## Doctoral Dissertation

# Evaluation of predicted protein complexes emerged from protein interaction network and proposition of prediction method toward topological structure of protein complexes Kensaku Nishikata 

February 4, 2010

Department of Bioinformatics and Genomics
Graduate School of Information Science
Nara Institute of Science and Technology

> A Doctoral Dissertation
> submitted to the Graduate School of Information Science, Nara Institute of Science and Technology
> in partial fulfillment of the requirements for the degree of Doctor of Engineering.

| Thesis Committee: |  |
| :--- | :--- |
| Professor Shigehiko Kanaya | (Supervisor) |
| Professor Naotake Ogasawara | (Co-supervisor) |
| Associate Professor Md.Altaf-Ul-Amin | (Co-supervisor) |
| Assistant Professor Hiroki Takahashi | (Co-supervisor) |

# Evaluation of predicted protein complexes emerged from protein interaction network and proposition of prediction method toward topological structure of protein complexes * 

Kensaku Nishikata


#### Abstract

Proteins interact with other proteins or biomolecules to perform their functions, and protein complexes are the fundamental functional units of these macromolecular systems: protein interactions play a key role in many cellular processes. Therefore, elucidating protein-protein interactions (PPIs) leads to understanding the protein functions required for various biological processes in cells. More or less over the past 10 years, vast amount of PPI data have been generated by high-throughput methods for detecting protein interactions. However, there being no complete and accurate detection method, each experimental strategy generates a significant number of false-negatives and false-positives. Additionally, experimental methods which identify protein complexes such as affinity purification-mass spectrometry (MS) equally detect direct and non-direct interactions i.e. "bait-prey" and "prey-prey" interactions. These false-positives and confusion non-direct interactions with direct interactions are serious problems because they cause erroneous results and misleading conclusions.

The goal of this dissertation based on above background is to predict and evaluate certain protein complexes, more specifically i) to predict members of protein complexes and to evaluate that by annotation and ii) to propose a method to predict the topological structure of protein complexes. In analysis on human PPI network presented in Chapter 2, totally, 1,264 protein complexes were predicted by finding densely connected regions with their cluster properties in the network, and these predicted complexes were annotated and evaluated using integrated data such as literatures and research papers, ternary structures, description of proteins, localizations, expression profiles etc. The study on Arabidopsis interactome of


Chapter 3 proposes a method to predict the topological structure of protein complexes by using domain-domain interactions (DDIs). As the first step, I extracted 312 statistically significant DDIs out of 1,162 DDIs underlying 3,118 protein-protein interactions (PPIs). Next, 67 protein complexes were obtained by protein interaction network analysis. Finally, I discussed the topological structure of protein complexes based on DDI information extracted in the first step.

## Keywords:

topological structure of protein complexes, protein-protein interactions (PPIs), domain-domain interactions (DDIs), graph theory, false-positive and false-negative, "bait-prey" and "prey-prey" interactions

[^0]
## Acknowledgements

First of all, I wish to express my deepest gratitude to Professor Shigehiko Kanaya, who has been the adviser for the research described in this dissertation. Regardless of his busy schedule, he has supported my wishes through many kinds of efforts for the accomplishment of this dissertation. I could have never got such wonderful experiences without his warmest supports. I never forget his warmest assistance for me.

I would also like to thank Associate Professor Md. Altaf-Ul-Amin for his valuable suggestions and comments, and reviewing my papers and this dissertation, to Assistant Professor Hiroki Takahashi and Professor Naotake Ogasawara for their suggestions and comments, and review of this dissertation.

I'm highly thankful to the members of Comparative genomics laboratory, especially Mr. Masayoshi Wada, who really supported domain detection by InterProScan, and Ms. Aki Hirai, who constantly supported my clerical ministrations.

I'm sincerely thankful to the members of now-defunct JBIRC (Japan Biological Information Research Center), particularly Dr. Shingo Kikugawa, who and I designed and performed human PPI analysis together, Dr. Yoshiharu Sato, Dr. Katsuhiko Murakami, and Team Leader Tadashi Imanishi. This organization gave me the opportunity to research human PPI.

Finally, I would like to offer deep and special thanks to my family. My parents and my sister babysat my baby son. Under their supporting, I have been able to spend considerable time on this work. And, above all, Naoko, who is my wife, understood my wish for taking the doctorate, and she has always warmly encouraged me to accomplish my work and this dissertation. Without her understanding and supporting, I couldn't have got this opportunity and achieved this work beside my business. I must never forget this appreciation of her supporting in my life.

February 2010
Kensaku Nishikata

## Contents

List of Figures ..... vi
List of Tables ..... vii
Chapter 1 Introduction ..... 1
1.1. Dissertation outline ..... 1
1.2. Protein-protein interaction and the detection ..... 2
1.2.1. Yeast two-hybrid (Y2H) method ..... 4
1.2.2. Affinity purification-MS method .....  .7
1.3. Problems of PPIs analysis ..... 10
1.3.1. Problems of false-negative and false-positive ..... 10
1.3.2. Problems of affinity purification-MS Data ..... 12
1.4. Suggestive ideas for approach of this dissertation ..... 15
1.4.1. Handling of affinity purification-MS data ..... 15
1.4.2. Graph clustering software, DPClus ..... 19
1.5. The goal of this dissertation ..... 20
Chapter 2 Prediction and evaluation of human protein complexes from the integrated protein interaction network ..... 21
2.1. Abstract ..... 21
2.2. Introduction ..... 22
2.3. Materials and Methods ..... 24
2.3.1. Integration of PPI data onto the $\mathrm{H}-\mathrm{InvDB}$ proteins ..... 24
2.3.2. Parameter optimizations of DPClus in prediction of protein complexes ..... 25
2.3.3. Manual annotation of the predicted protein complexes ..... 29
2.3.4. Consistency of GO terms assigned to the members in a complex ..... 30
2.3.5. Similarity of gene expression profiles among proteins in the same complex ..... 31
2.3.6. Judgement of the duplicated gene pair in a given protein complex ..... 32
2.4. List of abbreviations used ..... 32
2.5. Results ..... 33
2.5.1. Integration of PPI data onto the H -InvDB proteins ..... 33
2.5.2. Prediction of protein complexes ..... 36
2.5.3. Manual annotation of the predicted protein complexes ..... 39
2.5.4. Consistency of GO terms assigned to members in a complex ..... 44
2.5.5. Similarity of gene expression profiles among proteins in the same complexes ..... 48
2.5.6. Relationship between the formation of protein complexes and gene duplication ..... 53
2.5.7. Functional assignments for hypothetical proteins in the annotated complexes ..... 54
2.5.8. Development of "PPI view" and "PPI map" ..... 57
2.6. Discussion ..... 63
2.7. Conclusion ..... 65
Chapter 3 Predicting topological structure of protein complexes by determining statistically significant domain-domain interactions. ..... 67
3.1. Abstract ..... 67
3.2. Introduction ..... 68
3.3. Methods ..... 72
3.3.1. Step 1: Detection of domains in proteins ..... 75
3.3.2. Step 2: Statistically extracting domain-domain interactions from PPI data ..... 75
3.3.3. Step 3: Validation of protein complexes generated using DPClus ..... 76
3.4. Results and Discussion ..... 77
3.4.1. Collection and integration of PPIs data ..... 77
3.4.2. Significant domain-domain interactions extracted from PPI data ..... 77
3.4.3. Prediction of protein complexes ..... 81
3.5. Conclusion and Remarks ..... 88
Chapter 4 Conclusion and Remarks. ..... 89
References ..... 93
Achievements ..... 100
Appendix ..... 103

## List of Figures

Figure 1.1 Y2H method illustration. ..... 6
Figure 1.2 Affinity purification-MS illustration. ..... 9
Figure 1.3 Topological structure diversity of protein complexes ..... 14
Figure 1.4 PPIs models generated from affinity purification-MS data ..... 17
Figure 1.5 Advantages of adopting spoke model for affinity purification-MS data ..... 18
Figure 2.1 Precision-Recall curve for each of the 100 parameter combinations tested during parameter optimization ..... 28
Figure 2.2 Overlap of human PPIs in six PPI databases. ..... 35
Figure 2.3 View of Integrated PPI network. ..... 38
Figure 2.4 Properties of annotated complexes ..... 43
Figure 2.5 Distributions of GO consistency index in PCset1, PCset2, and random set. ..... 47
Figure 2.6 Relative percentage of gene expression levels of troponin complex. ..... 51
Figure 2.7 Boxplot of gene expression profile similarity and the number of protein members in a complex ..... 52
Figure 2.8 Database of PPIs and predicted and annotated protein complexes, "PPI view ..... 59
Figure 2.9 Visualization tool of protein and protein complex network, "PPI map" ..... 60
Figure 2.10 System architecture of PPI map ..... 61
Figure 2.11 An overhead view of complex-complex interactions (CCIs) with the subcellar localizations of the annotated complexes ..... 62
Figure 3.1 Topological structure diversity of protein complexes in case that proteins A, B and C are identified by MS ..... 71
Figure 3.2 Procedure of validation of protein complexes based on extracted DDIs ..... 73
Figure 3.3 Validation of statistically extracted DDIs using iPfam ..... 80
Figure 3.4 Complex-complex interaction network emerged from PPI network by adopting DPClus ..... 85
Figure 3.5 Prediction of the topological structure of protein complexes using significant DDIs information ..... 86

## List of Tables

Table 1.1 Different experimental methods of PPI detection. ........................................... 3
Table 2.1 Protein and the complex annotation summary. .............................................. 42
Table 2.2 The protein complexes consisting of members with significantly similar gene
expression profiles................................................................................. 50
Table 2.3 The hypothetical proteins in the annotated complexes. .................................. 56

## Appendix Tables

Appendix Table 1 List of PPIs supported by statistically significant domain-domain interactions.103
Appendix Table 2 Predicted protein complexes with statistically significant domain-domain interaction information. ..... 175

## Chapter 1 Introduction

### 1.1.Dissertation outline

Chapter 1 explains the background and the problem of protein-protein interaction (PPI) analysis, and indicates the goal of this dissertation. Section 1.2 explains detection method of PPIs and Section 1.3 elucidates problems of PPIs analysis caused by PPI detection method especially focused on the affinity purification-MS, which is one of the PPI detection methods. Section 1.4 gives some suggestive ideas for a common approach of studies of this dissertation. Taking these matters into consideration, Section 1.5 indicates the goal of this dissertation. Aiming to the goal of this dissertation, I tried two kinds of studies: i) predicting members of protein complexes in human and evaluating that by annotation (in Chapter 2) and ii) proposing a method to predict the topological structure of protein complexes by determining statistically significant domain-domain interactions (in Chapter 3).

Chapter 2 considers integrating PPI data from public PPI databases and predicting protein complexes, and evaluating the predicted complexes by annotation using integrated data such as such as literatures and research papers, ternary structures, descriptions of protein, localizations, expression profiles etc.

Chapter 3 proposes a method to predict the topological structure of protein complexes by using domain-domain interactions (DDIs). It is important to discuss the topological structure of protein complexes, because high-throughput methods for detecting protein interactions 1
generate a significant number of false-positives, and additionally, the affinity purification-MS detects non-direct interactions: "prey-prey" interactions.

Finally, in Chapter 4 concluding remarks of this dissertation are described.

### 1.2. Protein-protein interaction and the detection

Proteins interact with other proteins or biomolecules to perform their functions, and protein complexes are the fundamental functional units of these macromolecular systems: protein interactions play a key role in many cellular processes. Therefore, comprehensive analysis of protein-protein interactions (PPIs) provides a valuable framework for understanding the protein functions required for various biological processes in cells.

To elucidate complete picture of PPIs in several organism, more or less over the past 10 years, vast amount of PPIs data have been generated by high-throughput methods for detecting protein interactions (Uetz et al. 2000; Ito et al. 2001; Ho et al. 2002; Gavin et al. 2002). Protein interactions can be analyzed by different genetic, biochemical, and physical methods, which are listed in Table 1.1 (Shoemaker and Panchenko 2007). Although a number of methods are available for high-throughput analysis of PPIs, the most commonly used are the yeast two-hybrid ( Y 2 H ) method and the combinatorial method of protein complex purification and protein identification by mass spectrometry (MS) (affinity purification-MS). The mechanisms of these two representative methods ( Y 2 H and affinity purification-MS) are respectively outlined below.

Table 1.1 Different experimental methods of PPI detection.

High-throughput techniques are indicated with pluses (second column), and those which can provide information on interactions in vivo are shown in the third column. Fourth column indicates whether the method supplies data on physically interacting proteins in a complex ("complex") or only pairwise interactions ("binary"). Methods inferring interactions through functional association are shown as well. The type of protein interaction characterization is shown in the last column.

|  | High- <br> Throughput <br> Approach | Living <br> Cell <br> Assay | Type of Interactions | Type of Characterization |
| :--- | :--- | :--- | :--- | :--- |
|  | + | In vivo | Physical interactions (binary) | Identification |
| Y2H | + | In vitro | Physical interaction (complex) | Identification |
| Affinity purification-MS | + | Invitro | Functional association | Identification |
| DNA microarrays/Gene coexpression | + | In vitro | Physical interaction (complex) | Identification |
| Protein microarrays | + | In vivo | Functional association | Identification |
| Synthetic lethality | + | In vitro | Physical interaction (complex) | Identification |
| Phage display | + | In vitro | Physical interaction (complex) | Structural and biological characterization |
| X-ray crystallography, NMR spectroscopy | - | In vivo | Physical interactions (binary) | Biological characterization |
| Fluorescence resonance energy transfer | - | In vitro | Physical interaction (complex) | Kinetic, dynamic Biological |
| Surface plasmon resonance | - | In vitro | Physical interactions (binary) | Mechanical, dynamic characterization |
| Atomic force microscopy | - | In vitro | Physical interaction (complex) | Structural and biological |
| Electron microscopy |  |  |  |  |

### 1.2.1. Yeast two-hybrid (Y2H) method

Yeast two-hybrid (Y2H) method analyzes the physical interaction between two proteins by detecting the expression of the reporter gene (Shoemaker and Panchenko 2007) (Figure 1.1). The premise behind the test is the activation of downstream reporter gene(s) by the binding of a transcription factor onto an upstream activating sequence (UAS). In Y2H method, the transcription factor is split into two separate fragments, called the binding domain (BD) and activating domain (AD). The BD is the domain responsible for binding to the UAS and the AD is the domain responsible for activation of transcription. For the purposes of testing the interactions, a protein of interest is fused to BD ("bait"). This chimeric protein is cloned in an expression plasmid, which is then transfected into a yeast cell. A similar procedure creates a chimeric sequence of another protein fused to AD ("prey"). If two proteins physically interact, the reporter gene is activated and expressed, and then the expression is detected. The most broadly used Y2H method is GAL4/LexA-based, where the GAL4 protein controls in yeast the expression of the LacZ gene encoding beta-galactosidase, though numerous variations of Y 2 H have been developed including systems with several reporter genes.

For screening entire genomes, the Y 2 H method has been advanced into two main approaches: matrix-based and library-based. In the matrix approach, a matrix of "prey" clones is created where each clone expresses a particular "prey" protein in one well of a plate. Then each "bait" strain is mated with an array of "prey" strains, and those diploids where two chimeric proteins interact are selected based on the expression of a reporter gene and the
position on a plate. In the library approach, each "bait"s is screened against an undefined "prey" library containing random cDNA fragments or open reading frames (ORFs). Diploid positives are selected based on their ability to grow on specific substrates; and interacting proteins are determined by DNA sequencing.


## Figure 1.1 Y2H method illustration.

Yeast two-hybrid ( Y 2 H ) method analyzes the physical interaction between proteins X ("bait") and Y ("prey") by detecting the expression of the reporter gene. If two proteins physically interact, the reporter gene is activated and expressed.

### 1.2.2. Affinity purification-MS method

The affinity purification-MS is the combinatorial method of protein complex purification and protein identification by mass spectrometry (MS), i.e., a protein complex is purified by affinity purification, and then proteins in the complex are identified by mass spectrometry (MS).

The open reading frame (ORF) of a target protein is fused with the DNA sequences encoding the "tag" such as FLAG (Ho et al. 2002), TAP-tag (Gavin et al. 2002) or His-tag (Arifuzzaman et al. 2006), and expressed where it can form native complexes with other proteins. For purification, target protein complexes are fixed in affinity column through the tag of the target protein, and the contaminants are washed out. After washing, the target protein complex is released from affinity column by breaking away the interaction between the tag of target protein and affinity column. The components of each protein complex are screened by polyacrylamide gel electrophoresis, cleaved to fragments by proteases, and identified using subsequent analysis of the fragments by MS. The principle of the MS method is to produce ions which can be detected based on their mass-to-charge ratios, thereby allowing the identification of polypeptide sequences. The problem of converting protein/peptide molecules from the condensed phase into ions in the gas phase is solved by using Electrospray Ionization (ESI) and Matrix Assisted Laser Desorption Ionization (MALDI). Different algorithms have been developed to analyze mass spectra and to identify
proteins by their sequence (Shoemaker and Panchenko 2007).


Figure 1.2 Affinity purification-MS illustration.

Target protein with tag is expressed in cells and form protein complexes. Protein complexes is fixed in affinity column and washed, then released from affinity column. And each protein in complexes is identified by MS.

### 1.3.Problems of PPIs analysis

### 1.3.1. Problems of false-negative and false-positive

The first genome-wide PPIs studies of a free-living organism using the yeast two-hybrid ( Y 2 H ) method have been published by Uetz and coworkers and Ito and collaborators using the yeast Saccharomyces cerevisiae (Uetz et al. 2000; Ito et al. 2001). Soon after these analyses, Ho and coworkers and Gavin and colleagues used the combinatorial method of protein complex purification and protein identification by mass spectrometry (MS), so-called affinity purification-MS (Ho et al. 2002; Gavin et al. 2002).

These two different experimental approaches for detecting protein interactions have already taught us one important lesson. That is two data sets obtained from Y2H and affinity purification-MS are strikingly different but are also highly complementary, i.e., each experimental strategy generates false-negatives and these two experimental methods have propensity to detect different kind of protein interactions: Y2H more often find transient interactions, whereas affinity purification-MS identifies stable interactions such as those in protein complexes (Titz et al. 2004). This variation is caused by the differences in mechanisms of protein interaction detection. Y2H test an interaction of pairs of proteins, and can detect transient interactions, but not detecting highly cooperative and weak interactions which constitute a protein complex. And, artificial domain fused to "bait" protein or "prey" protein in Y 2 H , i.e., the binding domain ( BD ) or activating domain ( AD ) of transcription factor, may interrupt the interactions between "bait" and "prey". On the other hand, in affinity 10
purification-MS, major bottleneck for MS analysis is low abundance of proteins. Proteins that are only weakly associated with protein complexes such as in transient interactions tend to be lost in the process of protein complexes purification.

Furthermore, each experimental strategy generates not only false-negatives but also a significant number of false-positives (Titz et al. 2004). False-positives are usually a more serious problem because they cause erroneous results and misleading conclusions, making PPI analysis complicated and difficult. In Y2H studies, some "bait" constructs activate the reporter gene without interacting with a "prey" and so may generate large numbers of technical false-positives, and over-expression can be a result of non-specific interactions. Moreover, some proteins might specifically interact when they are co-expressed in the yeast, although in reality they are never present in the same cell at the same time. Conversely, in affinity purification-MS, too weak washing during purification can generate technical false-positives: exquisite balance of washing strength is required because too weak washing may generate technical false-positives and too strong washing may generate technical false-negatives. Continuously, it is undeniable that tag of target protein might directly interacts with "prey" proteins, and as is the case with Y2H, some "bait" proteins might specifically interact with "prey" proteins when "bait" proteins are forced-expressed in a host cell, although in reality they are never present in the same cell at the same time.

Affinity purification-MS is a powerful method of studying PPIs, however, in addition to above problems of technical false-positives, it has critical problems about handling
the data caused by fundamental principle of affinity purification-MS. That would be described the next section.

### 1.3.2. Problems of affinity purification-MS Data

An experimental method which identifies protein complexes such as affinity purification-MS, in the strict sense, detects all proteins in multiple complexes with a target protein ("bait"), but does not detect only direct PPIs. Isolation of protein complexes in these experimental procedure allow the purification of the "bait" together with all of the "prey" proteins that belong to the same multi-protein complexes. The problem seems to lie in the fact that every "prey" protein doesn't directly interacts with the "bait" protein; rather, the topological structure of the complex will include both "bait-prey" and "prey-prey" interactions (Hakes et al. 2007). It implies that the topological structure of the protein complexes cannot be determined from the individual experiments only. In case that proteins $\mathrm{A}, \mathrm{B}, \mathrm{C}$ and D are identified in isolation as a protein complex using protein $A$ with a tag as a "bait", it is difficult to conclude what is the true topological structure of the complex out of the four sample cases as shown in Figure 1.3 i) ~iv): if we presume that the topological structure of protein complexes is linear, the order of proteins cannot be decide (the case i) or ii)). In addition, first of all, it is not determinable that topological structure of protein complexes is linear or non-linear (the case i), ii) or iii)). Furthermore, the member of protein complexes cannot be determined, i.e., the all of proteins identified by MS can construct a single complex or multiple complexes (the case i)~iii) or iv)).

Though understanding the topological structure of protein complexes is necessary in order to obtain useful information about them, little attention has been given by the bioinformatics researchers to decipher the topological structure of the protein complexes. Also the experimental technologies used to detect PPIs do not focus on this matter.


Figure 1.3 Topological structure diversity of protein complexes

Some topologies of protein complexes are conceivable in case that proteins $\mathrm{A}, \mathrm{B}, \mathrm{C}$ and D are identified by MS after affinity purification.

### 1.4.Suggestive ideas for approach of this dissertation

### 1.4.1. Handling of affinity purification-MS data

Generally, in practice to analyze PPIs of affinity purification-MS data, pairwise interactions are assigned by applying either a "spoke model" or "matrix model" to each complex (Figure 1.4). In a spoke model, only interactions of a "bait" and each "prey" are applied to pairwise interactions, while in a matrix model, all possible interactions among a "bait" and all "preys" are applied to pairwise interactions.

Here, we would think which model (spoke model or matrix model) should be adopted in this study using Figure 1.5. In the case of i) in Figure 1.5, proteins A~F construct a single protein complex, while in the case of ii), proteins construct multiple protein complexes (proteins A, B, C, D complex and proteins A, E, F complex). In the case of adopting matrix model, not only case i) but also case ii) results in obtaining complete graph in that a "bait" protein A interacts with "preys" proteins B $\sim$ F. Conversely, in the case of adopting spoke model, each two case i) and ii) results in obtaining specific graph by integration of pairwise interactions of each "bait" proteins A~F and corresponding "prey" proteins. That is graphs after integration of applying spoke model of affinity purification-MS data for each "bait" protein in protein complexes can represent specific topologies depending on the separation patterns of protein complexes. Therefore, in the study of this dissertation, I adopted spoke model.

