. Results (Fig. 1b) show that the overall correlation is weak (PCC = 0.1), but very significant ($P = 1.2\times10^{-13}$). Further, for disorder pairs with higher phenotypic overlap, the correlation becomes stronger, and there is a linear relationship between the correlation coefficient and the phenotypic overlap score (Fig. 1b). For disorder pairs with phenotypic overlap scores larger than 0.6, the correlation coefficient is larger than 0.4. These results confirm that phenotypic overlap is a general indicator of shared pathogenesis.

3.2 Align human interactome and phenome networks

With the consistency between phenotypic overlap and genetic overlap justified, we perform the first heterogeneous alignment of human interactome and phenome networks, to identify pairs of matched sub-networks, or bi-modules. To obtain meaningful results, and also to make it computationally feasible, we remove phenotype links with phenotypic overlap scores <0.5, resulting in a smaller phenotype network (4256 phenotypes and 30 551 edges). The alignment identifies several hundred bi-modules, but there are significant overlap of nodes and edges between them. Using the program's default filtering procedure, we obtain 39 bi-modules with <80% duplications. Two representative bi-modules are shown in Figure 2 (see Supplementary Material 1 for all 39 bi-modules). We find that most diseases in the same module belong to the same category. For example, in Figure 2a, 12 of the 13 diseases in the module are neurological diseases, and in Figure 2b, all diseases are metabolic diseases. The enrichment for specific disease category is not surprising, given that diseases in the same module share significant phenotypic overlap with each other. We also find that genes in the same module are enriched in specific biological processes. For example, the eight genes in Figure 2a are enriched in neurotransmitter secretion and its regulation, dopamine/catecholamine metabolic process and apoptosis, while the six genes implicated in metabolic diseases in Figure 2b highlight the cholesterol/sterol metabolic and transport process (Supplementary Table S1 and S2). We also find that these genes are enriched in specific molecular function, and cellular component (Supplementary Material 2). These enriched common features are consistent with the pathogenesis of diseases in the module, suggesting that the causative gene network may serve as a common pathway for the disease family. To see if these observations are general for bi-modules, we perform gene function enrichment analysis and disease category enrichment analysis for each bi-module. Table 1 lists the most enriched category and function (Gene Ontology biological process terms) for each bi-module. From the table, we can see that all bi-modules are enriched with a specific category and a specific function at a significance level of 0.1, and 38 of the 39 bi-modules are further enriched at a level of 0.02, for both function and category. These results confirm that the identified bi-modules are biologically meaningful. Again, we can see reasonable