The following issue, in example of Figure 1.5, is how to determine the members of protein complexes from respectively specific topologies of graph. In the case of $\mathbf{i}$, a single complex composed of proteins A~F should be predicted, while in the case of ii), multiple complexes composed of proteins A, B, C, D and proteins A, E, F should be predicted from the topology of graph. Figure $\mathbf{1 . 5}$ demonstrates clusters of proteins densely interacting each other would correspond to protein complexes. In next section, clustering of nodes focusing the density in the graph theory would be described, in order to predict members of protein complexes from these topologies depending on the separation patterns of protein complexes.


Figure 1.4 PPIs models generated from affinity purification-MS data

Two models (spoke model and matrix model) can be generally applied when pairwise interactions are assigned from affinity purification-MS data.


Figure 1.5 Advantages of adopting spoke model for affinity purification-MS data

Graphs after integration of applying spoke model of affinity purification-MS data for each "bait" protein in protein complexes can represent specific topological structure depending on the separation pattern of protein complexes.

### 1.4.2. Graph clustering software, DPClus

A PPI network can be represented as a graph in that nodes indicate proteins and edges indicate interactions. The discussion of foregoing section suggest that the densely connected clusters in a protein interaction network generated from affinity purification-MS data should be predicted as protein complexes. DPClus is a graph clustering software that can extract densely connected clusters in a network using an algorithm that is based on density and periphery tracking of clusters (Altaf-Ul-Amin et al. 2006). While using DPClus, it is necessary to provide a value of minimum density for the generated clusters (d), a minimum value for cluster property for the nature of periphery tracking ( $c p$ ), and a minimum number of nodes in a cluster.

Here, in an undirected simple graph $\mathrm{G}=(N, E)$ that consists of an finite set of nodes $N$ and a finite set of edges $E$, density $d_{k}$ and cluster property $c p_{n k}$ are respectively defined below.

The density $d_{k}$ of any cluster $k$ is the ratio of the number of edges present in the cluster $\left(\left|E_{k}\right|\right)$ and maximum possible number of edges in the cluster $\left(\left|E_{k}\right|\right)$ and the maximum possible number of edges in cluster $\left(\left|E_{k}\right| \max \right)$, and is represented by (1). $\left|N_{k}\right|$ is the size of the cluster, i.e. the number of nodes in the cluster. The density of a cluster is a real number ranging from 0 to 1 .

$$
\begin{equation*}
d_{k}=\frac{\left|E_{k}\right|}{\left|E_{k}\right|_{\max }}=\frac{2 \times\left|E_{k}\right|}{\left|N_{k}\right| \times\left(\left|N_{k}\right|-1\right)} \tag{1}
\end{equation*}
$$

The cluster property $c p_{n k}$ of any node n with respect to any cluster $k$ of density $d_{k}$ and $\operatorname{size}\left|N_{k}\right|$ is defined by (2). $\left|E_{n k}\right|$ is the total number of edges between the node $n$ and each of the
nodes of cluster $k$. The cluster property is a real number ranging from 0 to 1 .

$$
\begin{equation*}
c p_{n k}=\frac{\left|E_{n k}\right|}{d_{k} \times\left|N_{k}\right|} \tag{2}
\end{equation*}
$$

In the cluster formation process, the cluster starts as a single node which is selected as cluster seed in the target graph, and then grows gradually by adding nodes one by one from its priority neighbors. The neighbors of a cluster are the nodes connected to any node of the cluster but not part of the cluster. In adding nodes of the cluster formation, the density $d_{k}$ and the cluster property $c p_{n k}$ are checked to determine that the node should be added or not to a cluster.

### 1.5.The goal of this dissertation

The goal of this dissertation based on above background is to predict certain protein complexes, more specifically i) to predict members of protein complexes and to evaluate that by annotation and ii) to propose a method to predict the topological structure of protein complexes. Analysis on human PPI network is presented in Chapter 2, which predicts certain new protein complexes in the PPI network. Protein complexes are predicted by finding densely connected regions with their cluster properties in the network, and these predicted complexes are annotated and evaluated using integrated data such as literatures and research papers, ternary structures, descriptions of protein, localizations, expression profiles etc. The study on Arabidopsis PPI network of Chapter 3 proposes a method to predict the topological structure of protein complexes by using domain-domain interactions (DDIs).

## Chapter 2

# Prediction and evaluation of human protein complexes from the integrated protein interaction network 

### 2.1.Abstract

Background: Comprehensive analysis of PPIs provides a valuable framework for understanding the protein functions required for various biological processes in cells, and in recent years, a number of well-organized public PPI databases have become available. However, overlap ratio of PPI data entities across databases is relatively low, so that we only have a partial map of the whole interactome space from each PPI databases, and it is essential to integrate PPI data in order to fill in as many holes in the interactome space as possible. And, the next important issue for the elucidation of the functional organization of the proteome is the extraction of information about protein complex formation and function from the PPI network.

Results: Here, I report the first attempt of comprehensive annotation of human protein complexes identified from the network of integrated protein-protein interactions (PPIs). I predicted 1,264 human protein complexes from protein interaction network emerged from 32,198 PPIs composed of 9,268 proteins by integrating PPI data from six databases. In addition, the predicted 1,264 protein complexes were evaluated by annotating with reference
papers. Out of these annotated complexes, 136 complexes were perfect matches to experimentally defined complexes, 405 complexes had at least two proteins common to the experimentally defined complexes, and 723 hypothetical complexes were predicted as protein complexes but had no evidence of paper supporting to construct protein complexes. To assess the quality of our protein complex annotation, I estimated the enrichment and the ratio of consistency of Gene Ontology (GO) terms among members of a complex. Next, I compared the expression profiles of the protein members in each complex. Furthermore, the ratio of duplicated genes in a complex was evaluated. Finally, I found 78 hypothetical proteins that were annotated as members of 82 complexes that included several well-known complexes.

Conclusion: I predicted and annotated 1,264 human protein complexes. GO analysis increased the reliability of both complex prediction and manual annotation. The analysis of expression profiles and duplicated genes made it clear that protein members in larger complexes tend to express similarly and are mutually paralogous in a complex. Comprehensive protein complex prediction and annotation will provide strong functional clues about hypothetical proteins. I constructed a new database "PPI view" to provide this large annotation of human protein complexes. PPI view is available at http://hinvitational.jp/hinv/ppi/

### 2.2. Introduction

Proteins interact with other proteins or biomolecules to perform their functions, and protein complexes are the fundamental functional units of these macromolecular systems.

Comprehensive analysis of PPIs provides a valuable framework for understanding the protein functions required for various biological processes in cells. Moreover, it can offer annotation clues for proteins with unknown function (Hishigaki et al. 2001; Kemmeren et al. 2002; Titz et al. 2004). The next important issue for the elucidation of the functional organization of the proteome is the extraction of information about protein complex formation and function from the PPI network.

In recent years, a number of well-organized public PPI databases have become available such as BIND (Biomolecular Interaction Network Database) (Bader et al. 2000; 2003), DIP (Database of Interacting Proteins) (Xenarios et al. 2002), MINT (Molecular INTeraction database) (Zanzoni et al. 2002; Chatr-aryamontri et al. 2007), HPRD (Human Protein Reference Database) (Peri et al. 2003), IntAct (Hermjakob et al. 2004), and GNP (Genome Network Project Y2H data; http://genomenetwork.nig.ac.jp/).

However, in the current PPI data, main focuses are on protein-binding partners or binary protein interactions. Knowledge about how gene products form a complex, interactions among complexes, or protein interconnectivity in a complex are still scarce. In addition, overlap ratio of PPI data entities across databases is relatively low, so that we only have a partial map of the whole interactome space, thus limiting the broader application of systems modeling. Therefore, it is essential to integrate PPI data in order to fill in as many holes in the interactome space as possible.

In this study, I integrated human PPI data from the databases. I then predicted human
protein complexes from the integrated PPI data set by finding densely connected regions with cluster properties in the PPI network based on the graph theory that were previously reported by Altaf-Ul-Amin et al. (2006). After prediction, these predicted protein complexes were manually annotated by experienced annotators according to our standardized procedures utilizing literature mining and the wealth of annotation data from the human full-length cDNA database "H-Invitational Database (H-InvDB)", which have been developed by Imanishi et al. (2004) and Yamasaki et al. (2008). By using the data from H-InvDB, several analyses of the annotated complexes were performed, which should increase the validity of our annotation. This is the first attempt of comprehensive annotation of human protein complexes identified from the network of PPIs.

### 2.3. Materials and Methods

### 2.3.1. Integration of PPI data onto the H-InvDB proteins

We used XML and flat files from PPI databases; BIND, DIP, MINT, HPRD, IntAct, and GNP on October 25, 2007. These databases excepting GNP stored experimentally determined PPIs from many organisms through literature curation, but GNP had original Y2H experimental data on humans, and any computationally predicted PPIs were excluded from this study. A standardized interaction data model is needed to store PPI data from different sources. With reference to the method described in the Atlas biological data warehouse (Shah et al. 2005), I designed data loading applications for each PPI database and a relational data storage
system compliant with the Proteomics Standards Initiative Molecular Interaction Standard (PSI-MI) controlled vocabulary (Hermjakob et al. 2004), which is a community standard XML format for the presentation of protein interaction data. This system allowed me to unify data from different sources. I used only human PPIs in this study but did not use cross-species PPI data such as human proteins interacting with mouse proteins or data with ambiguous taxonomic labels such as "Mammalia", commonly seen in the HPRD download file. I downloaded all protein sequences from GenBank (Benson et al. 2007), RefSeq (Pruitt et al. 2007), and UniProt (The UniProt Consortium 2007) databases by their accession numbers, removed redundancies and assigned them to H-InvDB proteins using BLASTCLUST (Altschul et al. 1997; Wheeler et al. 2007) with a threshold of $98 \%$ sequence similarity in $95 \%$ alignment length coverage for both sequences.

### 2.3.2. Parameter optimizations of DPClus in prediction of protein complexes

In a PPI network, nodes represent proteins and edges represent interactions. DPClus can detect clusters of a network that are separated by sparse regions, keeping track of the periphery of a cluster by monitoring cluster properties of a neighbour with respect to a cluster, i.e., the program runs with two parameters; "the network density" and "the cluster property". To evaluate optimal values of these two parameters to predict protein complexes, I used a set of experimentally determined protein complexes (reference complex set). 89 protein complexes were manually collected from scientific literature and 55 complexes were got from three-dimensional structures of human protein complexes recorded in the PDB (Sussman et al.
1998). Parameter optimization was performed to select the two best parameters on the basis of which the predicted set could be best matched with the reference complex set. DPClus was run using all possible combinations of the two parameters (network density, cluster property, $0.0-1.0$ in 0.1 increments). In the parameter optimization process, DPClus was limited to find complex sizes of three or more. In this case, a predicted complex needs at least two proteins in common with a known complex to be considered a match. I found two scores; the sum of recalls, which is a ratio of the number of matched proteins of a known complex to those of a predicted complex, and the sum of precisions, which is a ratio of the number of matched proteins of a predicted complex to those of a known complex, for each parameter set. When proteins of a known complex matched fewer than two proteins of a predicted complex, recall and precision would be zero. When proteins of a known complex matched perfectly to proteins of a predicted complex, recall and precision would be one. To avoid over prediction of duplicated complexes, which shared several proteins and matched an identical known complex, the best recall and precision scores were divided by their number. The best parameter choice for DPClus for the reference complex set is to find those parameters that generate a good combination of the two scores (the sum of recalls and the sum of precisions) that predicts a complex set matching with the largest protein members of the reference complex set. Precision-Recall curve (Wicker et al. 2010; Sankararaman et al. 2010; Kastrin et al. 2010; Qiu et al. 2010) of the sum of recalls versus the sum of precisions for each of the 100 parameter combinations tested during parameter optimization was examined (Figure 2.1). A parameter combination closest to the upper right portion of the curve maximizes overlaps of predicted complexes and known
complexes. A parameter combination with network density 0.6 and cluster property 0.5 optimizes DPClus on this data set. This parameter combination led DPClus to predict 1,264 complexes, which matched 92 of the 144 known complexes. The average recall and precision of these 92 matched complexes were 0.54 and 0.66 , respectively. I also calculated average number of complexes that share a common protein. A protein was found out in 1.24 complexes on an average from the reference complex set. So according to the optimized parameters, the result was the same as that by the predicted set.


Figure 2.1 Precision-Recall curve for each of the 100 parameter combinations tested during parameter optimization.

The $y$-axis indicates the sum of recalls, which is a ratio of the number of matched proteins of a known complex to those of a predicted complex, the $x$-axis indicates and the sum of precisions, which is a ratio of the number of matched proteins of a predicted complex to those of a known complex.

### 2.3.3. Manual annotation of the predicted protein complexes

We annotated the predicted complexes according to our defined procedures, and curators confirmed the supporting evidence regarding proteins of the predicted complexes to be the experimentally defined complex members or subunits from the literature in advance and then integrated data entities such as complex names, descriptions, localizations, complex-complex interactions (CCIs), and their subunits' functions, structures, expression profiles, gene loci, and PPIs among protein subunits. I defined three categories of proteins in a predicted complex; category I: a protein of a known complex reported in the literature, category II: a protein which has a related function to a complex, and category III: a protein which has not been reported in the literature but is predicted as a complex member from the PPI network. I defined a reliability index "CQI" for a whole annotated complex based on this protein categorization. CQI is written as [Number of category I proteins].[category II proteins].[category III proteins] /[Total number of proteins in a predicted complex].

I gave names to the predicted complexes in which I used scientific names from the literature, if the majority of proteins in a complex are common to a known complex and a name is available for it (e.g., exosome, spliceosome); otherwise, I used artificial descriptions using concatenated gene symbols [e.g., GLI1-STK36-SUFU complex, DBNL-ITK-PLCG1-SH3BP2 containing complex (using "containing complex" when all symbols of proteins were not available)]. Descriptions of complexes were quoted from references with their PubMed IDs. Functional categories and subcellular localizations were added if the descriptions were available from the literature.

### 2.3.4. Consistency of GO terms assigned to the members in a complex

All GO terms of "biological process", "cellular component" and "molecular function" assigned to the H-InvDB 5.0 transcripts were used for this study. The depth of GO terms from the root in the GO hierarchy was set to five. Hence, the GO terms representing nodes with depth less than five were ignored in the analysis. If the GO term assigned to the transcript had depth greater than five, the corresponding parental node with depth five was re-assigned and redundancy was removed. As a control set representing an entire proteome, GO terms assigned to all 36,073 representative transcripts in H-InvDB 5.0 were collected. All protein members in 1,264 complexes are used as one set of protein complexes (PCset1) for the assessment. To construct the manually curated set of protein complexes (PCset2), I collected only category I proteins from perfectly matched and partially matched complexes and discarded those category II or III proteins which have not been described as members of a complex in the literature. PCset2 contains 541 complexes.

Significance of enrichment of a given GO term in a complex was tested by one-sided Fisher's exact test using $2 \times 2$ contingency table (A, B, C, D). "A" represents the number of members expressing the given GO term, and " B " is the number of members not having the GO term in the protein complex. "C" and "D" represent the corresponding numbers estimated for the entire proteome.

To estimate the quality of protein complex annotation, another quality index, "GO consistency index", was defined. The GO consistency index for the given protein complex is
estimated by the following equation; GO consistency index $=\mathrm{N}_{\text {con }} / \mathrm{N}_{\text {all, }}$ where $\mathrm{N}_{\text {cons }}$ is the number of edges that connect two proteins sharing the same GO term, and $\mathrm{N}_{\text {all }}$ is the number of possible combinations (edges) for all members of the complex.

### 2.3.5. Similarity of gene expression profiles among proteins in the same complex

In the downloaded file of H-ANGEL (Tanino et al. 2005) ("H-ANGEL_matrix.txt", December, 2007 version), I used gene expression data measured by the iAFLP method (Kawamoto et al. 1999) for 10 tissue categories. As large number of iAFLP tags were used, for some loci, multiple iAFLP-tags correspond to the same locus; then, the different expression profiles for a single locus were averaged over the tags. Expression profile of a gene was expressed by a vector of 10 elements. The similarity of gene expression profiles between two loci was calculated using the cosine of the two vectors. The similarity of multiple gene expression profiles for members of a protein complex were defined by the averaged cosines of all combinations of all the different members. The cosines of a complex were evaluated by simulation described below. For every number $(k)$ of members in the complex, $k$-genes from genes that have expression profiles were randomly picked up. Then, the averages of the cosines of the expression profiles were calculated. The procedure was repeated 100,000 times for every number of members $(k)$, and used them for p -value estimation.

### 2.3.6. Judgement of the duplicated gene pair in a given protein complex

For all combination of members in a protein complex, it was judged whether the genes were paralogous, i.e., the two genes were copied by segmental duplication, using the method established by Gu et al (2002). The gene models, which were mapped onto the "random" or "haplotype" contigs, were not used in the analysis. FASTA package version 34 t 25 (Lipman et al. 1985) was used for the analysis. In addition, I also conducted another paralog analysis with BLASTP calculation using less stringent criteria for the judgment of duplicated genes. BLAST version 2.2.17 was used. If the gene pair has a similarity with E-value less than $1 \mathrm{E}-05$, I judged the gene pair as paralogous.

### 2.4.List of abbreviations used

BIND (Biomolecular Interaction Network Database), BLAST (Basic Local Alignment Search Tool), CAGE (Cap Analysis of Gene Expression), CCI (Complex-Complex Interaction), cDNA (Complementary DNA), CQI (Complex annotation Quality check Index), DIP (Database of Interacting Proteins), EST (Expressed Sequence Tag), FDR (False Discovery Rate), GNP (Genome Network Project), GO (Gene Ontology), H-ANGEL (Human Anatomic Gene Expression Library), H-InvDB (H-Invitational Database), HPRD (Human Protein Reference Database), iAFLP (introduced Amplified Fragment Length Polymorphism), MINT (Molecular INTeraction database), ORF (Open Reading Frame), PDB (Protein Data Bank), PPI (Protein-Protein Interaction), PSI-MI (Proteomics Standards Initiative Molecular Interaction

Standard), RFC (Replication Factor C), ROC (Receiver Operating Characteristic), XML (Extensible Markup Language).

### 2.5.Results

### 2.5.1. Integration of PPI data onto the H-InvDB proteins

Imanishi et al. and Yamasaki et al. developed a comprehensive database of human genes and transcripts, called H-InvDB (http://www.h-invitational.jp/) (Imanishi et al. 2004; Yamasaki et al. 2008). It is a unique database that integrates annotation of sequences, structure, function, expression, subcellular localization, evolution, and the diversity of the human genes and the encoded proteins into a single entity. It is useful as a platform for conducting in silico data mining. International collaboration for analysis of high-quality full-length cDNA clones, in addition to the EST assemblies and CAGE tags, now provides integrative annotation of 187,156 transcripts placed on 36,073 loci (Release 5.0). Based on the open reading frame (ORF) prediction of H-InvDB 5.0 transcript sequences, followed by the functional annotation of experienced annotators, 108,530 non-redundant human protein candidates (H-InvDB proteins) were determined. To survey human PPIs from the landscape of the human interactome, the PPI information was mapped onto the H -InvDB proteins.

A number of public PPI databases are currently available. I selected BIND (Bader et al. 2000; 2003), DIP (Xenarios et al. 2002), MINT (Zanzoni et al. 2002; Chatr-aryamontri et al 2007), HPRD (Peri et al. 2003), IntAct (Hermjakob et al. 2004) and GNP as major resources for

PPI integration. I removed PPI data redundancies by sequence similarity and then integrated human PPIs with the H-InvDB proteins (see also 2.3.1). As a result, I obtained 32,198 human PPIs composed of 9,268 proteins.

Figure 2.2 shows the overlap of human PPIs across the six databases. There are 6,234 non-redundant human PPIs in BIND, while DIP, MINT, HPRD, IntAct, and GNP have 1,037, 12,055, 2,913, 19,213, and 1,303 PPIs, respectively. Figure 2.2A shows pairwise overlaps of integrated PPIs across the databases; MINT and IntAct share 6,089 PPIs, which is the highest overlap among these databases. As shown in Figure 2.2B, 6,671, 1,786, 102, and 2 PPIs are shared in $2,3,4$, and 5 databases, respectively, but there are no PPI in common among all the 6 databases. There are 23,637 unique PPIs in the databases, occupying 73\% of the PPI dataset. The overlap across these databases was relatively small, which reflects the fact that the actual size of the human interactome will probably be much larger than that of the currently known PPIs (von Mering et al. 2002; Bader and Hogue 2002; Kumar and Snyder 2002). Thus, it is essential to integrate the PPI data to achieve a complete view of the human interactome.

## A

| BIND |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| DIP | 260 |  |  |  |  |  |
| MINT | 1,195 | 353 |  |  |  |  |
| HPRD | 150 | 12 | 1,385 |  |  |  |
| IntAct | 1,158 | 317 | 6,089 | 1,695 |  |  |
| GNP | 17 | 7 | 12 | 2 | 9 |  |
|  | $\begin{gathered} \hline \text { BIND } \\ (6,234) \\ \hline \end{gathered}$ | $\begin{gathered} \hline \text { DIP } \\ (1,037) \\ \hline \end{gathered}$ | $\begin{array}{c\|} \hline \text { MINT } \\ (12,055) \end{array}$ | $\begin{aligned} & \hline \text { HPRD } \\ & (2,913) \\ & \hline \end{aligned}$ | $\begin{gathered} \hline \text { IntAct } \\ (19,213) \\ \hline \end{gathered}$ | $\begin{gathered} \hline \text { GNP } \\ (1,303) \\ \hline \end{gathered}$ |

## B



Figure 2.2 Overlap of human PPIs in six PPI databases.
A) Pairwise overlaps of integrated PPIs across databases are shown in cells. The number of integrated PPIs is shown in parentheses for each database. B) Overlaps of PPIs sharing in common in single, two, three, four, five, and six databases are shown.

### 2.5.2. Prediction of protein complexes

View of integrated PPI network composed of 32,198 human PPIs was represented in Figure 2.3. The next important issue for the elucidation of the functional organization of the proteome is the extraction of information about protein complex formation and function from the PPI network.

In the past, the algorithm called "DPClus" was developed by Altaf-Ul-Amin et al. (2006), which extracts the densely connected regions in a network and demonstrated that many of these densely connected regions correspond to known protein complexes or protein functional units (Bader and Hogue 2003; Altaf-Ul-Amin et al. 2006). DPClus is a robust algorithm not affected by a high rate of false positives in data from high-throughput interaction-detection techniques (Altaf-Ul-Amin et al. 2006).

On prediction of protein complexes by DPClus, I adopted the "overlapping clustering mode", which allows identical proteins to be classified into different clusters, because it is biologically well established that proteins can be present in multiple complexes at different times and locations. For example, POLR2E/RPB5 (HIP000039507), POLR2F/RPB6 (HIP000096671), POLR2H/RPB8 (HIP000027404), POLR2K/RPB12 (HIP000043404), and POLR2L/RPB10 (HIP000064404) are conserved throughout RNA polymerases I, II, and III (Werner et al. 2007). Before complex prediction, I evaluated the optimal values of DPClus parameters by comparing the predicted complex set with the experimentally determined 144 reference complex set (see also 2.3.2).

As the result, 1,264 protein complexes were predicted in the PPI network composed of 32,198 human PPIs using DPClus with the best parameter set (network density 0.6 , cluster property 0.5 ) determined on the basis of the reference complex set.


Figure 2.3 View of Integrated PPI network.

Nodes and edges represent proteins and interactions respectively. How do we understand PPI network and biological process? The next important issue for the elucidation of the functional organization of the proteome is the extraction of information about protein complex formation and function from the PPI network.

### 2.5.3. Manual annotation of the predicted protein complexes

An annotation of the predicted complexes was performed with the following procedures; 1) curators searched for supporting evidence in the scientific literature, which could confirm that the proteins of the predicted complexes are experimentally defined complex members or subunits, 2) missing proteins were manually added to the predicted complexes if there was evidence in the literature that they were members of complexes but had been dropped in the complex prediction, 3) I integrated data such as complex names, descriptions, localizations, complex-complex interactions (CCIs), and their subunits' functions, structures, expression profiles, gene loci, and PPIs among protein subunits. I did not remove proteins that were predicted to be complex subunits but for which no evidence was found, because they are complex subunit candidates.

The protein members of the predicted complexes were categorized into three categories based on their annotation level: category I, proteins that are confirmed as subunits of a known complex in the literature or as ternary structures in the PDB (Sussman et al. 1998); category II, proteins for which no evidence of a complex membership exists in the literature, but have functions related to those of other category I members in the predicted complex according to their protein definitions or Gene Ontology (GO) terms (Ashburner et al. 2000); and category III, proteins that are predicted as a complex subunits by DPClus and do not overlap with functions of the other two categories, i.e., do not have functions related to those of other category I members in a predicted complex according to their protein definitions or Gene Ontology (GO) terms. Our protein complex prediction allowed the same proteins to be subunits
of different complexes; therefore, such shared proteins could be classified into the other categories as members of different complexes (see also 2.3.3).

In total, 1,264 protein complexes were predicted and annotated. A list of the all annotated complexes is available at http://h-invitational.jp/hinv/ppi/. As the result of annotation, 2,106 proteins, 299 proteins and 3,273 proteins were respectively classified into category I, category II, category III when protein subunit sharing in complexes was allowed (Table 2.1A). The average numbers of protein subunits per complex were 3.9 (category I proteins only), 4.3 (category I and II proteins), and 4.5 (all categories) respectively, and a large number of smaller complexes than average size (i.e., numbers of protein subunits per complex) and a small number of larger complexes than average size were observed in that the relationship between the size of a complex and its frequency follows an inverse power law (Figure 2.4A).

In this annotation, 136 predicted complexes matched perfectly with known complexes (all subunits are in category I) and 405 complexes had at least two subunits that matched with a known complex (at least two subunits are in category I), while 723 complexes had only predictive support (all subunits are in category III) (Table 2.1B). I defined three types of predicted complexes: perfectly matched complex, partially matched complex, and hypothetical complex; these correspond, respectively, to a complex with all subunits in category I, a complex with at least two proteins in category I, and a complex with all subunits in category III.

From information in the scientific literature, functional categories and subcellular
localization were assigned to the annotated complexes (Figure 2.4 C, D). The major functional categories were signal transduction (90 complexes, 19\%), transcription ( $61,14 \%$ ), cell cycle $(52,12 \%)$, and immune response $(49,11 \%)$; over $70 \%$ of the complexes are localized in the cell nucleus $(160,33 \%)$, membranes $(111,22 \%)$, and cytoplasm $(81,16 \%)$.

Table 2.1 Protein and the complex annotation summary.
A) Number of proteins in H-InvDB, the integrated PPI data set, the predicted complexes. The categorized proteins in the predicted complexes are described in the text. Since, complex sharing proteins could be classified into the other categories as members of different complexes, total number of the categorized proteins is more than number of proteins in the predicted complexes. B) Type of the predicted complexes. Three types of predicted complexes were defined by matching degree to known complexes (details are in the text). Total number of the predicted complexes in this study is 1,264 .

## A

|  | Number of proteins |
| :--- | ---: |
| H-InvDB proteins | 108,530 |
| Proteins in the PPI data set | 9,268 |
| Proteins in the predicted complexes | 4,513 |
| Category I proteins in the predicted complexes | 2,106 |
| Category II proteins in the predicted complexes | 299 |
| Category III proteins in the predicted complexes | 3,273 |

## B

| Type of the predicted complexes | Number of complexes |
| :--- | ---: |
| Perfectly matched complex | 136 |
| Partially matched complex | 405 |
| Hypothetical complex | 723 |
| Total | $\mathbf{1 , 2 6 4}$ |



Figure 2.4 Properties of annotated complexes
A) The relationship between complex size (number of different protein subunits of each category) and frequency. B) Percentage of category I and II protein occupancy of the annotated complexes. C) Distributions of functional categories. D) Distributions of subcellular localizations.

### 2.5.4. Consistency of GO terms assigned to members in a complex

Since proteins in a complex cooperatively exert a biological role, it is expected that they exist in the same location at a particular time and that they are players in the same biological process or pathway. To assess the quality of our protein complex annotation, the enrichment and the ratio of consistency of GO terms among members of a complex were estimated. This assessment is based on the assumption that the same GO terms are assigned to proteins in a single protein complex.

First, the enrichment of some GO terms in a complex compared to GO terms assigned to the proteome were estimated. The proteome set comprised 36,073 representative transcripts in H-InvDB 5.0. The enrichment of GO terms was estimated against two sets of protein complexes: PCset1 represents all 1,264 protein complexes comprising of all member proteins including category III proteins which is unknown for the complex-related function but predicted as complex subunits by DPClus based on PPI interconnectivity, and PCset2 represents 541 complexes whose subunits are limited to contain manually curated category I proteins obtained from perfectly matched and partially matched complexes. As a result, 450 protein complexes out of 1,264 PCset1 ( $35.6 \%$ ) complexes have at least one significantly enriched GO term (Fisher's exact test, p-value $\leq 0.01$ ). On the other hand, 254 protein complexes out of the 541 PCset 2 complexes ( $47 \%$ ) have at least one significantly enriched GO term. The ratio of protein complexes having enriched GO terms is greater in PCset2 than in PCset1, suggesting that the reliability of protein complex annotation was significantly refined
by manual check.

Second, the degree of consistency of GO terms among members in a complex; i.e., homogeneity of GO terms assigned for complex members were estimated. A consistency index (see 2.3.4) was used as an indicator of homogeneity. In order to estimate the degree of GO term consistency expected by chance, 100 sets of randomly selected genes from H-InvDB 5.0, all representative transcripts with complex sizes matched to our annotation of PCset1, were created and used as a control. Average consistency indexes were estimated to be $0.23,0.41$, and 0.04 for protein complexes of PCset1, PCset2, and the random set, respectively. The value is significantly higher in PCset 1 (Student's t test, p -value $2.9 \mathrm{E}-111$ ) than in the random set, and in PCset2 than in PCset1 (Student's t test, p -value $1.6 \mathrm{E}-25$ ). These results are still statistically significant after Bonferoni multiple test, which is relatively conservative. The histogram of consistency indexes for the three sets is shown in Figure 2.5. In particular, cases in which the consistency index equals 1.0 (i.e., all members share common GO terms with other members), increased dramatically after manual curation, indicating the relatively high quality of manual annotation and the advantage of protein complex prediction followed by manual annotation as opposed to only single computational prediction.

Intriguingly, 28 PCset 1 unique complexes with consistency index equal to 1.0 were found. Although the existence of the protein complexes has not yet been validated experimentally, the compatibility between the prediction of protein complexes by our clustering method and the consistency of GO terms offers us reliable candidates as novel functional
protein complexes to be validated by future experiments.


Figure 2.5 Distributions of GO consistency index in PCset1, PCset2, and random set.

Histogram of GO consistency index for protein complexes in PCset1, PCset2, and random set, shows a shift toward larger values in PCset1 and PCset 2 than those in the random set.

### 2.5.5. Similarity of gene expression profiles among proteins in the same complexes

Based on the idea that co-expressed genes are more likely to have similar functions, cluster analysis of gene expression data have been utilized to predict the functions of non-annotated proteins (Devos et al. 2000; Rost 2002). In reverse, I examined as to whether proteins in the same complex (involved in the same functions) have similar expression profiles. For each complex, the expression profiles of protein members in the complex were compared. If the members of a complex are similar in their expression profiles, the profile should indicate some functional information about the complex whose function is unknown.

Expression profiles of 729 complexes were obtained from the Human Anatomic Gene Expression Library (H-ANGEL) (Tanino et al. 2005), the satellite database of H-InvDB. Among 729 complexes, six complexes were found to have significant gene expression similarlity by a false discovery rate (FDR) criterion of 0.05 . FDR, the expected proportion of incorrectly rejected null hypotheses, is a widely used statistic for multiple tests (Kim and van de Wiel 2008). The six complexes are shown in Table 2.2. Some of the most interesting complexes in which the expression of the protein members was similar and tissue specific were found. The fibrinogen complex (Complex No. 130) was one of complexes in such a case, and in which the expression of the protein members was liver specific. Other examples were the AK5-CPNE6-TRIM46 complex (Complex No. 540) and the troponin complex (Complex No. 258). Though FDR of the two complexes was not significant, 0.22 and 0.68 , respectively, the gene expression profiles were very similar (the cosines were 0.99 and 0.95 ). For troponin, the
gene expression of the members is specific to muscle/heart tissue (average entropy, 1.12). The expression profiles of the troponin complex are shown in Figure 2.6. The similarity of their expression profiles suggests that they may function as a complex.

As shown above, the gene expression of the protein members was not significantly similar in most of the predicted protein complexes. However, instead I found that, for large complexes, the gene expression of complex members is more likely to be similar. For each complex, I calculated p-values of gene expression similarities (see also 2.3.5), then plotted the distribution of $p$-values for different number of proteins in a complex (Figure 2.7). The figure illustrates that gene expression of proteins in the same complex is more similar as the number of protein members (complex size) increases. This would be the first report of a relationship between expression similarity and complex size in human PPI and is consistent with the results for Yeast (Liu et al. 2009).

Table 2.2 The protein complexes consisting of members with significantly similar gene expression profiles.

| Complex <br> No. | CQI | Complex name | cosine | FDR | \# of genes |
| :--- | :--- | :--- | ---: | ---: | ---: |
| 30 | $21.1 .0 / 22$ | 19S proteasome of the 26S proteasome | 0.92 | 0.001 | 13 |
| 12 | $18.0 .4 / 22$ | 20S proteasome of the 26S proteasome | 0.88 | 0.006 | 17 |
| 41 | $12.1 .0 / 13$ | RNA polymerase II complex | 0.92 | 0.008 | 10 |
| 68 | $0.0 .11 / 11$ | COP9 signalosome (CSN) | 0.92 | 0.014 | 9 |
| 130 | $3.0 .8 / 11$ | Fibrinogen | 0.96 | 0.037 | 4 |
| 77 | $4.0 .8 / 12$ | 18S U11/U12 complex | 0.89 | 0.041 | 14 |



Figure 2.6 Relative percentage of gene expression levels of troponin complex.

The three gene loci of the troponin complex (Complex No. 258) member proteins are expressed specifically in muscle/heart tissue.


Figure 2.7 Boxplot of gene expression profile similarity and the number of protein members in a complex.

The $y$-axis indicates gene expression similarity (negative logarithm of $p$-value of average cosine of gene expression profiles) in a complex; a higher $y$-value means that the members of the complex have similar gene expression profiles. The x -axis indicates the number of protein members with expression data in the complex. The gene expression profiles are more similar as the number of proteins increases.

### 2.5.6. Relationship between the formation of protein complexes and gene duplication

To reveal how gene duplication contributes to the formation of protein complexes, the ratio of duplicated genes in a complex were examined.

The paralog judgment established by Gu et al. (2002) yielded 2,353 duplicated genes in a total of 4,191 genes that were the components of 1,264 complexes (see 2.3.6). Of the 1,264 complexes, 336 (26.5\%) were judged to have at least one paralog pair. Moreover, 218 complexes (17.2\%) were obtained in which more than half of the components were judged to be paralogous to another gene in the same complex. By using a less stringent method with BLASTP (E-value $\leq 1 \mathrm{e}-05$ ), these percentages were estimated to be $38.5 \%$ and $27.3 \%$, respectively.

The replication factor C (RFC) complex (Complex No. 105) is a good example of the formation of a protein complex induced by gene duplication. This complex consists of five RFC subunits and one binding partner, PCNA (Cai et al. 1996). The complex is known to be related to DNA synthesis (Cai et al. 1996), and the function and machinery are conserved between yeast and human (O'Donnell et al. 1993), indicating that this is an ancient protein complex. Paralog judgment by the Gu method suggested that three (RFC 36, 37, 40) out of five RFC subunits are paralogous, i.e., they originated from a common ancestor, whereas the result obtained by the less stringent BLASTP method suggested that all five subunits are mutually paralogous. The existence of the "RFC box" motif in all five proteins and the consistency of exon-intron boundaries also support the homologous relationships of these five subunits. These
results indicate that the enlargement of a protein complex is mainly mediated by homologous interactions, and gene duplication events greatly contribute to the establishment of protein complexes.

### 2.5.7. Functional assignments for hypothetical proteins in the annotated complexes

An important goal of proteomics is a functional assignment for proteins that cannot be annotated by homology alone. Several approaches for functional assignment from PPIs have been developed (Hishigaki et al. 2001; Kemmeren et al. 2002; Titz et al. 2004).

H-InvDB proteins were analyzed with standardized functional annotation by curators who classified the proteins into several categories; i) identical to known human proteins, ii) similar to known proteins (having $50 \%$ sequence similarity), iii) InterPro domain containing proteins, and iv) hypothetical proteins (no biological functions were inferred).

We found 78 hypothetical proteins in the 82 predicted complexes. Although the majority of hypothetical proteins ( 61 proteins, $78.2 \%$ ) were subunits of 67 hypothetical complexes (none of their subunits were reported as complexes in the literature), 13 hypothetical proteins were subunits of 12 complexes whose functions were strongly deduced because at least half of their subunits were annotated as common to known complexes. A protein complex is thought to be a functional unit where proteins get together and perform biological functions; therefore, a hypothetical protein can be assigned a function related to that of the complex it joins. For example, two hypothetical proteins HIP000013164 and HIP000053526 were in the
"DREAM complex (Complex No. 24)", which is tightly bound to E2F-regulated promoters in G0 and dissociates from these promoters in $S$ phase of the cell cycle. In addition, some subunits of the complex can also interact specifically with MYB and may be involved in expression of MYB-dependent genes important in G2/M progression (Litovchick et al. 2007). I expected that these two hypothetical proteins would then join the DREAM complex and might play roles in the cell cycle. Moreover, I found that annotated complexes such as "Fanconi anemia (FA) core complex (Complex No. 61)", "INO80 complex (Complex No. 75)", and "Lamins complex (Complex No. 101)" include hypothetical proteins (HIP000177716 for the FA core complex, HIP000079962 for the INO80 complex, and HIP000024165 for the Lamins complex). These complexes have DNA repair, DNA repair and transcription, and nuclear organization functions, respectively. Therefore, these hypothetical proteins might also have functions associated with those complexes. Table 2.1 summarizes the 12 complexes and their CQIs, which include hypothetical proteins as subunits and at least half of whose subunits are common to known complexes.

Table 2.3 The hypothetical proteins in the annotated complexes.

12 complexes which had hypothetical proteins as the subunits and in which at least half of subunits were common to known complexes are shown.

| Complex No. | CQI | Hypothetical proteins | Name |
| :---: | :---: | :---: | :---: |
| 24 | 10.1.2/13 | HIP000013164 | dREAM complex |
| 61 | 8.0.1/9 | HIP000053526 HIP00017716 | Fanconi anemia (FA) core complex |
| 75 | 11.0.2/13 |  | INO80 complex |
| 101 | 4.0.1/5 | HIP000024165 | Lamins complex |
| 200 | 3.0.2/5 | HIP000046613 | C8orf32-EFCBP2-RUNX1T1-ZNF652 containing complex |
| 673 | 3.0.2/5 | HIP000038372 | BCL2A1-BCL2L1-BCL2L2-HRK-PMAIP1 complex |
| 922 | 4.0.3/7 | HIP000089800 | HIF-1alpha-pVHL-ElonginB-ElonginC complex |
| 940 | 2.0.1/3 | HIP000027799 | SRGAP3-WASF1 containing complex |
| 960 | 3.0.1/4 | HIP000060581 | C19orf25-KNTC1-ZW10 containing complex |
| 967 | 3.0.2/5 | HIP000015491 | NONO-PSPC1-WBP4-ZNRD1 containing complex |
| 1156 | 2.0.2/4 | HIP000114159 | NUTF2-RAN complex |
| 1310 | 4.0.3/7 | HIP000091971 | SCF (Skp1, cullin 1, F-box) ubiquitin E3 ligase complex |

### 2.5.8. Development of "PPI view" and 'PPI map"

Database of PPIs and protein complexes called "PPI view (http://hinvitational.jp/hinv/ppi/)" and visualization tool of protein complexes "PPI map" were developed. PPI view provides detailed information of PPIs and predicted and annotated protein complexes (Figure 2.8). As PPIs information, we can obtain interacting proteins with a target protein, in addition we can check 1) the experimental methods that detected the PPI, 2) the experimental scale (large scale or small scale) detecting the PPI 3) the number of papers which report the PPI. This information would be important and useful to judge the reliability of PPIs. In information of predicted and annotated protein complexes, detailed complex annotations (complex name, functions, localization, etc.) and CQIs (Complex annotation Quality check Index) are provided. CQI is a reliability index for a whole annotated complex based on the protein category. For example, CQI is a simple index of 5.2.1/8 which means that the complex has eight subunits in which five, two, and one proteins are classified into categories I, II, and III, respectively.

The visualization tool "PPI map" in PPI view can show protein interconnectivity of a complex, complex-outside protein interactions: interactions among proteins in a complex and an outside protein, and complex-complex interactions: interactions among proteins in a complex and proteins in the other complex, and has two kinds of window, "detailed network" window and "whole network" window as shown in Figure 2.9. Detailed network window shows member proteins of a target complex in the center and directly interacting proteins or protein complexes in the neighborhood. Other hand, whole network window displays whole
of proteins and protein complexes network. On PPI map, each node (circle) represents an individual complex and edges represent their interactions. A node size represents number of proteins in a complex and thicknesses of edges connecting complexes, which is exponential to the number of PPIs between connected nodes. On using PPI map, a user can change scale of visualization by slide bar in the upper right of windows, and move displaying part of proteins and complexes interaction network by pressing-and-holding the mouse button and dragging. On scale down of visualization or moving displaying part of network, to avoid over-traffic on internet and reduce the time of visualization, PPI map request just information of new nodes and new edges to visualize to the WWW server and get the information by XML through the internet (Figure 2.10). Thus, PPI map can be used to show CCIs of interest, with the ability to seamlessly scale, move, and change the represented thicknesses of edges connecting complexes. Using the "PPI map" I had drawn an overhead view of CCIs with the subcellar localizations of the annotated complexes (Figure 2.11). Node colors indicate subcellular localization of the annotated complexes. To avoid a CCI network to be too complicated to understand, 541 complexes (perfectly matched and partially matched complexes) and the complex-complex interactions or the protein-complex interactions comprising more than 10 PPIs are here shown. From this "map", we can understand the relationship of known and partially-known protein complexes.

As far as I know, PPI view is the first database that can show CCIs in the human interactome with detailed annotation.


Figure 2.8 Database of PPIs and predicted and annotated protein complexes, "PPI view"

We can search an interesting protein from "top search page (left part)" and know information of interacting proteins with the target protein in the "results page (right part)". In addition to information of interacting proteins, we can check 1) an experimental methods to detect the PPI, 2) the experiment al scale (large scale or small scale) detecting the PPI 3) the number of papers which report the PPI.


Figure 2.9 Visualization tool of protein and protein complex network, "PPI map"
"Detailed network" window (upper part) shows member proteins of a target complex in the center and directly interacting proteins or protein complexes in the neighborhood. "Whole network" window (lower part) displays whole of proteins and protein complexes interaction network. Scale of visualization can be changed by slide bar in the upper right of windows (left and right in the lower part).


Figure 2.10 System architecture of PPI map

On scale down of visualization or moving displaying part of network, to avoid over-traffic on internet and reduce the time of visualization, PPI map request just information of new nodes and new edges to visualize to the WWW server and get the information by XML through the internet.


Figure 2.11 An overhead view of complex-complex interactions (CCIs) with the subcellar
localizations of the annotated complexes.

Node colors indicate subcellular localization of the annotated complexes; dark red: nucleus, blue:
cytoplasm, green: membrane, purple: nucleus and cytoplasm, yellow: Golgi apparatus, blue-green: cytoplasm and membrane, light blue: cytoplasm, membrane and nucleus, orange: mitochondria,
light red: endoplasmic reticulum, light green: endosome, grey: other subcellular localization, black:

[^1]
### 2.6.Discussion

Here, I report the first attempt of comprehensive annotation of human protein complexes identified from the network of integrated PPIs of six well-organized databases. I predicted 1,264 human protein complexes from 32,198 PPIs comprising 9,268 proteins by finding densely connected regions with their cluster properties in the PPI network. After evaluating the predicted complexes and comparing with the manually collected 144 reference complex set, the annotation of the predicted complexes was carried out according to our original procedures. In this annotation, 136 complexes were perfect matches to the experimentally defined complexes; 405 complexes included at least two proteins common to the experimentally defined complexes; and 723 hypothetical complexes with no supporting evidence of complexes in the literature were predicted as protein complexes.

To assess the quality of protein complex annotation, I estimated the enrichment and the ratio of consistency of GO terms among members of a complex. This assessment is based on the assumption that the same GO terms are assigned to the proteins in a single protein complex. The ratios of protein complexes having enriched GO terms and the degree of GO term consistency were greater in the manually curated set of protein complexes (PCset2) than in all the predicted complexes (PCset1) or the random set, indicating the relatively high quality of manual annotation and the advantage of protein complex prediction followed by manual annotation as opposed to only single computational prediction.