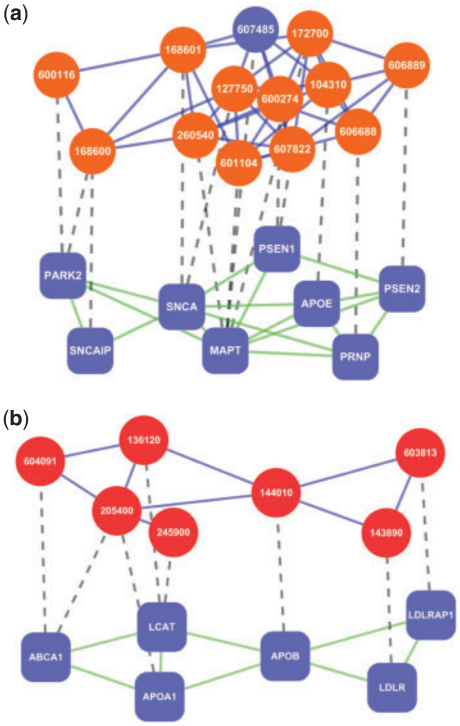

Fig. 2. Representative bi-modules. Circles are diseases and rectangles are genes. Orange, red and blue circles indicates neurological, metabolic and unclassified diseases. Edge between diseases indicates that the two-ended diseases share significant phenotypic overlap (score >0.5). Edge between genes indicates physical interaction between protein products of two-ended genes. Dashed line between gene and disease indicates a causal relationship. (**a**) A neurological bi-module. 607822: AD 3, 260540: Parkinson-dementia syndrome, 104310: AD 2, 600274: Frontotemporal dementia, 607485: Frontotemporal lobar degeneration with ubiquitin-positive inclusions, 606688: Spongiform encephalopathy with neuropsychiatric features, 606889: AD 4, 168601: Parkinson disease, familial, type 1, 601104: Supranuclear palsy, progressive, 1, 127750: Dementia, lewy body, 168600: Parkinson disease, 172700: Pick disease of brain, 600116: Parkinson disease 2, autosomal recessive juvenile. (**b**) a metabolic bi-module. 136120: Fish-eye disease, 144010: Hypercholesterolemia, autosomal dominant, type b, 143890: Hypercholesterolemia, autosomal dominant, 604091: Hypoalphalipoproteinemia, primary, 603813: Hypercholesterolemia, autosomal recessive, 205400: Tangier disease, 245900: Lecithin:cholesterol acyltransferase deficiency.

Table 1. The most enriched category and function

ID	Category	*P*-value	Function	*P*-value
1	Hematological	0.0E + 00	Epidermis development	4.1E − 08
2	Hematological	0.0E + 00	Epidermis development	2.8E − 07
3	Muscular	0.0E + 00	Muscle development	8.5E − 13
4	Muscular	0.0E + 00	Muscle system process	3.1E − 10
5	Muscular	0.0E + 00	Muscle development	8.9E − 18
6	Muscular	3.5E − 04	Muscle system process	5.0E − 12
7	Multiple	5.0E − 04	Response to DNA damage stimulus	2.2E − 07
8	Multiple	5.0E − 04	Multicellular organismal process	1.7E − 06
9	Neurological	5.4E − 10	Regulation of neurotransmitter secretion	1.2E − 05
10	Neurological	4.8E − 06	Mechanosensory behavior	3.9E − 04
11	Neurological	1.7E − 02	Peroxisome organization and biogenesis	2.1E − 16
12	Hematological	5.2E − 03	Blood coagulation	2.1E − 10
13	Metabolic	3.3E − 07	Cholesterol metabolic process	2.6E − 12
14	Muscular	8.3E − 04	Synaptic transmission, cholinergic	4.0E − 06
15	Metabolic	3.3E − 07	Cholesterol metabolic process	2.6E − 12
16	Neurological	9.7E − 02	Protein import into peroxisome matrix	3.1E − 07
17	Dermatological	1.6E − 04	NucleotidE-excision repair	1.0E − 10
18	Ophthamological	1.1E − 07	Visual perception	1.0E − 05
19	Renal	2.3E − 09	Visual behavior	2.6E − 03
20	Bone	6.5E − 04	Skeletal development	2.3E − 07
21	Muscular	8.3E − 04	Synaptic transmission, cholinergic	4.0E − 06
22	Skeletal	2.4E − 03	Regulation of transcription, DNA-dependent	9.9E − 05
23	Skeletal	2.4E − 03	Skeletal development	2.3E − 07
24	Neurological	9.7E − 02	Protein import into peroxisome matrix	3.1E − 07
25	Bone	3.8E − 04	Skeletal development	1.3E − 05
26	Muscular	8.3E − 04	Synaptic transmission, cholinergic	4.0E − 06
27	Renal	1.3E − 08	Visual behavior	2.0E − 03
28	Dermatological	1.0E − 03	Melanin metabolic process	3.9E − 04
29	Cancer	1.2E − 06	DNA-dependent DNA replication	1.5E − 04
30	Dermatological	5.3E − 06	Epidermis development	9.2E − 05
31	Connective tissue	5.3E − 03	Cell adhesion	9.8E − 02
32	Muscular	3.4E − 04	Muscle system process	1.1E − 02
33	Bone	5.3E − 07	Phosphate transport	6.1E − 03
34	Bone	5.3E − 07	Phosphate transport	6.1E − 03
35	Connective tissue	2.7E − 07	Phosphate transport	2.2E − 07
36	Connective tissue	2.7E − 07	Phosphate transport	2.2E − 07
37	Ear,Nose,Throat	7.8E − 03	Sensory perception of sound	8.9E − 03
38	Immunological	2.3E − 05	B cell proliferation	1.1E − 05
39	Cancer	1.3E − 02	Positive regulation of DNA metabolic process	2.6E − 03

correspondences from these results. For example, bi-module 38 is enriched for 'Immunological' disease and the function of 'B cell proliferation'. Another interesting relationship is that, both of the two bi-modules enriched for 'Renal' disease (19 and 27) are associated with genes for 'visual behavior'.

3.3 Predict disease genes via network alignment

The network alignment identifies modular sub-structure between human interactome and phenome networks, reduces the complexity, and facilitates their analysis. One limitation is that these sub-structures are identified within gene–disease relationships that are already known. To make novel discovery, one can incorporate candidate genes that are assumed to involve in a particular disease yet to be confirmed, such as those predicted by computational approaches or those reside in a genomic region identified by linkage analysis or association studies. The candidate gene–disease relationships can be treated as inter-network links and some of them may be retained in bi-modules after the alignment. According to the characteristics of the bi-module, these retained candidate genes share many features with, and are closely connected to, other genes that cause the same or similar diseases. They can explain the phenotypic overlap between these diseases and thus are likely to be true disease genes. We test this hypothesis by a benchmark test with known gene–disease relationships and simulated linkage loci (see Section 2). We call this novel framework AlignPI, which is short for Align Phenome & Interactome. The performance of this approach at different thresholds of phenotypic overlap scores is summarized in Table 2. For example, at the threshold of 0.6, there are 653 known gene–disease links tested, of which 178 cases have at least one test gene matched with the test disease (i.e. retained after alignment), and the average number of matched genes is 3.3. In 129 (72.5%) of the 178 loci, the 3.3 gene list contains (hits) the true disease gene, and the true disease gene can be correctly predicted (retained in the bi-module with the highest score) in 111 cases, yielding a relatively high precision of 0.623 (111/178), and an overall recall rate 0.17 (111/653). In summary, the novel approach greatly reduces the number of candidate genes (from 109 to 3.3) and is able to find the disease gene with high precision.

Table 2. Performance at different score threshold

Score threshold	0.50	0.55	0.60	0.65
Phenotypes	4256	3483	2646	1938
Edges between phenotypes	30 551	16 862	9840	6027
Cases/Known gene–phenotype links	1149	887	653	474
Match rate	0.488	0.318	0.273	0.21
(no. of cases with matched genes)	(561)	(282)	(178)	(100)
Average no. of matched genes	4.5	3.79	3.3	3.43
Hit rate	0.576	0.628	0.725	0.79
(No. of cases)	(323)	(177)	(129)	(79)
Precision	0.383	0.486	0.623	0.69
(No. of true positive)	(215)	(137)	(111)	(69)
Recall	0.187	0.154	0.170	0.146

The statistics for thresholds of 0.50, 0.55 and 0.65 are also provided in Table 2, from which we can further assess the impact of this parameter on the performance of the proposed framework. We show that the threshold indeed has impact on the values of precision and recall. However, one cannot say which threshold is the best, because the threshold just introduces a tradeoff between precision and recall—that is, the precision is generally higher for larger threshold, but the recall will be lower.

We assess the significance of these results by performing a permutation test. The known gene–disease links are randomly rewired to remove the modularity in gene–disease relationship. Then the same benchmark is performed for each randomized dataset. We repeat this procedure 30 times for the threshold at 0.60, and summarize the results in Figure 3. We can see that without modularity the performance drops drastically. Much fewer genes are found; the precision and recall are much lower; the ability to

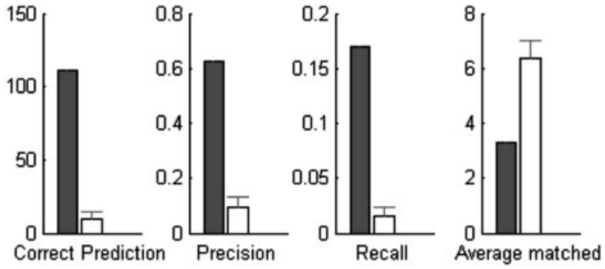

Fig. 3. Comparison of performance for real and randomized gene–disease links. The number of correctly predicted locus, precision, recall and average matched genes are shown. In each panel, the left bar indicates results on known gene–disease links, and the right bar shows the mean and SD of the results with randomized gene–disease links.

enrich disease genes to a short list is also significantly weakened. These results confirm that AlignPI is able to explore the modularity of gene–disease relationship for disease gene discovery.