Next, for each complex, I compared the expression profiles of the protein members in
the complex based on the idea that proteins in the same complex would have similar functions and that co-expressed genes are more likely to have similar functions. Members of large complexes tend to express similarly. Furthermore, the ratio of duplicated genes in a complex was evaluated. The results indicated that the enlargement of a protein complex is considerably mediated by homologous interactions, and that gene duplication events greatly contribute to the establishment of protein complexes.

Recent statistics of manually curated H-InvDB proteins showed that $43 \%$ of H -InvDB representative transcripts were hypothetical proteins (Yamasaki et al. 2008). Hence, assigning functions to hypothetical proteins of unknown function is one of the most important problems in proteome analysis. Since subunits of a complex generally tend to have the same biological function, prediction of a protein complex allows increased confidence in the annotation of hypothetical proteins. Protein complex prediction and the following annotation found that 78 hypothetical proteins were in the 82 predicted complexes. Although the majority of hypothetical proteins tended to be subunits of hypothetical complexes that I defined and whose subunits had not been reported in the literature as parts of a complex but were just predicted as complexes by DPClus, 13 hypothetical proteins were subunits of 12 function annotatable complexes (Table 2.3). These hypothetical proteins are probably involved in biological processes related to other subunits of the complexes they joined. Thus, I could expect that hypothetical proteins HIP000013164 and HIP000053526 in the DREAM complex might function in the cell cycle, and HIP000177716 (FA core complex), HIP000079962 (INO80 complex), and HIP000024165 (Lamins complex) might function in DNA repair, DNA repair
and transcription, and nuclear organization, respectively. The number of hypothetical proteins to whom functions were assigned in this study was very small compared to the total number of proteins in the proteome; however, information about protein complex prediction and annotation will offer annotation clues for hypothetical proteins.

Therefore, protein complexes determined solely on the basis of interaction data can help us predict the functions of proteins, and can also be useful in understanding and explaining certain biological processes. In this study, I could assign protein complex data for about $10 \%$ of the total human locus stored in H-InvDB (34,700 gene clusters [HIX]). Thus, as a rough approximation, I estimated that the total number of human protein complexes will be 10,000-20,000.

The new database "PPI view" was constructed to provide this large annotation of human protein complexes, URL; http://h-invitational.jp/hinv/ppi/. PPI view includes a new visualization tool "PPI map", which allows users to show not only PPIs among the protein subunits, but also complex-complex interactions (CCIs) through a seamless and detailed annotation of each protein complex and its subunits. PPI view will be a useful platform for understanding protein function from the viewpoint of a complex as another level of functional unit, and will provide annotation clues for proteins of unknown function.

### 2.7.Conclusion

1,264 human protein complexes were predicted and annotated from the integrated PPI
data. GO analysis increased the reliability of both complex prediction and manual annotation. The analysis of expression profiles and duplicated genes made it clear that protein members tend to express similarly and are mutually paralogous in a complex. Comprehensive protein complex prediction and annotation will provide strong functional annotation clues about hypothetical proteins. The new database "PPI view" was constructed to provide this large annotation of human protein complexes. PPI view is available at http://h-invitational.jp/hinv/ppi/

## Chapter 3

# Predicting topological structure of protein complexes by determining statistically significant domain-domain interactions 

### 3.1.Abstract

Background: More or less over the past 10 years, vast amount of protein-protein interaction data have been generated by high-throughput methods for detecting protein interactions, such as the affinity purification- mass spectrometry (affinity purification-MS) and the yeast two-hybrid (Y2H) system (Uetz et al. 2000; Ito et al. 2001; Ho et al. 2002; Gavin et al. 2002). However, high-throughput methods for detecting protein interactions generate a significant number of false-positives, and that is serious problem because they cause erroneous results and misleading conclusions. Additionally, the affinity purification-MS detects non-direct interactions: "prey-prey" interactions, it implies that the topological structure of the protein complexes cannot be determined from the individual experiments only.

Results: The present study proposes a method to predict the topological structure of protein complexes by using domain-domain interactions (DDIs). As the first step, I extracted 312 statistically significant DDIs out of 1,162 DDIs underlying 3,118 protein-protein interactions (PPIs). Significant DDIs were determined based on statistical analysis of $2 \times 2$
contingency tables, estimated using pairs of domains reported to interact in determined structures of protein complexes in iPfam (Finn et al. 2005) as true positive DDIs. Next, 67 protein complexes were obtained by finding densely connected regions with their cluster properties in the PPI network. Finally, I discussed the topological structure of protein complexes based on DDI information extracted in the first step.

Conclusion: This work has proposed a method to predict the topological structure of protein complexes by using domain-domain interactions (DDIs). Significant DDIs were determined based on statistical analysis of $2 \times 2$ contingency tables, 312 significant DDIs underlying 3,118 protein-protein interactions (PPIs) were obtained. In the present study, 1,629 out of 3,118 PPIs were supported by statistically significant domain-domain pairs. Furthermore, I validated 67 protein complexes detected in the protein interaction network composed of 1,302 proteins in view of significant DDIs. The present study makes it possible to interpret the topological structure of protein complexes by configuring interactions supported by DDIs.

### 3.2.Introduction

Experimental methods to detect protein-protein interactions (PPIs), such as Yeast two-hybrid method (Y2H) and affinity purification-mass spectrometry method (affinity purification-MS), generates a significant number of false-negatives and false-positives (Titz et al. 2004), and false-positives are especially a more serious problem in that they cause erroneous results and misleading conclusions. In addition to that, on affinity purification-MS, all of the
"prey" proteins in complexes containing a "bait" protein are identified as interacting proteins. It implies PPIs detected by affinity purification-MS include "bait-prey" and "prey-prey" interactions, and topological structure of the protein complexes cannot be determined from the individual experiments only. In case that proteins A, B and C are identified in isolation as a protein complex using protein A with a tag as a "bait", it is difficult to conclude what is the true topological structure of the complex out of the four possible cases as shown in i) ~iv) in Figure 3.1. Though understanding the topological structure of protein complexes is necessary in order to obtain useful information about them, little attention has been given by the bioinformatics researchers to decipher the true topological structure of the protein complexes. Also the experimental technologies used to detect PPIs do not focus on this matter.

The bioinformatics analysis of PPIs has mainly followed two different approaches after the high throughput experiments started to produce huge amount of data. One of the approaches is the analysis of protein interaction networks based on graph theory, aiming to detection of protein complexes from PPIs networks (Bader and Hogue 2003; Altaf-Ul-Amin et al. 2006). These studies have reported that the densely connected regions in a network correspond to known protein complexes or protein functional units. The other approach is the analysis of DDIs coming out from PPIs data by statistics or machine learning intended to predict unknown PPIs (Sprinzak and Margalit 2001; Riley et al. 2005; Singhal and Resat 2007; Liu et al. 2009). These studies have shown that the concept of DDIs statistically extracted from large-scale PPI data can explain the makeup of PPIs to some extent.

Taking these two recent approaches into account, the present study focuses on statistically significant DDIs i.e. the direct interactions to predict the topological structure of protein complexes by avoiding the effects of false-positives or non-direct interactions. Corresponding to Figure 3.1, if DDI analysis supports that protein B and C directly interacts with different domains of protein A then I would predict the topological structure of ii), and on the other hand, if protein B and C directly interacts with an identical domain of protein A competitively then I would predict the topological structure of iv). Further detail prediction of interactions in protein complexes were performed based on DDIs.
i)

$P_{B}(a)-P_{A}(b)$
$P_{B}(c)-P_{C}(d)$
ii)

iv)


$$
\begin{aligned}
& \mathrm{P}_{A}(a)-\mathrm{P}_{B}(b) \\
& \mathrm{P}_{A}(a)-\mathrm{P}_{C}(c)
\end{aligned}
$$

Figure 3.1 Topological structure diversity of protein complexes in case that proteins A, B and C are identified by MS.

Notations in boxes indicate example of direct protein interactions by domains, i.e., $\mathrm{P}_{\mathrm{A}}(a)-\mathrm{P}_{\mathrm{B}}(b)$ represents that the domain $a$ of protein A directly interacts the domain $b$ of protein $B$.

### 3.3.Methods

In the present study, I propose a procedure for inferring topological structure of protein complexes with reliable PPIs by using information of DDIs (Figure 3.2), which comprises three steps, (Step 1) Detection of domains in proteins, (Step 2) Extraction of statistically significant DDIs, and (Step 3) validation of protein complexes based on extracted DDIs.

## Step 1: Detection of domains in proteins

protein


Step 2: Statistically Extracting domain-domain interactions from PPIs data

|  | Domain $Y$ |  |  |
| :---: | :---: | :---: | :---: |
|  | $Y$ | $\bar{Y}$ |  |
| Domain $X$ | $X$ | $A$ | $B$ |
|  | $X$ | $C$ | $D$ |


| ID Protein | Proteinj | A | $\mathbf{B}$ | $\mathbf{C}$ | $\mathbf{D}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $1 P_{i}(X, Y)$ | $P_{j}(X, Y)$ | 2 | 0 | 0 | 0 |
| $2 P_{i}(X, Y)$ | $P_{j}(\bar{X}, Y)$ | 1 | 0 | 1 | 0 |
| $3 P_{i}(X, Y)$ | $P_{j}(X, \bar{Y})$ | 1 | 1 | 0 | 0 |
| $4 P_{i}(X, Y)$ | $P_{j}(\bar{X}, \bar{Y})$ | 0 | 1 | 1 | 0 |
| $5 P_{i}(\bar{X}, Y)$ | $P_{j}(X, Y)$ | 1 | 0 | 1 | 0 |
| $6 P_{i}(\bar{X}, Y)$ | $P_{j}(\bar{X}, Y)$ | 0 | 0 | 2 | 0 |
| $7 P_{i}(\bar{X}, Y)$ | $P_{j}(X, \bar{Y})$ | 1 | 0 | 0 | 1 |
| $8 P_{i}(\bar{X}, Y)$ | $P_{j}(\bar{X}, \bar{Y})$ | 0 | 0 | 1 | 1 |
| $9 P_{i}(X, \bar{Y})$ | $P_{j}(X, Y)$ | 1 | 1 | 0 | 0 |
| $10 P_{i}(X, \bar{Y})$ | $P_{j}(\bar{X}, Y)$ | 1 | 0 | 0 | 1 |
| $11 P_{i}(X, \bar{Y})$ | $P_{j}(X, \bar{Y})$ | 0 | 2 | 0 | 0 |
| $12 P_{i}(X, \bar{Y})$ | $P_{j}(\bar{X}, \bar{Y})$ | 0 | 1 | 0 | 1 |
| $13 P_{i}(\bar{X}, \bar{Y})$ | $P_{j}(X, Y)$ | 0 | 1 | 1 | 0 |
| $14 P_{i}(\bar{X}, \bar{Y})$ | $P_{j}(\bar{X}, Y)$ | 0 | 0 | 1 | 1 |
| $15 P_{i}(\bar{X}, \bar{Y})$ | $P_{j}(X, \bar{Y})$ | 0 | 1 | 0 | 1 |
| $16 P_{i}(\bar{X}, \bar{Y})$ | $P_{j}(\bar{X}, \bar{Y})$ | 0 | 0 | 0 | 2 |

Step 3: Prediction of protein complexes using DPClus


Figure 3.2 Procedure of validation of protein complexes based on extracted DDIs.

Firstly, domains were extracted by InterProScan (Step 1), then statistically co-occurred domain-domain pairs were extracted by Fisher's exact test for $2 \times 2$ contingency table (Step 2), and, in Step 3, validation of protein complexes were carried out using the domain-domain pairs determined in Step 2.

### 3.3.1. Step 1: Detection of domains in proteins

The InterProScan, which is a tool that combines different protein signature recognition methods into one resource, can detect protein families, domains, repeats and functional sites containing post-translational modification sites (Zdobnov and Apweiler 2001). In the present study the domains of all proteins were detected by InterProScan. The protein signature by InterProScan has the hierarchical structure, that is, a parent/child relationship between two signatures is defined in the output of InterProScan, and the parent is the entry containing a more general signature, while the children are more specific to certain members of the signature. In the context of protein interaction, domains or smaller peptide motifs act as recognition elements, therefore domains, repeats and functional sites of second depth in hierarchy but not families by InterProScan were simply used as domains in this DDI analysis.

### 3.3.2. Step 2: Statistically extracting domain-domain interactions from PPI data

Statistical analysis based on $2 \times 2$ contingency table was applied to detect significant relation between a domain pair by judging their presence and absence in a set of interacting protein pairs (Figure 3.2). Concerning the presence and absence of two domains say, $X$ and $Y$ in two interacting proteins say, $i$ and $j$ there could be 16 combinations and Figure 3.2 shows how I counted $A, B, C$ and $D$ of the contingency table corresponding to each combination. The null hypothesis is that the occurrence of domain $X$ in a protein and the occurrence of domain $Y$ in the other protein in a PPI are independent of each other. So, the test of independence between domain $X$ and domain $Y$ was performed using Fisher's exact test with significance level $\alpha=0.01$,
taking multiple hypotheses into consideration, that is, Bonferroni's correction was adopted in order to avoid statistical significance that might occur by chance. To determine significant DDIs, I statistically tested potential DDIs, containing self-DDIs, for which a protein has a domain and the other protein has another domain in at least one PPI, i.e., the count $A$ on the contingency table in Figure 3.2 is not lower than one.

### 3.3.3. Step 3: Validation of protein complexes generated using DPClus

Firstly, protein complexes were predicted by applying DPClus (Altaf-Ul-Amin et al. 2006) to the whole PPI network by setting three parameters as 0.7 for network density, 0.5 for cluster property and 3 for least number of members in a cluster. DPClus detects densely connected regions of a graph comprising nodes and edges as clusters which correspond to protein complexes in case of a PPI network. Prediction of protein complexes based on network density demonstrates a tendency that protein complexes with fewer number of members are more detected in a protein interaction network: by setting 3 as the parameter for least number of members in a cluster, complexes composed of 3 protein members are most found out. A protein complex of 3 members with network density 0.7 and over indicates complete graph among 3 members, whereas the density between $0.6 \sim 0.7$ density indicates just single-linkage chain of 3 members. Here, to discuss the topological structure of protein complexes using DDIs as evidence of true-positive PPIs, I applied 0.7 for network density in this study. After prediction of protein complexes with these parameter set, the PPIs in the complexes are validated by significant domain-domain pairs obtained in Step 2 which in turn helps to cast
insight into probable topological structure of the complexes.

### 3.4.Results and Discussion

### 3.4.1. Collection and integration of PPIs data

The PPIs of the Arabidopsis interactome were collected by following two procedures: collecting from public PPI databases and manual collecting from research papers. Any computationally predicted PPI was excluded from this study. In the first procedure, the PPI data was assembled from BIND (Bader et al. 2000; 2003), DIP (Xenarios et al. 2002), MINT (Zanzoni et al. 2002; Chatr-aryamontri et al. 2007), HPRD (Peri et al. 2003) and IntAct (Hermjakob et al. 2004) which are major PPIs data resources accepting experimentally determined PPIs from research papers. In the second procedure, 946 PPIs were manually gathered by reading experimental research papers. PPI data redundancies were removed by mapping PPI information onto the Arabidopsis genes (AGI codes), and, as a result an integrated PPI data, 3,118 PPIs composed of 1,302 Arabidopsis proteins was obtained.

### 3.4.2. Significant domain-domain interactions extracted from PPI data

Proteins must physically bind to other proteins, either stably or transiently, to perform their functions. Interaction specificity results from the binding of a modular domain to another domain or smaller peptide motif in the target protein (Pawson and Nash 2003). For example, some cytoskeletal proteins bind to actin through their modular gelsolin repeat domains
(McGough et al. 2003), and Src-homology 3 domains (SH3) bind to proline rich peptides that have a PxxP consensus sequence (Lim et al. 1994). In the context of protein interaction, such domains and peptides act as recognition elements; I refer to these binding domains or recognized peptides simply as 'domains' in this study. Over the past few years with developments of high-throughput PPI detection technologies, many researchers have shown an interest in extracting domain-domain interactions (DDIs) from large-scale PPI data by statistical methods, demonstrating that the idea of DDIs explain the cause of PPIs in some measure (Sprinzak and Margalit 2001; Riley et al. 2005; Singhal and Resat 2007; Liu et al. 2009). Here, I statistically extracted DDIs from integrated PPI data of Arabidopsis by original procedure described in the 'Method' section. Total 312 significant DDIs were obtained (Fisher's exact test, $\alpha=0.01$ with Bonferroni correction) out of 1,162 potential DDIs for which a protein has a domain and the other protein has another domain in at least one PPI.

PPI detection technologies experimentally provide the information about existence of interaction, but usually no direct information about the domains and peptides which act as recognition elements or binding sites, and determining binding domains and peptides in proteins require further analysis. Therefore, as a benchmark for true-positive DDIs, I used pairs of domains reported to interact in determined structures of protein complexes in iPfam (Finn et al. 2005). In iPfam, two domains are defined as interacting if they are close enough to form at least one interaction based on available PDB structures. It should be noted that known set of interacting domain pairs determined from structures are only a small fraction of all DDIs that may exist, i.e. though these are the gold standard DDIs, it is certainly possible that predicted

DDIs are also true and structures which contains predicted DDIs have not been determined yet. According to a recent study (Itzhaki et al. 2006), DDIs in iPfam and 3DID (Stein et al. 2005) databases could explain no more than $20 \%$ of the PPIs for any of the E. coli, S. cerevisiae, C. elegans, D.melanogaster, and H. sapience, suggesting that the number of known DDIs is rather small. So I used iPfam to assess true-positive rate with respect to P -value after Bonferroni's correction for DDIs and $70 \%$ of the known DDIs have been recognized within the threshold of 0.01 (Figure 3.3). Using this threshold, 312 statistically significant pairs of interacting domains were obtained (Figure 3.3B). So I used iPfam to assess true-positive rate with respect to P-value after Bonferroni's correction for DDIs and $70 \%$ of the known DDIs have been recognized within the threshold of 0.01 showing the effectiveness of our method (Figure 3.3A). Using this threshold, 312 statistically significant pairs of interacting domains were obtained (Figure 3.3B) and if I add 20 other gold DDIs not detected by our method then the number of total significant DDIs is 332. Now $20 \%$ of the PPIs can be explained by 66 DDIs and if I consider a linear relation then 330 DDIs are required to explain $100 \%$ PPIs which almost matches with the number 332 implying that the proposed DDI prediction method is a good one.


Figure 3.3 Validation of statistically extracted DDIs using iPfam.
(A) True-positive DDIs with the p-value; (B) true-positive DDIs out of statistically significant DDIs. By iPfam, 66 reported DDIs were obtained out of 1,162 potential domains. Of them, 46 DDIs were included in statistically significant cooccured domain-domain pairs in Step 2 in Figure 2. So the true-positive rate is estimated to be as 0.70 .

### 3.4.3. Prediction of protein complexes

In the past, the algorithm called "DPClus" was developed by Altaf-Ul-Amin et al. (2006), which extracts the densely connected regions in a network and demonstrated that many of these densely connected regions correspond to known protein complexes or protein functional units. DPClus is a robust algorithm not affected by high rate of false positives in data from high-throughput interaction-detection techniques. While predicting the protein complexes by DPClus, I adopted the "overlapping clustering mode", which allows identical proteins to be classified into different clusters, because it is biologically well established that proteins can be present in multiple complexes at different times and locations. By setting three parameters as 0.7 for network density, 0.5 for cluster property and 3 for least number of members in a cluster, 67 protein complexes were obtained from 3,118 PPIs (Figure 3.4). In the present study, 1,629 out 3,118 PPIs were supported by statistically significant domain-domain pairs (Appendix Table 1), suggesting that high-potential PPIs of true-positives in that DDIs can explain mechanisms of protein recognition and interaction. Additionally, using significant DDIs information, I provided considerable information for deducing the topological structure of all protein complexes detected by DPClus (Appendix Table 2). Appendix Table 2 shows protein members of complexes with domains detected by InterProScan and the locations of domains in proteins, PPIs and DDIs supporting the PPIs in the complexes. The $\left(^{*}\right)$ notation with a PPI indicates that the PPI is supported by statistically significant DDIs. In this method predicting topological structure of complexes using DDIs, multiple existences of identical domains in disparate locations of proteins were taken account.

Out of the all predicted 67 complexes, 16 complexes (cluster ID: 9, 20, 27, 33, 37, 39, $40,43,45,46,47,55,57,60,61,65)$ had no PPIs supported by statically significant DDIs. The proteins of these complexes have very few domains detected by InterProScan. For examples, in case of cluster 9, only two kinds of domains (IPR001932: Protein phosphatase 2C-related and IPR003124: Actin-binding WH2) were detected in 9 proteins, and in case of cluster 27 which has 4 protein members, only one domain was detected. In these cases, it seems wrong to interpret that all PPIs in predicted complexes are false-positives because these proteins in predicted complexes densely (density $\geq 0.7$ ) interact among them. It should be interpreted that DDIs explaining PPIs could not be detected by present study. On prediction of DDIs, relations of pairwise domains detected by InterProScan in interacting proteins were statistically tested, therefore, DDIs cannot be predicted if proteins do not have domains or InterProScan cannot detects domains in proteins. These results and consideration leads us to presume that there can be unknown mechanisms of protein recognition and interaction.

Other hand, in 19 complexes (cluster ID: 5, 11, 12, 16, 17, 21, 22, 24, 25, 28, 34, 42, $44,49,50,54,59,64,66$, all PPIs composing the complex were supported by significant DDIs. Interpretation of complex topological structure demands a lot of attention though these all PPIs seem like apparent true-positives. Large number of these complexes in which all PPIs were supported by significant DDIs are complexes of DNA- or RNA- binding proteins such as transcription factor complex (cluster ID: 5, 12, 17, 21, 24, 25, 42, 44, 49, 50, 54, 64). These complexes are mixture of protein and DNA or RNA and proteins interacts other proteins in the complex through DNA or RNA, inferring high possibility that ever predicted DDIs from
these DNA- or RNA- binding proteins are false-positives. In other protein complexes in which all PPIs were supported by significant DDIs, there is a tendency that most protein members have some identical domains with that of other protein members and PPIs among these proteins are explainable by a few DDIs. This tendency can be explained by considering competition of interacting domain: for instance of cluster 28 , all 4 proteins have one IPR008271 (Serine/threonine protein kinase, active site) and one IPR000719 (Protein kinase, core), and according to DDIs prediction IPR000719 interacts with IPR008271, i.e., in the complex IPR000719 of 3 proteins competitively contact with IPR008271 of another protein.