3.4 Predict novel disease gene

The OMIM database collects 876 genetic loci previously associated with particular disease but without the causal gene identified. For example, in 1997, three groups (Bowden *et al*., 1997; Ji *et al*., 1997; Zouali *et al*., 1997) reported linkage to a 20-cM region of 20q12–q13.1 for type II diabetes. Ten years later, we still do not know which of the 175 genes in this region accounts for the linkage. These genetic loci are great treasure to be explored for human disease genetics. We are able to map at least one gene for each of 591 such loci that are included in our data. The average and median number of genes in these loci are 337.2 and 268, respectively. The proposed framework makes predictions for 70 disease loci. Averagely, 7.4 genes are matched with the test phenotype (See Supplementary Material 3 for all predictions).

Here, we show the example of Late-onset familial Alzheimer disease (AD) (MIM: 608907), mapped to 19p13.2 (Wijsman *et al*., 2004). Two of the 207 genes within this locus, LDLR and ICAM5, reside in a bi-module that has the highest score (Supplementary Fig. S1). LDLR (low density lipoprotein receptor) has already been speculated as an AD gene, because it is a receptor of the AD gene APOE, and modulates the homeostasis of cholesterol, which itself appears associated with AD. Previous population studies (Gopalraj *et al*., 2005) supported that the LDLR haplotype is associated with reduced odds of AD.

Compared with LDLR, the link between ICAM5 and AD is less studied. The ICAM5 protein (intercellular adhesion molecule 5, or TLN—telencephalin) is expressed in the somadendritic region of neurons of the mammalian brain, and may be a critical component in neuron–microglial cell interactions in the course of normal development or as part of neurodegenerative diseases (Gahmberg *et al*., 2008). It is involved in immune privilege of the brain and acts as an anti-inflammatory agent (Tian *et al*., 2008). More directly, the immunoreactivity of ICAM5 is markedly decreased in the brain of AD patients, particularly in the hippocampal formation (Hino *et al*., 1997). Soluble ICAM5 has been detected in brain ischemia (Guo *et al*., 2000), encephalitis (Lindsberg *et al*., 2002) and epilepsy (Rieckmann *et al*., 1998). Further, ICAM5 directly binds to two AD genes: PSEN1 and PSEN2, and other member of the ICAM family has been implicated in AD (Combarros *et al*., 2005). These evidences strongly support a role of ICAM5 in AD.

3.5 Crohn's disease: genome-wide screen and multi-loci effect

The above locus by locus prediction scheme seeks to find from a single locus a gene that is probably part of an *existing* bi-module that contains the disease under investigation. The term 'existing' is used because the bi-module is largely shaped by already established gene–disease relationships. The novel locus is assumed to contain only one true disease gene, thus has limited impact in defining bi-modules.

For complex diseases with heterogeneous origins, there are often multiple loci identified without the causative genes specified. It is likely that the implicated genes from these loci interact with each other and form a novel modular structure/pathway for the disease. In such a scenario, the locus by locus scheme would fail to find the causative genes from these loci. To account for the potential effect of unknown interacting loci, we could fuse candidate genes in multiple loci as if they came from a single locus so that all genes inside could be aligned simultaneously and the potential interacting effect could be automatically considered.

We test this multi-loci scheme for Crohn's disease (CD). Recently, a meta-analysis of three genome-wide association studies (Burton *et al*., 2007; Libioulle *et al*., 2007; Rioux *et al*., 2007) reported 40 susceptibility loci for CD (Barrett *et al*., 2008). These loci correspond to 37 distinct chromosome regions (Tables 2 and 3 in Barrett *et al.*'s paper) containing 6154 genes in total. We first perform the single locus scheme for each locus and no significant bi-modules are identified. The result suggests that current functional (interactome) data does not support the idea that genes in these novel loci are part of a known CD-related bi-module. However, as pointed out earlier, there is a possibility that the combination of several genes in these loci renders some local structure to be significant enough to become a novel bi-module. To test all possible combinations, we fuse all 6154 genes in the 37 regions into one region, and align CD (MIM 266600) with all genes simultaneously. This genome-wide alignment identifies 48 candidate genes that might be associated with CD (Table 3). Three of the 48 genes (STAT3, JAK2 and PTPN2, darkgray rows in Table 3) are inside the critical region defined by genome-wide association studies (Barrett *et al*., 2008), and all three genes are also proposed as the potential causative genes by Barrett *et al*. Two (STAT3 and JAK2) of these three genes are inside the same bi-module with the highest score. Beside these three genes, nine genes (light gray rows in Table 3) are <1 Mb away from the critical region, such as STAT5A (23-kb upstream), STAT5B (60-kb upstream) and MST1R (30-kb downstream). Of the nine genes near the critical regions, three (SUMO4, GRB10 and CARD6) are near a critical region that contains no genes. Besides these candidate genes that are consistent with genome-wide association studies, we also identify many other genes that are plausible candidates. Of particular interest are the other two genes in the most highly scored bi-module: IL12RB1 and FYN. Further works are needed to verify the role of these genes in CD pathogenesis. Nonetheless, the above results not only demonstrate the usefulness of our novel method, but also illustrate the ability of the method to perform genome-wide prediction and to handle multi-loci effect.