In other 32 complexes (cluster ID: $1,2,3,4,6,7,8,10,13,14,15,18,19,23,26,29$, $30,31,32,35,36,38,41,48,51,52,53,56,58,62,63,67)$, some PPIs have evidence of DDIs, but other PPIs did not have that. Figure 3.5 shows some examples, in which each node and each edge represent a protein and an interaction respectively: blue and red edges represent interactions determined by PPI experiments and interactions supported by statistically significant domain-domain pairs (suggesting direct interaction) respectively. The complex represented by cluster 41 in Figure 3.5 is composed of At1g16970, At1g4805 and At4g13870, and according to the present DDIs analysis, domain IPR005160 (Ku70/Ku80 C-terminal arm) and IPR006164 (DNA helicase, ATP-dependent, Ku type) of At1g16970 and At1g4805 contact with identical domain IPR002562 (3'-5' exonuclease) of At 4 g 13870 respectively. Thus this complex is explainable by the competitive interactions of two proteins, At1g16970 and At1g4805, with domain IPR002562 of At 4 g 13870 . The cluster 23 consists of 5 proteins and 3 of its 7 PPIs are supported by significant DDIs. At1g16240 has two domains (IPR000727:

Target SNARE coiled-coil region and IPR010989: t-SNARE) to interact with At1g28490 and At5g46860, respectively, and At1g28490 also has two domains (IPR000727 and IPR010989) to interact with At1g16240 and At5g26980. This can be explained by the consecutively ordered interaction of four proteins At5g46860, At1g16240, At1g28490 and At5g26980. In the complex 35, it is presumed from the graph that At1g74740 and At3g19290 play a vital role. In fact, Atlg74740 and At3g19290 have characteristic domains respectively: At1g74740 has IPR008271 (serine/threonine protein kinase active site in protein kinase core) and At3g19290 has IPR004827 (Basic-leucine zipper (bZIP) transcription factor), whereas other all proteins in the complex has only common domains (IPR002048: Calcium-binding EF-hand and IPR011992: EF-Hand type). Based on DDIs analysis, At1g74740 and At3g19290 play a central part of complex by directly interacting with other each protein (At3g43810, At3g51920, At3g56800, At4g14640, At5g21274) and these each 5 proteins interact each other directly.


Figure 3.4 Complex-complex interaction network emerged from PPI network by adopting DPClus.

Nodes represent complexes and edges represent their interactions. Size of nodes shows the number of proteins in complexes and thickness of edges suggests the number of PPIs between complexes. Protein complex prediction allowed identical proteins to be classified into different clusters. Blue nodes represent to be clusters containing such shared proteins and Green nodes represent not to be.

## Cluster 41



## Cluster 35



Cluster 51


Cluster 15


## Cluster 8



Figure 3.5 Prediction of the topological structure of protein complexes using significant DDIs information.

Each node and edge represents protein and interaction respectively. Blue edges show original interaction, and red edges show confidential interactions supported by significant DDIs. Blue and red circles represent self-interactions.

### 3.5. Conclusion and Remarks

This work has originally proposed a method to predict the topological structure of protein complexes by using domain-domain interactions (DDIs). Significant DDIs were determined based on statistical analysis of $2 \times 2$ contingency tables, 312 significant DDIs underlying 3,118 protein-protein interactions (PPIs) were obtained. In the present study, 1,629 out of 3,118 PPIs were supported by statistically significant DDIs. Furthermore, I validated 67 protein complexes detected in the protein interaction network composed of 1,302 proteins in view of significant DDIs. The present study makes it possible to interpret the topological structure of protein complexes by configuring interactions supported by DDIs. A further study of topological structure of protein complexes from PPI data should be conducted, which can also help computer simulations of protein complexes to develop new drugs. The more understanding of conformation of protein complexes would give new clues to development of drugs.

## Chapter 4 Conclusion and Remarks

Introduction (Chapter 1) of this dissertation explained the background of protein-protein interaction (PPI) analysis and two representative PPI detection methods (Y2H and affinity purification-MS). Then, I elucidated problems of high-throughput methods for detecting protein interactions: each detection method generates a significant number of false-positives, and that is serious problem because they cause erroneous results and misleading conclusions. Additionally, affinity purification-MS detects non-direct interactions: "prey-prey" interactions. It implies that the topological structure of the protein complexes cannot be determined from the individual experiments only. Moreover, detecting "prey-prey" interactions makes determining certain members of protein complexes complicated and difficult. Therefore, based on above background, I aimed to predict and evaluate certain protein complexes, more specifically i) to predict members of protein complexes and to evaluate that by annotation (in Chapter 2) and ii) to propose a method to predict the topological structure of protein complexes (in Chapter 3). Aiming to the goal of this dissertation, I tried two kinds of studies respectively.

Analysis on human PPI network is presented in Chapter 2, which predicts 1,264 protein complexes from 32,198 PPIs comprising 9,268 proteins by finding densely connected regions with their cluster properties in the network. These predicted complexes were annotated and evaluated using integrated data such as literatures and research papers, ternary structures, descriptions of protein, localizations, expression profiles, gene loci, and PPIs among protein
subunits. In this annotation of predicted 1,264 protein complexes, I found that 136 complexes were well-known complexes; 405 complexes were partially-known complexes; and 723 were unknown complexes with no supporting evidence of complexes in the literature.

In addition, I found 78 hypothetical proteins that were annotated as members of 82 complexes that included several well-known complexes. Recent statistics of manually curated H-InvDB proteins showed that $43 \%$ of $\mathrm{H}-\mathrm{InvDB}$ representative transcripts were hypothetical proteins (Yamasaki et al. 2008). Hence, assigning functions to hypothetical proteins of unknown function is one of the most important problems in proteome analysis. Since subunits of a complex generally tend to have the same biological function, prediction of a protein complex allows increased confidence in the annotation of hypothetical proteins. The number of hypothetical proteins to whom functions were assigned in this study was very small compared to the total number of proteins in the proteome; however, comprehensive protein complex prediction and annotation will provide strong functional clues about hypothetical proteins.

To provide this large annotation of human protein complexes, the new database "PPI view" was constructed, URL; http://h-invitational.jp/hinv/ppi/. PPI view includes a new visualization tool "PPI map", which allows users to show not only PPIs among the protein subunits, but also complex-complex interactions (CCIs) through a seamless and detailed annotation of each protein complex and its subunits. PPI view will be a useful platform for understanding protein function from the viewpoint of a complex as another level of functional
unit.

The study on Arabidopsis PPI network of Chapter 3 originally proposed a method to predict the topological structure of protein complexes by using domain-domain interactions (DDIs). As the first step, I extracted 312 statistically significant DDIs out of 1,162 DDIs underlying 3,118 protein-protein interactions (PPIs). Significant DDIs were determined based on statistical analysis of $2 \times 2$ contingency tables, estimated using pairs of domains reported to interact in determined structures of protein complexes in iPfam (Finn et al. 2005) as true positive DDIs. And, 1,629 out of 3,118 PPIs were supported by statistically significant domain-domain pairs. Next, 67 protein complexes were obtained by finding densely connected regions with their cluster properties in the PPI network. Finally, I validated 67 protein complexes and discussed the topological structure based on DDI information extracted in the first step. The present study makes it possible to interpret the topological structure of protein complexes by configuring interactions supported by DDIs.

In present study, using Arabidopsis PPI network composed of 1,302 proteins and 3,118 PPIs, I tried to predict the topological structure of protein complexes based on DDIs and proposed the method. Here, human PPI network is composed of 9,268 proteins and 32,198 PPIs, so proteins and PPIs of human are respectively 7.12 times and 10.33 times as much as that of Arabidopsis. On human PPI network of about 10 times the large of Arabidopsis one, analysis of the topological structure of protein complexes will require a lot of machine power: especially, calculation of statistical analysis of DDIs grows at a rate of second power of
number of domains in interacting proteins, because all potential domain pairs are tested by Fisher's exact test. Therefore, progress of machine power or contrivance of computation is required in order to analyze the topological structure of protein complexes on human.

In this dissertation, I discussed prediction of protein complex based on protein interaction network and topological structure of protein complexes, taking into account the fact that experimental methods to detect PPIs generates a significant number of false-positives and affinity purification- MS equally detect direct and non-direct interactions. To understand the protein functions required for various biological processes in cells, comprehensive analysis of PPIs and protein complexes as fundamental functional units of macromolecular systems would provide a valuable framework. And, in order to elucidate PPIs, it is important to unravel the mechanisms of protein recognition at the molecular level. The understanding of mechanisms of protein recognition and interaction would release us from problems of false-negative and false-positive. Concurrently a further study of topological structure of protein complexes from PPI data should be conducted, which can also help computer simulation of protein complexes to develop new drugs. The more understanding of conformation of protein complexes and mechanisms of protein recognition and interaction would give new clues to development of drugs.

## References

Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S (2006) Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinformatics 7:207
Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389-3402.

Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, Saito R, Ara T, Nakahigashi K, Huang HC, Hirai A et al (2006) Large-scale identification of protein-protein interaction of Escherichia coli K-12. Genome Res. 16(5):686-691
Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. Nat. Genet. 25(1):25-29.

Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31(1): 248-250

Bader GD, Hogue CW (2000) BIND-a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics 16(5): 465-477

Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics Jan 13;4:2. Epub 2003 Jan 13.
Bader GD, Hogue CW (2002) Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol 20(10):991-997.
Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2007) GenBank. Nucleic Acids Res 35(Database issue):D21-25.

Cai J, Uhlmann F, Gibbs E, Flores-Rozas H, Lee CG, Phillips B, Finkelstein J, Yao N, O'Donnell M, Hurwitz J (1996) Reconstitution of human replication factor C from its five subunits in baculovirus-infected insect cells. Proc Natl Acad Sci U S A 93(23):12896-12901.
Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35(Database
issue):D572-574
Chen XW and Liu M (2006) Domain Based Predictive Models for Protein-Protein Interaction Prediction. EURASIP Journal on Applied Signal Processing 2006
Deng M, Mehta S, Sun F, Chen T (2002) Inferring domain-domain interactions from protein-protein interactions. Genome Res. 12: 1540-1548

Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41(1):98-107.
Finn RD, Marshall M, Bateman A (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 21: 410-412

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A (2008) The Pfam protein families database. Nucleic Acids Res. 36: D281-288

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, 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 MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141-7
Gong S, Park C, Choi H, Ko J, Jang I, Lee J, Bolser DM, Oh D, Kim DS, Bhak J (2005) A protein domain interaction interface database: InterPare. BMC Bioinformatics 6: 207
Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH (2002) Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. Mol Biol Evol 19(3):256-262.
Guimaraes KS, Jothi R, Zotenko E, Przytycka TM (2006) Predicting domain-domain interactions using a parsimony approach. Genome Biol. 7: R104

Hakes L, Robertson DL, Oliver SG, Lovell SC (2007) Protein interactions from complexes: a structural perspective. Comp Funct Genomics:49356

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C et al (2004) The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol. 22(2):177-183

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron

M, Roechert B, Roepstorff P, Valencia A et al (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res 32(Database issue): D452-455
Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T (2001) Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast 18(6):523-531 Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868):180-3
Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino $M$ et al (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2(6):e162

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98(8):4569-74

Itzhaki Z, Akiva E, Altuvia Y, Margalit H (2006) Evolutionary conservation of domain-domain interactions. Genome Biol. 7: R125
Jothi R, Cherukuri PF, Tasneem A, Przytycka TM (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. J. Mol Biol 362: 861-875

Kann MG, Jothi R, Cherukuri PF, Przytycka TM (2007) Predicting protein domain interactions from coevolution of conserved regions. Proteins 67: 811-820

Kastrin A, Peterlin B, Hristovski D (2010) Chi-square-based Scoring Function for Categorization of MEDLINE Citations. Methods Inf Med. 2010 Jan 20;49(2)

Kawamoto S, Ohnishi T, Kita H, Chisaka O, Okubo K (1999) Expression profiling by iAFLP: A PCR-based method for genome-wide gene expression profiling. Genome Res 9(12):1305-1312

Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FC (2002)

Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol. Cell 9(5):1133-1143
Kim KI, van de Wiel MA (2008) Effects of dependence in high-dimensional multiple testing problems. BMC Bioinformatics 9:114
Kumar A, Snyder M (2002) Protein complexes take the bait. Nature 415(6868):123-124
Lee H, Deng M, Sun F, Chen T (2006) An integrated approach to the prediction of domain-domain interactions. BMC Bioinformatics 7: 269

Lim WA, Richards FM, Fox RO (1994) Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. Nature 372:375-379

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227(4693):1435-1441

Litovchick L, Sadasivam S, Florens L, Zhu X, Swanson SK, Velmurugan S, Chen R, Washburn MP, Liu XS, DeCaprio JA (2007) Evolutionarily conserved multisubunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence. Mol Cell 26(4):539-551

Littler SJ and Hubbard SJ (2005) Conservation of orientation and sequence in protein domain-domain interactions. J. Mol. Biol. 345: 1265-1279

Liu CT, Yuan S, Li KC (2009) Patterns of co-expression for protein complexes by size in Saccharomyces cerevisiae. Nucleic Acids Res 37(2):526-532

Liu M, Chen XW, Jothi R (2009) Knowledge-guided inference of domain-domain interactions from incomplete protein-protein interaction networks. Bioinformatics. Oct 1;25(19):2492-9

Liu Y, Liu N, Zhao H (2005) Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. Bioinformatics 21: 3279-3285

McGough AM, Staiger CJ, Min JK, Simonetti KD (2003) The gelsolin family of actin regulatory proteins: modular structures, versatile functions. FEBS Lett. 552:75-81

Mrowka R, Patzak A, Herzel H (2001) Is there a bias in proteome research? Genome Res. 11: 1971-1973
Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol. 3: e405

Ng SK, Zhang Z, Tan SH (2003) Integrative approach for computationally inferring protein domain interactions. Bioinformatics 19: 923-929
Nye TM, Berzuini C, Gilks WR, Babu MM, Teichmann SA (2005) Statistical analysis of domains in interacting protein pairs. Bioinformatics 21: 993-1001
O'Donnell M, Onrust R, Dean FB, Chen M, Hurwitz J (1993) Homology in accessory proteins of replicative polymerases--E. coli to humans. Nucleic Acids Res 21(1):1-3

Pagel P, Wong P, Frishman D (2004) A domain interaction map based on phylogenetic profiling. J. Mol. Biol. 344: 1331-1346

Pawson T and Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300:445-452

Pazos F and Valencia A (2008) Protein co-evolution, co-adaptation and interactions. The EMBO Journal 27: 2648-2655

Pereira-Leal JB, Teichmann SA (2005) Novel specificities emerge by stepwise duplication of functional modules. Genome Res. 15:552-559

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 13: 2363-2371

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35(Database issue):D61-65

Qiu P, Gentles AJ, Plevritis SK (2010) Reducing the Computational Complexity of Information Theoretic Approaches for Reconstructing Gene Regulatory Networks. J Comput Biol. 2010 Jan 15
Raghavachari B, Tasneem A, Przytycka TM, Jothi R (2008) DOMINE: a database of protein
domain interactions. Nucleic Acids Res. 36: D656-661
Riley R, Lee C, Sabatti C, Eisenberg D (2005) Inferring protein domain interactions from databases of interacting proteins. Genome Biol. 6: R89
Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318(2):595-608
Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjolander K (2010) Active Site Prediction using Evolutionary and Structural Information. Bioinformatics. 2010 Jan 14.

Schuster-Bockler B and Bateman A (2007) Reuse of structural domain-domain interactions in protein networks. BMC Bioinformatics 8: 259

Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF (2005) Atlas - a data warehouse for integrative bioinformatics. BMC Bioinformatics 6:34

Shoemaker BA, Panchenko AR, Bryant SH (2006) Finding biologically relevant protein domain interactions: conserved binding mode analysis. Protein Sci. 15: 352-361
Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Comput Biol. 2007 Mar 30;3(3):e42.

Singhal M and Resat H (2007) A domain-based approach to predict protein-protein interactions. BMC Bioinformatics 8: 199

Sleno L and Emili A (2008) Proteomic methods for drug target discovery. Curr. Opin. Chem. Biol. 12: 46-54

Sprinzak E and Margalit H (2001) Correlated sequence-signatures as markers of protein-protein interaction. J. Mol. Biol. 311: 681-692
Stein A, Russell RB, Aloy P (2005) 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Res. 33: D413-417

Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr 54(Pt 6 Pt 1):1078-1084

Tanino M, Debily MA, Tamura T, Hishiki T, Ogasawara O, Murakawa K, Kawamoto S, Itoh K, Watanabe S, de Souza SJ et al (2005) The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. Nucleic Acids Res 33(Database issue):D567-572
The UniProt Consortium (2007) The Universal Protein Resource (UniProt). Nucleic Acids Res 35(Database issue):D193-197

Titz B, Schlesner M, Uetz P (2004) What do we learn from high-throughput protein interaction data? Expert Rev Proteomics 1(1):111-121
Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770):623-7
von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417: 399-403

Wang H, Segal E, Ben-Hur A, Li QR, Vidal M, Koller D (2007) InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. Genome Biol. 8: R192.

Werner F (2007) Structure and function of archaeal RNA polymerases. Mol. Microbiol. 65(6):1395-1404.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S et al (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35(Database issue):D5-12

Wicker J, Fenner K, Ellis L, Wackett L, Kramer S (2010) Predicting Biodegradation Products and Pathways: A Hybrid Knowledge-Based and Machine Learning Based Approach. Bioinformatics. 2010 Jan 26
Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30(1): 303-305

Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, Takeda J, Taniya T, Sakate R, Kikugawa S, Shimada M et al (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Nucleic Acids Res 36(Database issue):D793-799

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) MINT: a Molecular INTeraction database. FEBS Lett. 513(1): 135-140

Zdobnov EM and Apweiler R (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17(9): 847-8.

## Achievements

## Reviewed publications and manuscripts in preparation

1. Kensaku Nishikata, Masayoshi Wada, Hiroki Takahashi, Kensuke Nakamura, Shigehiko KanAya, Md Altaf-Ul-Amin, Predicting conformation of protein complexes by determining statistically significant Domain-Domain interactions, Plant Biotechnology 26: 495-501
2. Chisato Yamasaki, Katsuhiko Murakami, Yasuyuki Fujii, Yoshiharu Sato, Erimi Harada, Jun-ichi Takeda, Takayuki Taniya, Ryuichi Sakate, Shingo Kikugawa, Makoto Shimada, Motohiko Tanino, Kanako O. Koyanagi, Roberto A. Barrero, Craig Gough, Hong-Woo Chun, Takuya Habara, Hideki Hanaoka, Yosuke Hayakawa, Phillip B. Hilton, Yayoi Kaneko, Masako Kanno, Yoshihiro Kawahara, Toshiyuki Kawamura, Akihiro Matsuya, Naoki Nagata, Kensaku Nishikata, Akiko Ogura Noda, Shin Nurimoto, Naomi Saichi, Hiroaki Sakai, Ryoko Sanbonmatsu, Rie Shiba, Mami Suzuki, Kazuhiko Takabayashi, Aiko Takahashi, Takuro Tamura, Masayuki Tanaka, Susumu Tanaka, Fusano Todokoro, Kaori Yamaguchi, Naoyuki Yamamoto, Toshihisa Okido, Jun Mashima, Aki Hashizume, Lihua Jin, Kyung-Bum Lee, Yi-Chueh Lin, Asami Nozaki, Katsunaga Sakai, Masahito Tada, Satoru Miyazaki, Takashi Makino, Hajime Ohyanagi, Naoki Osato, Nobuhiko Tanaka, Yoshiyuki Suzuki, Kazuho Ikeo, Naruya Saitou, Hideaki Sugawara, Claire O’Donovan, Tamara Kulikova, Eleanor Whitfield, Brian Halligan, Mary Shimoyama, Simon Twigger, Kei Yura, Kouichi Kimura, Tomohiro Yasuda, Tetsuo Nishikawa, Yutaka Akiyama, Chie Motono, Yuri Mukai, Hideki Nagasaki, Makiko Suwa, Paul Horton, Reiko Kikuno, Osamu Ohara, Doron Lancet, Eric Eveno, Esther Graudens, Sandrine Imbeaud, Marie Anne Debily, Yoshihide Hayashizaki, Clara Amid, Michael Han, Andreas Osanger, Toshinori Endo, Michael A. Thomas, Mika Hirakawa, Wojciech Makalowski, Mitsuteru Nakao, Nam-Soon Kim, Hyang-Sook Yoo, Sandro J. De Souza, Maria de Fatima Bonaldo, Yoshihito Niimura, Vladimir Kuryshev, Ingo Schupp, Stefan Wiemann, Matthew Bellgard, Masafumi Shionyu, Libin Jia, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, Qinghua Zhang, Mitiko Go, Shinsei Minoshima, Masafumi Ohtsubo, Kousuke Hanada, Peter Tonellato, Takao Isogai, Ji Zhang, Boris Lenhard, Sangsoo Kim, Zhu Chen, Ursula Hinz, Anne Estreicher, Kenta Nakai, Izabela

Makalowska，Winston Hide，Nicola Tiffin，Laurens Wilming，Ranajit Chakraborty， Marcelo Bento Soares，Maria Luisa Chiusano，Yutaka Suzuki，Charles Auffray，Yumi Yamaguchi－Kabata，Takeshi Itoh，Teruyoshi Hishiki，Satoshi Fukuchi，Ken Nishikawa， Sumio Sugano，Nobuo Nomura，Yoshio Tateno，Tadashi Imanishi，Takashi Gojobori （2008）＂The H－Invitational Database（H－InvDB），a comprehensive annotation resource for human genes and transcripts＂，Nucleic Acids Res．36（Database issue）：D793－D799

3．Arifuzzaman M．，Maeda M．，Itoh A．，Nishikata K．，Takita C．，Saito R．，Ara T．， Nakahigashi K．，Huang，H．C．，Hirai A．，Tsuzuki K．，Nakamura S．，Altaf－Ul－Amin M．， Oshima T．，Baba T．，Yamamoto N．，Kawamura T．，Ioka－Nakamichi T．，Kitagawa M．， Tomita M．，Kanaya S．，Wada C．and Mori H．（2006）＂Large－scale identification of protein－protein interaction of Escherichia coli K－12＂，Genome Research 16：686－691
4．Kensaku Nishikata，Shingo Kikugawa，Yoshiharu Sato，Katsuhiko Murakami，Mami Suzuki，Md Altaf－Ul－Amin，Shigehiko Kanaya，and Tadashi Imanishi，Integrative Annotation of Predicted Protein Complexes in the Network of Human Protein－Protein Interactions in H－Invitational Human Transcriptome Database，in preparation．

## Other publications

## Japanese journal

1．Md．Altaf－Ul－Amin，和田眞昌，西潟憲策，旭弘子，黒川顕，金谷重彦，生体ネッ トワークにおけるタンパク質間相互作用の研究法，生体の科学， 58（5）：391－394（2007）