Table 3. Potential CD genes identified

Rank	Gene	Score	Region	Critical region	Distance to critical region
1	STAT3	19.99	17q21	34.63–35.34, 37.74–37.95	Inside
2	IL12RB1	19.99	19p13	1.05–1.15	>1Mb
3	FYN	19.99	6q21	106.52–106.62	>1Mb[a]
4	JAK2	19.99	9p24	4.94–5.26	Inside
5	PTK2	19.98	8q24	126.60–126.62	>1Mb
6	ERBB2	10.69	17q12	29.57–29.70	>1Mb
7	BRCA1	4.024	17q21	34.63–35.34, 37.74–37.95	500 kb down
8	RELA	4.024	11q13	75.80–76.02	>1Mb
9	PRKCD	4.024	3p21	48.73–49.87	>1Mb
10	NR3C1	4.024	5q31	131.44–131.90	>1Mb
11	SUMO4	4.023	6q25	149.54–149.65	110 kb down[a]
12	INSR	4.023	1q23	157.65–157.72	>1Mb
13	HTATIP	4.023	11q13	75.80–76.02	>1Mb
14	JAK1	4.023	1p31	67.4	>1Mb
15	VAV1	4.023	19p13	1.05–1.15	>1Mb
16	NCOA1	4.023	2p23	27.30–27.77	>1Mb
17	HDAC3	4.023	5q31	131.44–131.90	>1Mb
18	STAT5A	4.022	17q21	34.63–35.34, 37.74–37.95	23 kb up
19	PDGFRB	4.022	5q33	150.15–150.32	640 kb up
20	CDKN1A	4.022	6p21	32.44–32.79	>1Mb
21	PPP2CA	4.022	5q31	131.44–131.90	>1Mb
22	STAT5B	4.022	17q21	34.63–35.34, 37.74–37.95	60 kb up
23	CCND1	4.022	11q13	75.80–76.02	>1Mb
24	CCR5	4.022	3p21	48.73–49.87	>1Mb
25	EPOR	4.022	19p13	1.05–1.15	>1Mb
26	PTMA	4.022	2q37	230.9	>1Mb
27	JAK3	4.022	19p13	1.05–1.15	>1Mb
28	NGFR	4.022	17q21	34.63–35.34, 37.74–37.95	>1Mb
29	YES1	4.021	18p11	12.73–12.88	>1Mb
30	GRB10	4.021	7p12	50.03–50.11	510 kb down[a]
31	IFNAR1	4.021	21q22	44.43–44.48	>1Mb
32	MST1R	4.021	3p21	48.73–49.87	30 kb down
33	HSP90AB1	4.021	6p21	32.44–32.79	>1Mb
34	CCR1	4.021	3p21	48.73–49.87	>1Mb
35	PTPN2	4.021	18p11	12.73–12.88	Inside
36	TRAF5	4.021	1q32	197.60–197.77	>1Mb
37	PRLR	4.021	5p13	40.32–40.48	>1Mb
38	HTR2A	4.021	13q14	43.13–43.54	>1Mb
39	PPP2R5A	4.021	1q32	197.60–197.77	>1Mb
40	LEPR	4.021	1p31	67.4	>1Mb
41	IL7R	4.021	5p13	40.32–40.48	>1Mb
42	IFNAR2	4.021	21q22	44.43–44.48	>1Mb
43	IFNGR2	4.021	21q22	44.43–44.48	>1Mb
44	IL12RB2	4.021	1p31	67.4	145 kb down
45	OSMR	4.021	5p13	40.32–40.48	>1Mb
46	IL12B	4.021	5q33	150.15–150.32	>1Mb
47	NDUFA13	4.021	19p13	1.05–1.15	>1Mb
48	CARD6	4.021	5p13	40.32–40.48	400 kb down[a]

[a] The critical region contains no genes.