## Public DB and tool

2．＂PPI view＂and＂PPI map＂http：／／www．h－invitational．jp／hinv／ppi／

## International Conferences

1．Shingo Kikugawa，Kensaku Nishikata，Md．Altaf－Ul－Amin，Shigehiko Kanaya，and Tadashi Imanishi，＂Integration and Analysis of Human Protein－Protein Interaction Data on the Human Transcriptome Database，H－InvDB＂，HUPO 6th Annual World Congress， October 6－10，2007，Seoul，Korea．
2．Md．Altaf－Ul－Amin，Kensaku Nishikata，Toshihiro Koma，Teppei Miyasato，Yoko Shinbo，Md．Arifuzzaman，Chieko Wada，Maki Maeda，Taku Oshima，Hirotada Mori，

Shigehiko Kanaya，＂Prediction of Protein Functions Based on K－cores of Protein－Protein Interaction Networks and Amino Acid Sequences＂，The 14th International Conference on Genome Informatics，December 14－17，2003，Yokohama，Japan．

## Local Conferences

1．西潟 憲策，喜久川 真悟，渡辺 賢，金谷 重彦，今西 規「タンパク質間相互作用に関わるドメインの統計的抽出」 B MB 2007 第 30 回日本分子生物学会年会•第 80 回日本生化学会大会 合同大会，2007年12月11日－15日，横浜

## Appendix

## Appendix Table 1

List of PPIs supported by statistically significant domain-domain interactions.
A) PPIs supported by statistically significant domain-domain interactions.












|  | IPR001092 | 009057 |
| :---: | :---: | :---: |
| No. 0257 | At1g26260 | IPR001092 |
|  | At2g35940 | IPR009057 \| IPR006563 |
|  | IPR001092 | 009057 |
| No. 0258 | At1g26260 | IPR001092 |
|  | At4g32040 | IPR005541 \| IPR009057| |
|  | IPR001092 | 009057 |
| No. 0259 | At1g26260 | IPR001092 |
|  | At4g32980 | IPR009057 \| IPR006563 |
|  | IPR001092 | 009057 |
| No. 0260 | At1g26260 | IPR001092 |
|  | At4g36870 | IPR009057\| IPR006563 |
|  | IPR001092 | 009057 |
| No. 0261 | At1g26260 | IPR001092 |
|  | At5g02030 | IPR009057\|IPR006563 |
|  | IPR001092 | 009057 |
| No. 0262 | At1g26260 | IPR001092 |
|  | At5g25220 | IPR005541 \| IPR009057| |
|  | IPR001092 | 009057 |
| No. 0263 | At1g26260 | IPR001092 |
|  | At5g41410 | IPR009057\|IPR006563 |
|  | IPR001092 | 009057 |
| No. 0264 | At1g26310 | IPR002487 \| IPR002100 |
|  | At1g 48150 | IPR002100 |
|  | IPR002100 | 002100 \| IPR002100-IPR |
| No. 0265 | At1g26310 | IPR002487 \| IPR002100 |
|  | At2g22540 | IPR002100 \| IPR002487 |
|  | IPR002100 | 002100 \| IPR002100-IPR |
| No. 0266 | At1g26310 | IPR002487 \| IPR002100 |
|  | At2g 45660 | IPR002487 \\| IPR002100 |
|  | IPR002100 | 002100 \| IPR002100-IPR |
| No. 0267 | At1g26310 | IPR002487 \| IPR002100 |
|  | At4g24540 | IPR002487\| IPR002100 |
|  | IPR002100 | 002100\|IPR002100-IPR |
| No. 0268 | At1g26830 | IPR011991 |
|  | At2g26990 | IPR000717 |
|  | IPR000717 | 011991 |
| No. 0269 | At1g26830 | IPR011991 |
|  | At2g39760 | IPR008974 \| IPR000210 |
|  | IPR000210 | 011991\|IPR008974-IPR |
| No. 0270 | At1g26830 | IPR011991 |
|  | At3g06190 | IPR008974 \| IPR000210 |
|  | IPR000210 | 011991 \| IPR008974-IPR |
| No. 0271 | At1g26830 | IPR011991 |
|  | At3g43700 | IPR008974 \| IPR000210 |
|  | IPR000210 | 011991\|IPR008974-IPR |
| No. 0272 | At1g26830 | IPR011991 |
|  | At3g48360 | IPR000210 \| IPR000197 |
|  | IPR000210 | 011991 |
| No. 0273 | At1g26830 | IPR011991 |
|  | At3g51770 | IPR001440 \| IPR000210 |
|  | IPR000210 | 011991 |
| No. 0274 | At1g26830 | IPR011991 |
|  | At4g08455 | IPR000210 |
|  | IPR000210 | 011991 |
| No. 0275 | At1g26830 | IPR011991 |
|  | At4g37610 | IPR000210 \| IPR000197 |
|  | IPR000210 | 011991 |
| No. 0276 | At1g26830 | IPR011991 |
|  | At5g 13060 | IPR000210\|IPR000225 | |
|  | IPR000210 | 011991 |
| No. 0277 | At1g26830 | IPR011991 |
|  | At5g 19000 | IPR008974 \| IPR000210 |
|  | IPR000210 | 011991\|IPR008974-IPR |
| No. 0278 | At1g26830 | IPR011991 |
|  | At5g48510 | IPR000210 |
|  | IPR000210 | 011991 |
| No. 0279 | At1g26830 | IPR011991 |
|  | At5g63160 | IPR000210\|IPR000197 |
|  | IPR000210 | 011991 |
| No. 0280 | At1g27190 | IPR001611\|IPR011009| |
|  | At2g41090 | IPR011992 \| IPR002048 |
|  | $\begin{aligned} & \text { IPR000719 } \\ & \text { \|IPR01100 } \end{aligned}$ | $\begin{aligned} & \text { R002048\|IPR000719-IPR0 } \\ & \text { R011992 } \end{aligned}$ |
| No. 0281 | At1g27190 | IPR001611\|IPR011009| |
|  | At3g43810 | IPR011992 \| IPR002048 |
|  | $\begin{aligned} & \text { IPR000719 } \\ & \text { \|IPR01100 } \end{aligned}$ | $\begin{aligned} & \text { R002048\|IPR000719-IPR0 } \\ & \text { R011992 } \end{aligned}$ |
| No. 0282 | At1g27190 | IPR001611\|IPR011009 | |




|  | IPR001680-IPR001810\|IPR001680-IPR011043 |
| :---: | :---: |
| No. 0328 | At1g31630 IPR002100 |
|  | At1g65360 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0329 | At1g31630 IPR002100 |
|  | At1g69120 IPR002487\|IPR002100 |
|  | IPR002100-IPR002100\|IPR002100-IPR002487 |
| No. 0330 | At1g31630 IPR002100 |
|  | At2g03710 IPR002487\|IPR002100 |
|  | IPR002100-IPR002100\|IPR002100-IPR002487 |
| No. 0331 | At1g31630 IPR002100 |
|  | At2g45650 IPR002487\| IPR002100 |
|  | IPR002100-IPR002100\|IPR002100-IPR002487 |
| No. 0332 | At1g31630 IPR002100 |
|  | At4g36590 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0333 | At1g31630 IPR002100 |
|  | At5g27960 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0334 | At1g31630 IPR002100 |
|  | At5g58890 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0335 | At1g31630 IPR002100 |
|  | At5g60440 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0336 | At1g31640 IPR002100 |
|  | At1g65360 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0337 | At1g31640 IPR002100 |
|  | At1g69120 IPR002487\|IPR002100 |
|  | IPR002100-IPR002100\|IPR002100-IPR002487 |
| No. 0338 | At1g31640 IPR002100 |
|  | At2g03710 IPR002487\|IPR002100 |
|  | IPR002100-IPR002100\|IPR002100-IPR002487 |
| No. 0339 | At1g31640 IPR002100 |
|  | At2g45650 IPR002487\| IPR002100 |
|  | IPR002100-IPR002100\|IPR002100-IPR002487 |
| No. 0340 | At1g31640 IPR002100 |
|  | At3g18650 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0341 | At1g31640 IPR002100 |
|  | At5g23260 IPR002487\|IPR002100 |
|  | IPR002100-IPR002100\|IPR002100-IPR002487 |
| No. 0342 | At1g31640 IPR002100 |
|  | At5g27960 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0343 | At1g31640 IPR002100 |
|  | At5g60440 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0344 | At1g31640 IPR002100 |
|  | At5g65330 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0345 | At1g31930 IPR011025 |
|  | At3g15730 IPR001736\| IPR008973 |
|  | IPR001736-IPR011025\|IPR008973-IPR011025 |
| No. 0346 | At1g32070 IPR000182 |
|  | At1g32070 IPR000182 |
|  | IPR000182-IPR000182 |
| No. 0347 | At1g32440 IPR011037 |
|  | At3g22960 IPR011037 |
|  | IPR011037-IPR011037 |
| No. 0348 | At1g43700 IPR004827 |
|  | At1g43700 IPR004827 |
|  | IPR004827-IPR004827 |
| No. 0349 | At1g45249 IPR004827 |
|  | At3g57530 IPR011992\|IPR008271| IPR002048|IPR011009|IPR000719 |
|  | IPR002048-IPR004827\|IPR004827-IPR011992 |
| No. 0350 | At1g46408 IPR002100 |
|  | At1g65330 IPR002100 |
|  | IPR002100-IPR002100 |
| No. 0351 | At1g46408 IPR002100 |
|  | At1g69120 IPR002487\|IPR002100 |
|  | IPR002100-IPR002100\|IPR002100-IPR002487 |
| No. 0352 | At1g46408 IPR002100 |
|  | At1g77080 IPR002487\|IPR002100 |
|  | IPR002100-IPR002100\|IPR002100-IPR002487 |
| No. 0353 | At1g46408 IPR002100 |
|  | At2g03710 IPR002487\| IPR002100 |
|  | IPR002100-IPR002100 \| IPR002100-IPR002487 |


| No. 0354 | At1g46408 | IPR002100 |
| :---: | :---: | :---: |
|  | At2g45650 | IPR002487\| IPR002100 |
|  | IPR002100 | 002100 \|IPR002100-IPR |
| No. 0355 | At1g46408 | IPR002100 |
|  | At3g18650 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0356 | At1g46408 | IPR002100 |
|  | At4g11250 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0357 | At1g46408 | IPR002100 |
|  | At4g 18960 | IPR002487\|IPR002100 |
|  | IPR002100 | 2002100 \| IPR002100-IPR |
| No. 0358 | At1g46408 | IPR002100 |
|  | At5g23260 | IPR002487 \\| IPR002100 |
|  | IPR002100 | 2002100 \| IPR002100-IPR |
| No. 0359 | At1g46408 | IPR002100 |
|  | At5g26870 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0360 | At1g46408 | IPR002100 |
|  | At5g27050 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0361 | At1g46408 | IPR002100 |
|  | At5g27070 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0362 | At1g46408 | IPR002100 |
|  | At5g27580 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0363 | At1g46408 | IPR002100 |
|  | At5g27960 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0364 | At1g46408 | IPR002100 |
|  | At5g38740 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0365 | At1g46408 | IPR002100 |
|  | At5g40120 | IPR002100 |
|  | IPR002100 | R002100 |
| No. 0366 | At1g 46408 | IPR002100 |
|  | At5g40220 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0367 | At1g 46408 | IPR002100 |
|  | At5g58890 | IPR002100 |
|  | IPR002100 | R002100 |
| No. 0368 | At1g46408 | IPR002100 |
|  | At5g60440 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0369 | At1g46408 | IPR002100 |
|  | At5g65330 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0370 | At1g 47260 | IPR001451 \| IPR011004 |
|  | At1g47260 | IPR001451 \| IPR011004 |
|  | IPR001451 | 2001451\|IPR001451-IPR |
| No. 0371 | At1g47260 | IPR001451 \| IPR011004 |
|  | At3g48680 | IPR001451 \| IPR011004 |
|  | IPR001451 | 2001451\|IPR001451-IPR |
| No. 0372 | At1g47260 | IPR001451 \| IPR011004 |
|  | At5g63510 | IPR001451 \| IPR011004 |
|  | IPR001451 | P001451\|IPR001451-IPR |
| No. 0373 | At1g47760 | IPR002100 |
|  | At3g18650 | IPR002100 |
|  | IPR002100 | R002100 |
| No. 0374 | At1g47760 | IPR002100 |
|  | At5g58890 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0375 | At1g47760 | IPR002100 |
|  | At5g65330 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0376 | At1g48050 | IPR005160 \| IPR006164 |
|  | At4g13870 | IPR002562 |
|  | IPR002562 | 005160 \| IPR002562-IPR |
| No. 0377 | At1g48150 | IPR002100 |
|  | At1g60040 | IPR002100 |
|  | IPR002100 | R002100 |
| No. 0378 | At1g48150 | IPR002100 |
|  | At1g60880 | IPR002100 |
|  | IPR002100 | R002100 |
| No. 0379 | At1g48150 | IPR002100 |
|  | At1g60920 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 0380 | At1g48150 | IPR002100 |






| No. 0480 | At1g60880 | IPR002100 |
| :---: | :---: | :---: |
|  | At4g11250 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0481 | At1g60880 | IPR002100 |
|  | At5g65330 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0482 | At1g60920 | IPR002100 |
|  | At3g 18650 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0483 | At1g60920 | IPR002100 |
|  | At4g11250 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0484 | At1g60920 | IPR002100 |
|  | At5g65330 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0485 | At1g62360 | IPR005541 |
|  | At1g62990 | IPR005541 |
|  | $\begin{aligned} & \text { IPR005539 } \\ & \text { \| IPR00554 } \\ & 7 \\ & \hline \end{aligned}$ | $\begin{aligned} & \text { R005539\|IPR } \\ & \text { PR005541\|IP } \end{aligned}$ |
| No. 0486 | At1g62360 | IPR005541 |
|  | At1g75410 | IPR009057 |
|  | IPR005539 \| IPR00554 | $\begin{aligned} & \hline \mathrm{R006563\mid IPR} \\ & \text { PR009057\|IP } \end{aligned}$ |
| No. 0487 | At1g62360 | IPR005541 |
|  | At2g23760 | IPR009057 |
|  | $\begin{aligned} & \text { IPR005539 } \\ & \text { \| IPR00554] } \end{aligned}$ | $\begin{aligned} & \mathrm{R} 006563 \mid \mathrm{IPR} \\ & \text { PR009057\|IP } \\ & \hline \end{aligned}$ |
| No. 0488 | At1g62360 | IPR005541 |
|  | At2g35940 | IPR009057 |
|  | IPR005539 \| IPR00554 | $\begin{aligned} & \mathrm{R} 006563 \mid \mathrm{IPR} \\ & \text { RR009057\|IP } \\ & \hline \end{aligned}$ |
| No. 0489 | At1g62360 | IPR005541 |
|  | At4g32980 | IPR009057 |
|  | IPR005539 \| IPR00554 | $\begin{aligned} & \text { R006563\|IPR } \\ & \text { PR009057।IP } \end{aligned}$ |
| No. 0490 | At1g62360 | IPR005541 |
|  | At4g34610 | IPR009057 |
|  | IPR005539 \| IPR00554 | $\begin{aligned} & \text { R006563\|IPR } \\ & \text { PR009057।IP } \\ & \hline \end{aligned}$ |
| No. 0491 | At1g62360 | IPR005541 |
|  | At4g36870 | IPR009057 |
|  | IPR005539 \| IPR00554 | $\begin{aligned} & \text { R006563\|IPR } \\ & \text { PR009057\|IP } \\ & \hline \end{aligned}$ |
| No. 0492 | At1g62360 | IPR005541 |
|  | At5g02030 | IPR009057 |
|  | $\begin{aligned} & \text { IPR005539 } \\ & \text { \| IPR00554 } \end{aligned}$ | $\begin{aligned} & \mathrm{R} 006563 \mid \mathrm{IPR} \\ & \text { PR009057\|IP } \\ & \hline \end{aligned}$ |
| No. 0493 | At1g62360 | IPR005541 |
|  | At5g41410 | IPR009057 |
|  | IPR005539 \| IPR00554 | $\begin{aligned} & \hline \mathrm{RO06563} \mathrm{\\| IPR} \\ & \text { PR009057\|IP } \end{aligned}$ |
| No. 0494 | At1g62990 | IPR005541 |
|  | At1g62990 | IPR005541 |
|  | $\begin{aligned} & \text { IPR005539 } \\ & \text { \| IPR005540 } \\ & 7 \\ & \hline \end{aligned}$ | $\begin{aligned} & \text { R005539\|IPR } \\ & \text { PR005541\|IP } \end{aligned}$ |
| No. 0495 | At1g62990 | IPR005541 |
|  | At2g16400 | IPR009057 |
|  | $\begin{aligned} & \text { IPR005539- } \\ & \text { \| IPR00554] } \end{aligned}$ | $\begin{aligned} & \text { R006563\|IPR } \\ & \text { PR009057\|IP } \\ & \hline \end{aligned}$ |
| No. 0496 | At1g62990 | IPR005541 |
|  | At2g27220 | IPR009057 |
|  | IPR005539 \| IPR00554 | $\begin{aligned} & \hline \mathrm{R006563\mid IPR} \\ & \text { PR009057\|IP } \end{aligned}$ |
| No. 0497 | At1g62990 | IPR005541 |
|  | At2g30400 | IPR006458 |
|  | IPR005539 | 2006458\|IPR |
| No. 0498 | At1g62990 | IPR005541 |
|  | At3g52525 | IPR006458 |
|  | IPR005539 | 2006458\|IPR |
| No. 0499 | At1g62990 | IPR005541 |
|  | At4g08150 | IPR005541 |
|  | $\begin{aligned} & \text { IPR005539 } \\ & \text { \| IPR00554 } \\ & 7 \\ & \hline \end{aligned}$ | $\begin{aligned} & \text { R005539\|IPR } \\ & \text { PR005541।IP } \\ & \hline \end{aligned}$ |
| No. 0500 | At1g62990 | IPR005541 |
|  | At4g32040 | IPR005541 |
|  | IPR005539 \| IPR005540 | $\begin{aligned} & \text { R005539\|IPR } \\ & \text { PR005541।IP } \\ & \hline \end{aligned}$ |



| No. 0526 | At1g65300 | IPR002100 |
| :---: | :---: | :---: |
|  | At5g60440 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0527 | At1g65330 | IPR002100 |
|  | At4g36590 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0528 | At1g65330 | IPR002100 |
|  | At5g04640 | IPR002100 |
|  | IPR002100 | R002100 |
| No. 0529 | At1g65330 | IPR002100 |
|  | At5g27130 | IPR002100 |
|  | IPR002100 | R002100 |
| No. 0530 | At1g65330 | IPR002100 |
|  | At5g60440 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 0531 | At1g66240 | IPR006121 |
|  | At5g44790 | IPR008250 |
|  | IPR005834 | R006121\|IPR |
| No. 0532 | At1g66340 | IPR003018 |
|  | At3g21510 | IPR008207 |
|  | IPR003594 | R008207\|IPR |
| No. 0533 | At1g66340 | IPR003018 |
|  | At3g29350 | IPR008207 |
|  | IPR003594 | R008207\|IPR |
| No. 0534 | At1g66340 | IPR003018 |
|  | At5g39340 | IPR008207 |
|  | IPR003594 | 2008207\|IPR |
| No. 0535 | At1g66370 | IPR001005 |
|  | At4g00480 | IPR001092 |
|  | IPR001005 | $2001092 \mid$ IPR |
| No. 0536 | At1g66370 | IPR001005 |
|  | At4g09820 | IPR001092 |
|  | IPR001005 | R001092\|IPR |
| No. 0537 | At1g66380 | IPR001005 |
|  | At4g00480 | IPR001092 |
|  | IPR001005 | 2001092 \| IPR |
| No. 0538 | At1g66380 | IPR001005 |
|  | At4g09820 | IPR001092 |
|  | IPR001005 | 2001092 \| IPR |
| No. 0539 | At1g66390 | IPR001005 |
|  | At4g00480 | IPR001092 |
|  | IPR001005 | R001092 \|IPR |
| No. 0540 | At1g66390 | IPR001005 |
|  | At4g09820 | IPR001092 |
|  | IPR001005 | R001092\|IPR |
| No. 0541 | At1g66390 | IPR001005 |
|  | At5g41315 | IPR001092 |
|  | IPR001005 | R001092 \| IPR |
| No. 0542 | At1g66750 | IPR008271 |
|  | At4g28980 | IPR008271 |
|  | IPR000719 | R008271\|IPR |
| No. 0543 | At1g66750 | IPR008271 |
|  | At5g27620 | IPR011028 |
|  | $\begin{aligned} & \hline \text { IPR000719 } \\ & \text { \| IPR01100 } \\ & \hline \end{aligned}$ | $\begin{aligned} & \hline \text { R006671 IIPR } \\ & \text { PR011028 } \\ & \hline \end{aligned}$ |
| No. 0544 | At1g67340 | IPR006597 |
|  | At5g42190 | IPR001680 |
|  | IPR001680 | 2001810 |
| No. 0545 | At1g67580 | IPR008271 |
|  | At4g28980 | IPR008271 |
|  | IPR000719 | R008271\|IPR |
| No. 0546 | At1g68050 | IPR011043 |
|  | At3g47500 | IPR003851 |
|  | IPR000014 | 2003851\|IPR |
| No. 0547 | At1g68050 | IPR011043 |
|  | At5g39660 | IPR003851 |
|  | IPR000014 | R003851 \| IPR |
| No. 0548 | At1g68050 | IPR011043 |
|  | At5g42190 | IPR001680 |
|  | IPR001680 | R001810 \| IPR |
| No. 0549 | At1g68050 | IPR011043 |
|  | At5g61380 | IPR011006 |
|  | IPR001610 | R010402\|IPR |
| No. 0550 | At1g68400 | IPR001611 |
|  | At2g41090 | IPR011992 |
|  | $\begin{aligned} & \text { IPR000719 } \\ & \text { \| IPR00204 } \\ & \hline \end{aligned}$ | $\begin{aligned} & \hline \text { R002048\|IPF } \\ & \text { PR011009\|IP } \\ & \hline \end{aligned}$ |
| No. 0551 | At1g68400 | IPR001611 |
|  | At3g43810 | IPR011992 |













| No. 0817 | $\begin{aligned} & \text { At2g32950 } \\ & \text { At5g64920 } \\ & \hline \end{aligned}$ | IPR001680\|IPR001841| IPR011046 IPR001841 |
| :---: | :---: | :---: |
|  | IPR001680-IPR001841 \|IPR001841-IPR001841 |  |
| No. 0818 | At2g33610 | IPR001005 \| IPR007526| IPR009057 |
|  | At2g33610 IPR001005\|IPR007526|IPR009057 |  |
|  | IPR001005-IPR001005 \| IPR001005-IPR007526|IPR001005-IPR009057|IPR007526-IPR007526|IPR009057-IPR009057 |  |
| No. 0819 | At2g33610 | IPR001005 \| IPR007526 | IPR009057 |
|  | At2g46020 | IPR001487 \| IPR000330 | IPR001650 |
|  | IPR001487-IPR007526 |  |
| No. 0820 | At2g33610 | IPR001005 \| IPR007526 | IPR009057 |
|  | At2g47620 | IPR001005 \| IPR007526 | IPR009057 |
|  | IPR001005-IPR001005\|IPR001005-IPR007526|IPR001005-IPR009057|IPR007526-IPR007526|IPR009057-IPR009057 |  |
| No. 0821 | At2g34180 | IPR004041 \| IPR008271|IPR011009|IPR000719 |
|  | At4g26570 IPR011992 \| IPR002048 |  |
|  | IPR000719-IPR002048\| IPR000719-IPR011992 | IPR002048-IPR004041|IPR002048-IPR008271|IPR002048-IPR011009\|IPR004041-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |  |
| No. 0822 | At2g34180 | IPR004041 \| IPR008271|IPR011009 | IPR000719 |
|  | At5g55990 IPR011992\|IPR002048 |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR004041|IPR002048-IPR008271|IPR002048-IPR011009 | IPR004041-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |  |
| No. 0823 | At2g34650 | IPR008271\|IPR011009|IPR000719 |
|  | At2g41100 IPR011992\|IPR002048 |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992\|IPR011009-IPR011992 |  |
| No. 0824 | At2g34660 IPR011527\|IPR003593 |  |
|  | At3g21640 | IPR001440 \| IPR001179 |
|  | IPR001179-IPR003593\|IPR001179-IPR011527|IPR001440-IPR003593|IPR001440-IPR011527 |  |
| No. 0825 | At2g35670 IPR007087 <br> At4g02020 IPR001005 IPR001214 |  |
|  |  |  |
|  | IPR001214-IPR007087 |  |
| No. 0826 | At2g35940 IPR009057 \| IPR006563 <br> At2g35940 IPR009057 \| IPR006563 |  |
|  |  |  |
|  | IPR006563-IPR006563\|IPR006563-IPR009057|IPR009057-IPR009057 |  |
| No. 0827 | At2g35940 IPR009057\|IPR006563 |  |
|  | At2g36050 IPR006458 |  |
|  | IPR006458-IPR006563\|IPR006458-IPR009057 |  |
| No. 0828 | At2g35940 IPR009057\|IPR006563 |  |
|  | At2g40000 IPR009869\|IPR009743 |  |
|  | IPR006563-IPR009743\|IPR006563-IPR009869 |  |
| No. 0829 | At2g35940 IPR009057\|IPR006563 |  |
|  | At4g18830 IPR006458 |  |
|  | IPR006458-IPR006563 \| IPR006458-IPR009057 |  |
| No. 0830 | At2g35940 IPR009057\|IPR006563 |  |
|  | At4g32040 IPR005541\|IPR009057|IPR005540|IPR005539 |  |
|  | IPR005539-IPR006563\|IPR005539-IPR009057|IPR005540-IPR006563|IPR005540-IPR009057|IPR005541-IPR006563 | IPR005541-IPR009057|IPR006563-IPR009057|IPR009057-IPR009057 |  |
| No. 0831 | At2g35940 IPR009057\|IPR006563 |  |
|  | At4g32980 IPR009057\|IPR006563 |  |
|  | IPR006563-IPR006563 \| IPR006563-IPR009057| IPR009057-IPR009057 |  |
| No. 0832 | At2g35940 IPR009057\|IPR006563 |  |
|  | At4g34610 IPR009057\| IPR006563 |  |
|  | IPR006563-IPR006563 \| IPR006563-IPR009057|IPR009057-IPR009057 |  |
| No. 0833 | At2g35940 IPR009057 \| IPR006563 |  |
|  | At4g36870 IPR009057\|IPR006563 |  |
|  | IPR006563-IPR006563 \| IPR006563-IPR009057|IPR009057-IPR009057 |  |
| No. 0834 | $\begin{array}{ll}\text { At2g35940 } & \text { IPR009057\|IPR006563 } \\ \text { At5g01840 } & \text { IPR006458 }\end{array}$ |  |
|  |  |  |
|  | IPR006458-IPR006563\|IPR006458-IPR009057 |  |
| No. 0835 | At2g35940 IPR009057\|IPR006563 |  |
|  | At5g25220 IPR005541\|IPR009057|IPR005540|IPR005539 |  |
|  | IPR005539-IPR006563 \| IPR005539-IPR009057| IPR005540-IPR006563| IPR005540-IPR009057|IPR005541-IPR006563\| IPR005541-IPR009057| IPR006563-IPR009057| IPR009057-IPR009057 |  |
| No. 0836 | At2g35940 IPR009057\|IPR006563 |  |
|  | At5g41410 IPR009057\| IPR006563 |  |
|  | IPR006563-IPR006563\|IPR006563-IPR009057|IPR009057-IPR009057 |  |
| No. 0837 | At2g35940 IPR009057\|IPR006563 |  |
|  | At5g58360 | IPR006458 |
|  | IPR006458-IPR006563\|IPR006458-IPR009057 |  |
| No. 0838 | At2g36250 IPR003008\|IPR008280 |  |
|  | At2g36250 IPR003008\| IPR008280 |  |
|  | IPR003008-IPR003008\|IPR003008-IPR008280|IPR008280-IPR008280 |  |
| No. 0839 | At2g36250 IPR003008\|IPR008280 |  |
|  | At5g55280 IPR003008\|IPR008280 |  |
|  | IPR003008-IPR003008\| IPR003008-IPR008280 | IPR008280-IPR008280 |  |
| No. 0840 | At2g36270 IPR004827 <br> At2g36270 IPR004827 |  |
|  |  |  |
|  | At2g36270 IPR004827 |  |
| No. 0841 | At2g36350 IPR008271\| IPR011009|IPR000719 <br> At5g04510 IPR008271\| IPR011993|IPR011009|IPR000719 |  |
|  |  |  |








| No. 0986 | At3g04670 | IPR003657 |
| :---: | :---: | :---: |
|  | At5g21274 | IPR011992 \| IPR002048 |
|  | IPR002048-IPR003657\|IPR003657-IPR011992 |  |
| No. 0987 | At3g04730 | IPR011525 |
|  | At4g 14560 | IPR011525 |
|  | IPR011525-IPR011525 |  |
| No. 0988 | At3g05050 | IPR008271 \| IPR011009 | IPR000719 |
|  | At3g43810 | IPR011992 \| IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992\|IPR011009-IPR011992 |  |
| No. 0989 | At3g05050 | IPR008271\|IPR011009|IPR000719 |
|  | At3g51920 | IPR011992 \| IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992\| IPR011009-IPR011992 |  |
| No. 0990 | At3g05050 | IPR008271 \| IPR011009 | IPR000719 |
|  |  |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992 <br> \| IPR011009-IPR011992 |  |
| No. 0991 | At3g05050 | IPR008271 \| IPR011009 | IPR000719 |
|  | At4g 14640 | IPR011992 \| IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992 <br> \| IPR011009-IPR011992 |  |
| No. 0992 | At3g05050 | IPR008271\|IPR011009|IPR000719 |
|  | At5g21274 | IPR011992 \| IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992 <br> \| IPR011009-IPR011992 |  |
| No. 0993 | At3g05860 | IPR002100 |
|  | At4g36590 IPR002100-IPR002100 |  |
|  |  |  |
| No. 0994 | At3g06030 | IPR008271 \| IPR011009|IPR000719 |
|  | At3g43810 IPR011992\|IPR002048 |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992\| IPR011009-IPR011992 |  |
| No. 0995 | At3g06030 | IPR008271\|IPR011009|IPR000719 |
|  | At3g56800 IPR011992\|IPR002048 |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992 | IPR011009-IPR011992 |  |
| No. 0996 | At3g06030 IPR008271\|IPR011009|IPR000719 |  |
|  | At5g21274 | IPR011992 \| IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992\|IPR011009-IPR011992 |  |
| No. 0997 | At3g07370 IPR001440\|IPR003613 <br> At4g27960 IPR000608 |  |
|  |  |  |
|  | IPR000608-IPR003613 |  |
| No. 0998 | At3g07370 IPR001440\|IPR003613 <br> At5g41700 IPR000608 |  |
|  |  |  |
|  | IPR000608-IPR003613 |  |
| No. 0999 | At3g07370 IPR001440\|IPR003613 <br> At5g53300 IPR000608 |  |
|  |  |  |
|  | IPR000608-IPR003613 |  |
| No. 1000 | $\begin{array}{ll}\text { At3g07740 } & \text { IPR001005 \| IPR007526\| IPR009057\|IPR000433 } \\ \text { At3g54610 } & \text { IPR000182 \| IPR001487 }\end{array}$ |  |
|  |  |  |
|  | IPR000182-IPR000433\|IPR001487-IPR007526 |  |
| No. 1001 | At3g08730 IPR000961\|IPR008271|IPR011009|IPR000719 |  |
|  | At3g08730 IPR000961\|IPR008271|IPR011009|IPR000719 |  |
|  | IPR000719-IPR008271\|IPR008271-IPR008271|IPR008271-IPR011009 |  |
| No. 1002 | At3g10070 IPR009072 |  |
|  | At4g20280 | IPR009072 |
|  | IPR009072-IPR009072 |  |
| No. 1003 | At3g10070 IPR009072 |  |
|  | At4g34340 | IPR009072 \| IPR006565 |
|  | IPR006565-IPR009072\|IPR009072-IPR009072 |  |
| No. 1004 | At3g10070 IPR009072 |  |
|  | At5g25150 | IPR001680\|IPR011046|IPR007582 |
|  | IPR007582-IPR009072 |  |
| No. 1005 | At3g10070 IPR009072 |  |
|  | At5g43130 IPR007900 |  |
|  | IPR007900-IPR009072 |  |
| No. 1006 | At3g11400 IPR000504\|IPR001878 |  |
|  | At5g25780 IPR011042\|IPR000504 |  |
|  | IPR000504-IPR000504\|IPR000504-IPR001878 |  |
| No. 1007 | At3g11400 IPR000504\|IPR001878 |  |
|  | At5g27640 IPR000504 |  |
|  | IPR000504-IPR000504\|IPR000504-IPR001878 |  |
| No. 1008 | At3g11750 IPR006157 |  |
|  | $\text { At3g11750 } \quad \text { IPR006157 }$ |  |
|  | IPR006157 | 2006157 |
| No. 1009 | At3g12200 IPR008271\| IPR008266| IPR011009| IPR000719 <br> At3g43810 IPR011992\|IPR002048 |  |
|  |  |  |


|  | IPR000719-IPR002048\| IPR000719-IPR011992|IPR002048-IPR008266|IPR002048-IPR008271|IPR002048-IPR011009 | IPR008266-IPR011992 | IPR008271-IPR011992|IPR011009-IPR011992 |
| :---: | :---: |
| No. 1010 | At3g12200 IPR008271\|IPR008266|IPR011009|IPR000719 |
|  | At3g51920 IPR011992\|IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008266|IPR002048-IPR008271|IPR002048-IPR011009 | IPR008266-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1011 | At3g12200 IPR008271\|IPR008266|IPR011009|IPR000719 |
|  | At3g56800 IPR011992\|IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008266|IPR002048-IPR008271|IPR002048-IPR011009 | IPR008266-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1012 | At3g12200 IPR008271\|IPR008266|IPR011009|IPR000719 |
|  | At4g14640 IPR011992\|IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008266|IPR002048-IPR008271|IPR002048-IPR011009 |IPR008266-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1013 | At3g12200 IPR008271\|IPR008266|IPR011009|IPR000719 |
|  | At5g21274 IPR011992\|IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992 | IPR002048-IPR008266|IPR002048-IPR008271|IPR002048-IPR011009 |IPR008266-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1014 | At3g12200 IPR008271\|IPR008266|IPR011009|IPR000719 |
|  | At5g37780 IPR011992\|IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008266|IPR002048-IPR008271|IPR002048-IPR011009 <br> \| IPR008266-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1015 | At3g12250 IPR004827 |
|  | At4g19660 IPR002110\|IPR000210 |
|  | IPR002110-IPR004827 |
| No. 1016 | At3g12250 IPR004827 |
|  | At5g37780 IPR011992\|IPR002048 |
|  | IPR002048-IPR004827\|IPR004827-IPR011992 |
| No. 1017 | At3g12690 IPR008271\|IPR011009|IPR000719 |
|  | At5g04510 IPR008271\|IPR011993|IPR011009|IPR000719 |
|  | IPR000719-IPR008271\|IPR000719-IPR011993|IPR008271-IPR008271|IPR008271-IPR011009|IPR008271-IPR011993 | IPR011009-IPR011993 |
| No. 1018 | At3g13540 IPR001005\|IPR009057 |
|  | At4g00480 IPR001092 |
|  | IPR001005-IPR001092 \| IPR001092-IPR009057 |
| No. 1019 | At3g13540 IPR001005\|IPR009057 |
|  | At4g09820 IPR001092 |
|  | IPR001005-IPR001092 \| IPR001092-IPR009057 |
| No. 1020 | At3g13570 IPR000504 |
|  | At5g64200 IPR000504 |
|  | IPR000504-IPR000504 |
| No. 1021 | At3g13570 IPR000504 |
|  | At5g64220 IPR002110\|IPR000048|IPR005559 |
|  | IPR000048-IPR000504\|IPR000504-IPR005559 |
| No. 1022 | At3g14100 IPR000504 |
|  | At3g56860 IPR000504 |
|  | IPR000504-IPR000504 |
| No. 1023 | At3g14370 IPR008271\|IPR011009|IPR000719 |
|  | At5g04510 IPR008271\|IPR011993|IPR011009|IPR000719 |
|  | IPR000719-IPR008271\|IPR000719-IPR011993|IPR008271-IPR008271|IPR008271-IPR011009|IPR008271-IPR011993 |IPR011009-IPR011993 |
| No. 1024 | At3g15354 IPR001680\|IPR011046|IPR011009|IPR000719 |
|  | At5g15850 IPR010402\| IPR000315 |
|  | IPR000315-IPR001680\|IPR000315-IPR011046 | IPR010402-IPR011046 |
| No. 1025 | At3g15540 IPR011525 |
|  | At4g 14560 IPR011525 |
|  | IPR011525-IPR011525 |
| No. 1026 | At3g16620 IPR006703\|IPR005690 |
|  | At3g46740 IPR010827 |
|  | IPR005690-IPR010827\|IPR006703-IPR010827 |
| No. 1027 | At3g16857 IPR011006\|IPR006447|IPR009057 |
|  | At3g21510 IPR008207 |
|  | IPR006447-IPR008207 \| IPR008207-IPR009057|IPR008207-IPR011006 |
| No. 1028 | At3g16857 IPR011006\|IPR006447|IPR009057 |
|  | At3g29350 IPR008207 |
|  | IPR006447-IPR008207\|IPR008207-IPR009057|IPR008207-IPR011006 |
| No. 1029 | At3g16857 IPR011006\|IPR006447|IPR009057 |
|  | At5g39340 IPR008207 |
|  | IPR006447-IPR008207 \| IPR008207-IPR009057| IPR008207-IPR011006 |
| No. 1030 | At3g17210 IPR011008 |
|  | At3g17210 IPR011008 |
|  | IPR011008-IPR011008 |
| No. 1031 | At3g17510 IPR004041\|IPR008271|IPR011009|IPR000719 |
|  | At4g17615 IPR011992\|IPR002048 |
|  | IPR000719-IPR002048\| IPR000719-IPR011992 | IPR002048-IPR004041| IPR002048-IPR008271|IPR002048-IPR011009 |IPR004041-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1032 | At3g17510 IPR004041\|IPR008271|IPR011009|IPR000719 |
|  | At4g26570 IPR011992\|IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR004041|IPR002048-IPR008271|IPR002048-IPR011009 |IPR004041-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |




| No. 1079 | $\begin{aligned} & \text { At3g23310 } \\ & \text { At5g37780 } \\ & \hline \end{aligned}$ | IPR000961\|IPR008271|IPR011009|IPR000719 <br> IPR011992\|IPR002048 |
| :---: | :---: | :---: |
|  | $\begin{aligned} & \hline \text { IPR000719- } \\ & \text { \| IPR01100 } \end{aligned}$ | R002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992 PR011992 |
| No. 1080 | At3g23380 | IPR000095 |
|  | At3g51300 | IPR005225 |
|  | IPR000095 | R005225 |
| No. 1081 | At3g24140 | IPR001092 |
|  | At5g46690 | IPR001092 |
|  | IPR001092 | R001092 |
| No. 1082 | At3g24140 | IPR001092 |
|  | At5g65640 | IPR001092 |
|  | IPR001092 | R001092 |
| No. 1083 | At3g25250 | IPR000961 \| IPR008271 | IPR011009 | IPR000719 |
|  | At5g04510 | IPR008271 \| IPR011993 | IPR011009 | IPR000719 |
|  | $\begin{aligned} & \text { IPR000719- } \\ & \text { \| IPR01100 } \end{aligned}$ | $\begin{aligned} & \text { R008271\|IPR000719-IPR011993\|IPR008271-IPR008271\|IPR008271-IPR011009\|IPR008271-IPR011993 } \\ & \text { PR011993 } \end{aligned}$ |
| No. 1084 | At3g27380 | IPR006058\|IPR001041 |
|  | At5g66760 | IPR003952 \| IPR003953 | IPR004112 |
|  | IPR003952 | 2006058\|IPR003953-IPR006058|IPR004112-IPR006058 |
| No. 1085 | At3g27580 | IPR008271\|IPR011009 | IPR000719 |
|  | At5g04510 | IPR008271\|IPR011993|IPR011009 | IPR000719 |
|  | $\begin{aligned} & \hline \text { IPR000719 } \\ & \text { \|IPR01100 } \end{aligned}$ | $\begin{aligned} & \text { R008271\|IPR000719-IPR011993\|IPR008271-IPR008271\|IPR008271-IPR011009\|IPR008271-IPR011993 } \\ & \text { PR011993 } \end{aligned}$ |
| No. 1086 | At3g27920 | IPR001005 \| IPR009057 |
|  | At4g00480 | IPR001092 |
|  | IPR001005 | R001092 \| IPR001092-IPR009057 |
| No. 1087 | At3g27920 | IPR001005 \| IPR009057 |
|  | At5g41315 | IPR001092 |
|  | IPR001005 | R001092 \| IPR001092-IPR009057 |
| No. 1088 | At3g28450 | IPR001611 \| IPR011009 | IPR000719 |
|  | At3g43810 | IPR011992 \| IPR002048 |
|  | $\begin{aligned} & \hline \text { IPR000719 } \\ & \text { \| IPR01100 } \end{aligned}$ | R002048\|IPR000719-IPR011992 | IPR001611-IPR002048| IPR001611-IPR011992|IPR002048-IPR011009 PR011992 |
| No. 1089 | At3g28450 | IPR001611 \| IPR011009 | IPR000719 |
|  | At3g51920 | IPR011992 \| IPR002048 |
|  | IPR000719 <br> \| IPR01100 | R002048\|IPR000719-IPR011992|IPR001611-IPR002048|IPR001611-IPR011992|IPR002048-IPR011009 R011992 |
| No. 1090 | At3g28450 | IPR001611 \| IPR011009|IPR000719 |
|  | At3g56800 | IPR011992 \| IPR002048 |
|  | $\begin{aligned} & \text { IPR000719 } \\ & \text { \| IPR01100 } \end{aligned}$ | R002048\|IPR000719-IPR011992 | IPR001611-IPR002048| IPR001611-IPR011992|IPR002048-IPR011009 PR011992 |
| No. 1091 | At3g28450 | IPR001611\|IPR011009|IPR000719 |
|  | $\begin{aligned} & \text { IPR000719- } \\ & \text { \| IPR01100 } \end{aligned}$ | $\begin{aligned} & \hline \text { R002048\|IPR000719-IPR011992 \| IPR001611-IPR002048\|IPR001611-IPR011992\|IPR002048-IPR011009 } \\ & \text { PR011992 } \end{aligned}$ |
| No. 1092 | At3g28450 | IPR001611\|IPR011009|IPR000719 |
|  | $\begin{aligned} & \text { At5g21274 } \\ & \hline \text { IPR000719 } \\ & \text { \| IPR01100 } \end{aligned}$ | $\begin{aligned} & \text { R002048\|IPR000719-IPR011992\|IPR001611-IPR002048\|IPR001611-IPR011992\|IPR002048-IPR011009 } \\ & \text { PR011992 } \end{aligned}$ |
| No. 1093 | $\begin{aligned} & \text { At3g28450 } \\ & \text { At5g37780 } \\ & \hline \end{aligned}$ | $\begin{aligned} & \text { IPR001611\| IPR011009\| IPR000719 } \\ & \text { IPR011992 \| IPR002048 } \\ & \hline \end{aligned}$ |
|  | $\begin{aligned} & \hline \text { IPR000719- } \\ & \text { \| IPR01100 } \\ & \hline \end{aligned}$ | R002048\|IPR000719-IPR011992|IPR001611-IPR002048|IPR001611-IPR011992|IPR002048-IPR011009 | 1 -IPR011992 |
| No. 1094 | At3g28920 | IPR006455 \| IPR006456 |
|  | At3g50890 | IPR006455 \| IPR006456 |
|  | IPR006455 | 2006455 \| IPR006455-IPR006456 | IPR006456-IPR006456 |
| No. 1095 | At3g28920 | IPR006455 \| IPR006456 |
|  | At5g 15210 | IPR006455 \| IPR006456 |
|  | IPR006455 | R006455 \| IPR006455-IPR006456 | IPR006456-IPR006456 |
| No. 1096 | At3g28920 | IPR006455 \| IPR006456 |
|  | At5g65410 | IPR006455 \| IPR006456 |
|  | IPR006455 | 2006455 \| IPR006455-IPR006456|IPR006456-IPR006456 |
| No. 1097 | At3g29350 | IPR008207 |
|  | At3g48100 | IPR011006 |
|  | IPR008207 | 2011006 |
| No. 1098 | At3g29350 | IPR008207 |
|  | At3g57040 | IPR011006 |
|  | IPR008207 | 2011006 |
| No. 1099 | At3g29350 | IPR008207 |
|  | At4g16110 | IPR011006 \| IPR009057 |
|  | IPR008207 | R009057 \| IPR008207-IPR011006 |
| No. 1100 | At3g29350 | IPR008207 |
|  | At4g31920 | IPR011006 \| IPR006447 | IPR009057 |
|  | IPR006447 | 2008207 \| IPR008207-IPR009057|IPR008207-IPR011006 |
| No. 1101 | At3g29350 | IPR008207 |
|  | At5g35750 | IPR009082 \| IPR011006 | IPR005467|IPR006189|IPR003594 |
|  | IPR003594 | 2008207 \| IPR005467-IPR008207 | IPR006189-IPR008207 | IPR008207-IPR009082 | IPR008207-IPR011006 |
| No. 1102 | At3g30530 | IPR004827 |
|  | At3g30530 | IPR004827 |