4 DISCUSSION

As a proof-of-concept analysis, we have not investigated the impact of different network alignment algorithms, though there are a dozen of methods available (Berg and Lassig, 2006; Flannick *et al.*, 2006; Sharan and Ideker, 2006; Singh *et al.*, 2008). There are several reasons for the choice of the NetworkBlast algorithm used here. First, NetworkBlast is one of the pioneering works, and has successfully led to novel biological discoveries (Suthram *et al.*, 2005). Second, it is conceptually simple; especially no evolutionary model is imposed. Most of the methods developed later assume a dynamic evolutionary history between the aligned networks, which cannot be explained for the alignment of heterogeneous networks. Of course it is interesting to see if some of the evolutional operations (such as node duplication and edge deletion) can be used to explain the pathogenesis of disease families. Third, the alignment of NetworkBlast is local, which aims to find matched substructures between networks. We make use of this capability to identify bi-modules. Many later-developed methods perform global alignment, thus are not appropriate here.

Recently, a number of network-based methods have been proposed to predict or prioritize disease gene candidates (Ala *et al.*, 2008; Franke *et al.*, 2006; Koller *et al.*, 2008; Lage *et al.*, 2007; Oti *et al.*, 2006; Wu *et al.*, 2008). Though it is not our primary concern to develop a disease gene prediction method that outperforms existing ones, the good performance of the novel AlignPI framework renders it as one of the top methods in this field. Of these network-based methods, two are of particular interest: the Bayesian predictor proposed by Lage *et al.* (2007) and our previous regression model CIPHER (Wu *et al.*, 2008). These two methods are based on the same types of data as this study: phenotype similarity and protein interaction. In general, the precision of AlignPI is slightly better than CIPHER, though the recall is lower. CIPHER has a precision ranges from 0.47 to 0.66, and a recall ranges from 0.3 to 0.5, while AlignPI can achieve a precision of 0.69 at the score threshold of 0.65, where the recall is 0.15. As a comparison, the precision for the Bayesian predictor ranges from 0.23 to 0.65, and the recall ranges from 0.13 to 0.23.

Certainly, there are several limitations of this study. First, there are imprecision and subjectiveness in quantifying phenotypic overlap score. The standardization and quantification of phenotypic description is another issue that is out of the scope of this study (Biesecker, 2005). Second, though we have shown that the alignment algorithm designed for protein networks is also effective in aligning phenome and interactome networks; it is still worthwhile to design specific algorithms for this problem. For instance, the phenotype network is a weighted complete graph (all pairs are connected), while the protein network is binary and sparse. Specific algorithms are needed to address the alignment problem under this scenario.

Our framework could also be applied to model organisms, providing that there are systematic phenotype similarity and gene interaction data, for example, in *Caenorhabditis elegans* (Gunsalus *et al.*, 2005). Similarly, the framework could also be applied to other labeling problems, such as protein function prediction, as there are also observations of the correlation between protein functional distance (semantic similarity of GO annotations) and network distance (Sharan *et al.*, 2007).

ACKNOWLEDGEMENTS

We thank Brunner lab for providing the phenome data, and Dr MQ Zhang from CSHL, Dr FZ Sun from USC for critically reading the article. We are grateful to the anonymous reviewers whose suggestions and comments contributed to the significant improvement of this article.

Funding: Natural Science Foundation of China (60575014 and 60805010); Hi-Tech Research and Development Program of China (863project) (2006AA02Z325); National Basic Research Program of China (2004CB518605); a startup supporting plan at Tsinghua University.

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al.* (2007) FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.*, **35**, W91–W96.

Ala,U. *et al.* (2008) Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput. Biol.*, **4**, e1000043.

Barrett,J.C. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955–962.

Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Berg,J. and Lassig,M. (2006) Cross-species analysis of biological networks by Bayesian alignment. *Proc. Natl Acad. Sci. USA*, **103**, 10967–10972.

Biesecker,L.G. (2005) Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations. *Clin. Genet.*, **68**, 320–326.

Bowden,D.W. *et al.* (1997) Linkage of genetic markers on human chromosomes 20 and 12 to NIDDM in Caucasian sib pairs with a history of diabetic nephropathy. *Diabetes*, **46**, 882–886.

Burton,P.R. *et al.* (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.

Chen,Y. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**, 429–435.

Combarros,O. *et al.* (2005) Interaction between interleukin–6 and intercellular adhesion molecule–1 genes and Alzheimer's disease risk. *J. Neurol.*, **252**, 485–487.

Dennis,G. *et al.* (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Flannick,J. *et al.* (2006) Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res.*, **16**, 1169–1181.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

Fraser,H.B. and Plotkin,J.B. (2007) Using protein complexes to predict phenotypic effects of gene mutation. *Genome Biol.*, **8**, R252.

Gahmberg,C.G. *et al.* (2008) ICAM-5—A novel two-facetted adhesion molecule in the mammalian brain. *Immunol. Lett.*, **117**, 131–135.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Gopalraj,R.K. *et al.* (2005) Genetic association of low density lipoprotein receptor and Alzheimer's disease. *Neurobiol. Aging*, **26**, 1–7.

Gunsalus,K.C. *et al.* (2005) Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature*, **436**, 861–865.

Guo,H. *et al.* (2000) Release of the neuronal glycoprotein ICAM-5 in serum after hypoxic-ischemic injury. *Ann. Neurol.*, **48**, 590–602.

Hino,H. *et al.* (1997) Reduction of telencephalin immunoreactivity in the brain of patients with Alzheimer's disease. *Brain Res.*, **753**, 353–357.

Ji,L.N. *et al.* (1997) New susceptibility locus for NIDDM is localized to human chromosome 20q. *Diabetes*, **46**, 876–881.

Koller,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Lee,I. *et al.* (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat. Genet.*, **40**, 181–188.

Libioulle,C. *et al.* (2007) Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.*, **3**, 6.

Lim,J. *et al.* (2006) A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, **125**, 801–814.

Lindsberg,P.J. *et al.* (2002) Release of soluble ICAM-5, a neuronal adhesion molecule, in acute encephalitis. *Neurology*, **58**, 446–451.

McGary,K.L. *et al.* (2007) Broad network-based predictability of *S. cerevisiae* gene loss-of-function phenotypes. *Genome Biol.*, **8**, R258.

McKusick,V.A. (2007) Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.

Mishra,G.R. *et al.* (2006) Human protein reference database – 2006 update. *Nucleic Acids Res.*, **34**, D411–D414.

Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.

Oti,M. *et al.* (2006) Predicting disease genes using protein–protein interactions. *J. Med. Genet.*, **43**, 691–698.

Pujana,M.A. *et al.* (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.*, **39**, 1338–1349.

Rieckmann,P. *et al.* (1998) Telencephalin as an indicator for temporal-lobe dysfunction. *The Lancet*, **352**, 370–371.

Rioux,J.D. *et al.* (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.*, **39**, 596–604.

Rzhetsky,A. *et al.* (2007) Probing genetic overlap among complex human phenotypes. *Proc. Natl Acad. Sci. USA*, **104**, 11694–11699.

Sharan,R. and Ideker,T. (2006) Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.*, **24**, 427–433.

Sharan,R. *et al.* (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.

Sharan,R. *et al.* (2007) Network-based prediction of protein function. *Mol. Syst. Biol.*, **3**, 88.

Singh,R. *et al.* (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl Acad. Sci. USA*, **105**, 12763–12768.

Suthram,S. *et al.* (2005) The Plasmodium protein network diverges from those of other eukaryotes. *Nature*, **438**, 108–112.

Tian,L. *et al.* (2008) Shedded neuronal ICAM-5 suppresses T-cell activation. *Blood*, **111**, 3615–3625.

van Driel,M.A. *et al.* (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

Wijsman,E.M. *et al.* (2004) Evidence for a novel late-onset Alzheimer disease locus on chromosome 19p13.2. *Am. J. Hum. Genet.*, **75**, 398–409.

Wood,L.D. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.

Wu,X. *et al.* (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.

Zouali,H. *et al.* (1997) A susceptibility locus for early-onset non-insulin dependent (type 2) diabetes mellitus maps to chromosome 20q, proximal to the phosphoenolpyruvate carboxykinase gene. *Hum. Mol. Genet.*, **6**, 1401–1408.