| No. 1172 | At3g50890 | IPR006455 \| IPR006456 |
| :---: | :---: | :---: |
|  | At5g15210 | IPR006455 \| IPR006456 |
|  | IPR006455-IPR006455 \| IPR006455-IPR006456 | IPR006456-IPR006456 |  |
| No. 1173 | At3g50890 | IPR006455 \| IPR006456 |
|  | At5g65410 | IPR006455 \| IPR006456 |
|  | IPR006455-IPR006455 \| IPR006455-IPR006456| IPR006456-IPR006456 |  |
| No. 1174 | At3g51300 | IPR005225 |
|  | At5g 16490 | IPR000095 |
|  | IPR000095-IPR005225 |  |
| No. 1175 | At3g51920 | IPR011992 \| IPR002048 |
|  | IPR002048-IPR004827\|IPR004827-IPR011992 |  |
|  |  |  |
| No. 1176 | At3g51920 | IPR011992 \| IPR002048 |
|  | IPR002048-IPR004827\|IPR004827-IPR011992 |  |
|  |  |  |
| No. 1177 | At3g51920 | IPR011992 \| IPR002048 |
|  | At4g02410 IPR008271\|IPR008985|IPR011009|IPR001220|IPR000719 |  |
|  | IPR000719-IPR002048\| IPR000719-IPR011992|IPR001220-IPR002048|IPR001220-IPR011992|IPR002048-IPR008271 | IPR002048-IPR008985 | IPR002048-IPR011009|IPR008271-IPR011992|IPR008985-IPR011992|IPR011009-IPR01199 2 |  |
|  |  |  |
| No. 1178 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g02640 | IPR004827 |
|  | IPR002048-IPR004827\|IPR004827-IPR011992 |  |
| No. 1179 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g04570 | IPR002902 \| IPR008271| IPR011009 | IPR000719 |
|  |  | \| IPR002902-IPR011992 | IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1180 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g13020IPR000719 | IPR008271 \| IPR011009|IPR000719 |
|  |  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992 |IPR011009-IPR011992 |
| No. 1181 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g14340 | IPR008271 \| IPR011009|IPR000719 |
|  | \| IPR011009-IPR011992 |  |
| No. 1182 | At3g51920 IPR011992 \| IPR002048 |  |
|  | At4g18950 | IPR002110\|IPR011009|IPR000719 |
|  |  | R002048\|IPR000719-IPR011992|IPR002048-IPR011009|IPR011009-IPR011992 |
| No. 1183 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g23050 | IPR001610\|IPR008271 | IPR000014 | IPR011009|IPR000719 |
|  |  | $\begin{aligned} & \text { R002048\|IPR000719-IPR011992\|IPR002048-IPR008271\|IPR002048-IPR011009\|IPR008271-IPR011992 } \\ & \text { PR011992 } \end{aligned}$ |
| No. 1184 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g23170 | IPR002902 |
|  | IPR002048-IPR002902 \| IPR002902-IPR011992 |  |
| No. 1185 | At3g51920 IPR011992 \| IPR002048 |  |
|  | At4g23650 | IPR011992 \| IPR008271| IPR002048|IPR011009|IPR000719 |
|  |  | \| IPR011009-IPR011992 |
| No. 1186 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g23810 IPR003657 |  |
|  | IPR002048-IPR003657\|IPR003657-IPR011992 |  |
| No. 1187 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g26070 IPR008271\|IPR011009|IPR000719 |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992 | IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992IIPR011009-IPR011992 |  |
| No. 1188 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g30960 IPR004041\|IPR008271|IPR011009|IPR000719 |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR004041|IPR002048-IPR008271|IPR002048-IPR011009 | IPR004041-IPR011992 |IPR008271-IPR011992|IPR011009-IPR011992 |  |
| No. 1189 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g35600 IPR008271 \| IPR011009 | IPR000719 |  |
|  | IPR000719-IPR002048\| IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992\| IPR011009-IPR011992 |  |
| No. 1190 | At3g51920 IPR011992\|IPR002048 |  |
|  | At4g39540 IPR001680\|IPR000623 |  |
|  | IPR000623-IPR002048\|IPR000623-IPR011992 |  |
| No. 1191 | At3g51920 IPR011992\|IPR002048 |  |
|  | At5g01890 IPR001611\|IPR011009|IPR000719 |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR001611-IPR002048|IPR001611-IPR011992|IPR002048-IPR011009 | IPR011009-IPR011992 |  |
| No. 1192 | At3g51920 IPR011992\|IPR002048 |  |
|  | At5g10520 IPR008271\| IPR011009|IPR000719 |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992\| IPR011009-IPR011992 |  |
| No. 1193 | At3g51920 IPR011992\|IPR002048 |  |
|  | At5g15830 IPR004827 |  |
|  | IPR002048-IPR004827\|IPR004827-IPR011992 |  |
| No. 1194 | $\begin{aligned} & \hline \text { At3g51920 } \\ & \text { At5g24080 } \\ & \hline \end{aligned}$ | IPR011992 \| IPR002048 |
|  |  | IPR008271 \| IPR011009 | IPR000719 |







| 2 |  |  |
| :---: | :---: | :---: |
| No. 1317 | At4g02410 | IPR008271 \| IPR008985 | IPR011009 | IPR001220|IPR000719 |
|  | At5g37780 | IPR011992 \| IPR002048 |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR001220-IPR002048|IPR001220-IPR011992|IPR002048-IPR008271 |  |
|  | \| IPR002048-IPR $008985 \mid$ IPR002048-IPR011009\|IPR008271-IPR011992|IPR008985-IPR011992|IPR011009-IPR011992 |  |
| No. 1318 | At4g02440 | IPR001810 |
|  | At5g42190 | IPR001680 |
|  | IPR001680-IPR001810 |  |
| No. 1319 | At4g02510 | IPR006703 \| IPR005690 |
|  | At5g05000 IPR006703 |  |
|  | IPR005690-IPR006703\|IPR006703-IPR006703 |  |
| No. 1320 | At4g02510 IPR006703\|IPR005690 <br> At5g19620 IPR010827 |  |
|  |  |  |
|  | IPR005690-IPR010827\|IPR006703-IPR010827 |  |
| No. 1321 | At4g02520 IPR010987\|IPR004045 |  |
|  | At4g02520 | IPR010987\| IPR004045 |
|  |  | R004045 \| IPR004045-IPR010987 | IPR010987-IPR010987 |
| No. 1322 | At4g02640 IPR004827 |  |
|  | At4g02640 | IPR004827 |
|  |  | IPR004827-IPR004827 |
| No. 1323 | At4g02640 | IPR004827 |
|  | At4g14640 IPR011992\|IPR002048 |  |
|  |  |  |
| No. 1324 | At4g02640 IPR004827 |  |
|  | At4g34590 IPR004827 |  |
|  | IPR004827-IPR004827 |  |
| No. 1325 | At4g02640 IPR004827 <br> At5g21274 IPR011992 \| IPR002048 |  |
|  |  |  |
|  | IPR002048-IPR004827\|IPR004827-IPR011992 |  |
| No. 1326 | At4g02640 IPR004827 |  |
|  | IPR004827-IPR004827 |  |
|  |  |  |
| No. 1327 | At4g02640 IPR004827 |  |
|  | At5g28770 IPR004827 |  |
|  | IPR004827-IPR004827 |  |
| No. 1328 | At4g02640 IPR004827 <br> At5g37780 IPR011992 \| IPR002048 |  |
|  |  |  |
|  | IPR002048-IPR004827\|IPR004827-IPR011992 |  |
| No. 1329 | At4g02740 IPR001810 |  |
|  | At5g42190 | IPR001680 |
|  | IPR001680-IPR001810 |  |
| No. 1330 | At4g04570 IPR002902\|IPR008271|IPR011009|IPR000719 <br> At4g14640 IPR011992\|IPR002048 |  |
|  |  |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR002902|IPR002048-IPR008271|IPR002048-IPR011009\| IPR002902-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |  |
| No. 1331 | At4g04570 IPR002902\|IPR008271|IPR011009|IPR000719 |  |
|  |  |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR002902|IPR002048-IPR008271|IPR002048-IPR011009\|IPR002902-IPR011992 | IPR008271-IPR011992 |IPR011009-IPR011992 |  |
| No. 1332 | At4g04570 IPR002902\|IPR008271|IPR011009|IPR000719 |  |
|  | At5g37780 IPR011992\|IPR002048 |  |
|  | IPR000719-IPR002048\|IPR000719-IPR011992|IPR002048-IPR002902|IPR002048-IPR008271|IPR002048-IPR011009\| IPR002902-IPR011992|IPR008271-IPR011992|IPR011009-IPR011992 |  |
| No. 1333 | At4g05420 | IPR004871\|IPR011046|IPR000408 |
|  | At5g58760 | IPR001680 \| IPR011046 | IPR001878 |
|  | IPR001680-IPR011046 \| IPR011046-IPR011046 |  |
| No. 1334 | At4g05460 | IPR001810 |
|  | At5g42190 | IPR001680 |
|  | IPR001680-IPR001810 |  |
| No. 1335 | At4g08040 | IPR004839 \| IPR004838 |
|  | At4g08040 | IPR004839 \| IPR004838 |
|  | IPR004838-IPR004838\| IPR004838-IPR004839 | IPR004839-IPR004839 |  |
| No. 1336 | At4g08040 | IPR004839 \| IPR004838 |
|  | At4g37770 | IPR004839 \| IPR004838 |
|  | IPR004838-IPR004838\| IPR004838-IPR004839 | IPR004839-IPR004839 |  |
| No. 1337 | At4g08040 | IPR004839 \| IPR004838 |
|  | At5g65800 | IPR004839 \| IPR004838 |
|  | IPR004838-IPR004838\|IPR004838-IPR004839 | IPR004839-IPR004839 |  |
| No. 1338 | At4g08150 | IPR005541\|IPR009057|IPR005540|IPR005539 |
|  | At4g34610 | IPR009057\| IPR006563 |
|  | IPR005539-IPR006563 \| IPR005539-IPR009057| IPR005540-IPR006563|IPR005540-IPR009057|IPR005541-IPR006563\| IPR005541-IPR009057|IPR006563-IPR009057|IPR009057-IPR009057 |  |
| No. 1339 | At4g08150 IPR005541\|IPR009057|IPR005540|IPR005539 |  |
|  | At4g36870 | IPR009057\|IPR006563 |
|  | IPR005539-IPR006563 \| IPR005539-IPR009057| IPR005540-IPR006563|IPR005540-IPR009057|IPR005541-IPR006563\| IPR005541-IPR009057|IPR006563-IPR009057|IPR009057-IPR009057 |  |
| No. 1340 | At4g08150 IPR005541\|IPR009057| IPR005540|IPR005539 <br> At5g01840 IPR006458 |  |
|  |  |  |











|  | At5g49760 | IPR001611 \| IPR008271|IPR011009 | IPR000719 |
| :---: | :---: | :---: |
|  | IPR000719 <br> \| IPR00204 | R002048\|IPR000719-IPR011992|IPR001611-IPR002048|IPR001611-IPR011992|IPR002048-IPR008271 R011009|IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1553 | At5g21274 | IPR011992 \| IPR002048 |
|  | At5g53450 | IPR011009 \| IPR000719 |
|  | IPR000719 | R002048\|IPR000719-IPR011992|IPR002048-IPR011009|IPR011009-IPR011992 |
| No. 1554 | At5g21274 | IPR011992 \| IPR002048 |
|  | At5g56040 | IPR001611 \| IPR008271| IPR011009 | IPR000719 |
|  | IPR000719 | R002048\|IPR000719-IPR011992|IPR001611-IPR002048| IPR001611-IPR011992|IPR002048-IPR008271 R011009 |IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1555 | At5g21274 | IPR011992 \| IPR002048 |
|  | At5g56890 | IPR008271 \| IPR011009 | IPR000719 |
|  | $\begin{aligned} & \hline \text { IPR000719 } \\ & \text { \| IPR01100 } \\ & \hline \end{aligned}$ | R002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992 |
| No. 1556 | At5g21274 | IPR011992 \| IPR002048 |
|  | At5g63160 | IPR000210 \| IPR000197 |
|  | IPR000197 | 2002048 \| IPR000197-IPR011992 |
| No. 1557 | At5g21274 | IPR011992 \| IPR002048 |
|  | At5g66770 | IPR005202 |
|  | IPR002048 | 2005202 \| IPR005202-IPR011992 |
| No. 1558 | At5g21274 | IPR011992 \| IPR002048 |
|  | At5g66790 | IPR008271 \| IPR011009 | IPR000719 |
|  | $\begin{aligned} & \hline \text { IPR000719 } \\ & \text { \| IPR01100 } \\ & \hline \end{aligned}$ | (P002048\| IPR000719-IPR011992 | IPR002048-IPR008271| IPR002048-IPR011009|IPR008271-IPR011992 PR011992 |
| No. 1559 | At5g21274 | IPR011992 \| IPR002048 |
|  | At5g67480 | IPR000210 \| IPR000197 |
|  | IPR000197 | R002048\|IPR000197-IPR011992 |
| No. 1560 | At5g22580 | IPR011008 |
|  | At5g22580 | IPR011008 |
|  | IPR011008 | 2011008 |
| No. 1561 | At5g23170 | IPR008271 \| IPR011009 | IPR000719 |
|  | At5g37780 | IPR011992 \| IPR002048 |
|  | $\begin{aligned} & \hline \text { IPR000719 } \\ & \text { \| IPR01100 } \\ & \hline \end{aligned}$ | \| IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992 R011992 |
| No. 1562 | At5g23260 | IPR002487\|IPR002100 |
|  | At5g23260 | IPR002487\|IPR002100 |
|  | IPR002100 | R002100 \| IPR002100-IPR002487 | IPR002487-IPR002487 |
| No. 1563 | At5g24080 | IPR008271 \| IPR011009 | IPR000719 |
|  | At5g 37780 | IPR011992 \| IPR002048 |
|  | $\begin{aligned} & \text { IPR000719 } \\ & \text { \| IPR01100 } \\ & \hline \end{aligned}$ | R002048\|IPR000719-IPR011992|IPR002048-IPR008271|IPR002048-IPR011009|IPR008271-IPR011992 R011992 |
| No. 1564 | At5g24270 | IPR011992 \| IPR002048 |
|  | At5g35410 | IPR004041 \| IPR008271|IPR011009 | IPR000719 |
|  | $\begin{array}{r} \hline \text { IPR000719 } \\ \text { \| IPR00404 } \\ \hline \end{array}$ | R002048\| IPR000719-IPR011992 | IPR002048-IPR004041| IPR002048-IPR008271|IPR002048-IPR011009 R011992 | IPR008271-IPR011992 |IPR011009-IPR011992 |
| No. 1565 | At5g24270 | IPR011992 \| IPR002048 |
|  | At5g58380 | IPR004041 \| IPR008271|IPR011009 | IPR000719 |
|  | IPR000719 I IPR00404 | R002048\|IPR000719-IPR011992|IPR002048-IPR004041|IPR002048-IPR008271|IPR002048-IPR011009 R011992 |IPR008271-IPR011992|IPR011009-IPR011992 |
| No. 1566 | At5g24800 | IPR004827 |
|  | At5g24800 | IPR004827 |
|  | IPR004827 | 2004827 |
| No. 1567 | At5g24800 | IPR004827 |
|  | At5g28770 | IPR004827 |
|  | IPR004827 | 004827 |
| No. 1568 | At5g24800 | IPR004827 |
|  | At5g37780 | IPR011992 \| IPR002048 |
|  | IPR002048 | R004827\|IPR004827-IPR011992 |
| No. 1569 | At5g25150 | IPR001680 \| IPR011046 | IPR007582 |
|  | At5g25150 | IPR001680 \| IPR011046 | IPR007582 |
|  | IPR001680 | R001680\|IPR001680-IPR011046 | IPR011046-IPR011046 |
| No. 1570 | At5g25350 | IPR001810 |
|  | At5g 42190 | IPR001680 |
|  | IPR001680 | 2001810 |
| No. 1571 | At5g26870 | IPR002100 |
|  | At5g38620 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 1572 | At5g26870 | IPR002100 |
|  | At5g 49420 | IPR002100 |
|  | IPR002100 | 002100 |
| No. 1573 | At5g27050 | IPR002100 |
|  | At5g38620 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1574 | At5g27050 | IPR002100 |
|  | At5g49420 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1575 | At5g27070 | IPR002100 |
|  | At5g38620 | IPR002100 |
|  | IPR002100 | 2002100 |


| No. 1576 | At5g27070 | IPR002100 |
| :---: | :---: | :---: |
|  | At5g 49420 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1577 | At5g27070 | IPR002100 |
|  | At5g49490 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1578 | At5g27090 | IPR002100 |
|  | At5g 49490 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1579 | At5g27130 | IPR002100 |
|  | At5g38740 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1580 | At5g27130 | IPR002100 |
|  | At5g58890 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1581 | At5g27130 | IPR002100 |
|  | At5g65330 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1582 | At5g27580 | IPR002100 |
|  | At5g38620 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1583 | At5g27580 | IPR002100 |
|  | At5g 49420 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1584 | At5g27960 | IPR002100 |
|  | At5g60440 | IPR002100 |
|  | IPR002100 | 2002100 |
| No. 1585 | At5g28770 | IPR004827 |
|  | At5g28770 | IPR004827 |
|  | IPR004827 | 2004827 |
| No. 1586 | At5g35410 | IPR004041 |
|  | At5g37780 | IPR011992 |
|  | $\begin{aligned} & \hline \text { IPR000719 } \\ & \text { \| IPR00404 } \\ & \hline \end{aligned}$ | $\begin{aligned} & \text { R002048\|IPF } \\ & \text { PR011992 IF } \\ & \hline \end{aligned}$ |
| No. 1587 | At5g35550 | IPR001005 |
|  | At5g35550 | IPR001005 |
|  | IPR001005 | 2001005 \| IP |
| No. 1588 | At5g35750 | IPR009082 |
|  | At5g35750 | IPR009082 |
|  | $\begin{aligned} & \text { IPR003594 } \\ & \text { \| IPR00546 } \\ & 2 \mid \text { IPR0061 } \end{aligned}$ | $\begin{aligned} & \text { R003594 \| IPF } \\ & \text { PR006189\|IF } \\ & \text { IPR011006\|I } \end{aligned}$ |
| No. 1589 | At5g35750 | IPR009082 |
|  | At5g39340 | IPR008207 |
|  | IPR003594 | 008207\|IPR |
| No. 1590 | At5g35840 | IPR003018 |
|  | At5g35840 | IPR003018 |
|  | IPR000014 | 2000014\|IPR |
|  | \| IPR00001 | PR005467 \\| IP |
|  | 4 \| IPR0012 | PR005467। |
|  | $18 \mid \text { IPR003 }$ | -IPR003594 |
| No. 1591 | At5g37500 | IPR002110 |
|  | At5g37500 | IPR002110 |
|  | IPR000595 | 2000595 \|IPR |
| No. 1592 | At5g37780 | IPR011992 |
|  | At5g38800 | IPR004827 |
|  | IPR002048 | 004827 \| IPR |
| No. 1593 | At5g37780 | IPR011992 |
|  | At5g41920 | IPR005202 |
|  | IPR002048 | 2005202 \|IP |
| No. 1594 | At5g37780 | IPR011992 |
|  | At5g 49760 | IPR001611 |
|  | $\begin{aligned} & \text { IPR000719 } \\ & \text { \| IPR00204 } \end{aligned}$ | $\begin{aligned} & \text { R002048\|IPF } \\ & \text { PR011009 IP } \\ & \hline \end{aligned}$ |
| No. 1595 | At5g37780 | IPR011992 |
|  | At5g53450 | IPR011009 |
|  | IPR000719 | 002048\|IPR |
| No. 1596 | At5g37780 | IPR011992 |
|  | At5g56040 | IPR001611 |
|  | $\begin{aligned} & \text { IPR000719 } \\ & \text { \| IPR002048 } \end{aligned}$ | $\begin{aligned} & \text { R002048\|IPF } \\ & \text { PR011009\|IF } \end{aligned}$ |
| No. 1597 | At5g37780 | IPR011992 |
|  | At5g66770 | IPR005202 |
|  | IPR002048 | 005202 \|IPR |
| No. 1598 | At5g37780 | IPR011992 |
|  | At5g66790 | IPR008271 |
|  | IPR000719 <br> \| IPR01100 | $\begin{aligned} & \hline \text { R002048\|IPF } \\ & \text { PR011992 } \\ & \hline \end{aligned}$ |
| $\text { No. } 1599$ | At5g38620 | IPR002100 |



| No. 1626 | At5g57360 | IPR011043 \| IPR001610 |
| :---: | :---: | :---: |
|  | At5g61380 | IPR011006 \| IPR010402 |
|  | IPR001610 | R010402 \| IPR006652-IPR |
| No. 1627 | At5g65410 | IPR006455 \| IPR006456 |
|  | At5g65410 | IPR006455 \| IPR006456 |
|  | IPR006455 | R006455 \| IPR006455-IPR |
| No. 1628 | At5g65800 | IPR004839\|IPR004838 |
|  | At5g65800 | IPR004839 \| IPR004838 |
|  | IPR004838 | R004838\|IPR004838-IPR |
| No. 1629 | At5g67580 | IPR001005 \| IPR009057 |
|  | At5g67580 | IPR001005 \| IPR009057 |
|  | IPR001005 | R001005 \| IPR001005-IPR |

## B) Descriptions of IPRs (InterPro ID)

| IPR(InterPro ID) | Description |
| :---: | :---: |
| IPR000014 | PAS |
| IPR000048 | IQ calmodulin-binding region |
| IPR000061 | SWAP/Surp |
| IPR000073 | Alpha/beta hydrolase fold-1 |
| IPR000095 | PAK-box/P21-Rho-binding |
| IPR000127 | Ubiquitin-activating enzyme repeat |
| IPR000152 | EGF-type aspartate/asparagine hydroxylation conserved site |
| IPR000157 | Toll-Interleukin receptor |
| IPR000182 | GCN5-related N -acetyltransferase |
| IPR000194 | ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding |
| IPR000197 | Zinc finger, TAZ-type |
| IPR000210 | BTB/POZ-like |
| IPR000225 | Armadillo |
| IPR000283 | NADH dehydrogenase 75 kDa subunit, conserved site |
| IPR000299 | FERM domain |
| IPR000315 | Zinc finger, B-box |
| IPR000330 | SNF2-related |
| IPR000342 | Regulator of G protein signalling |
| IPR000357 | HEAT |
| IPR000375 | Dynamin central region |
| IPR000403 | Phosphatidylinositol 3- and 4-kinase, catalytic |
| IPR000408 | Regulator of chromosome condensation, RCC1 |
| IPR000433 | Zinc finger, ZZ-type |
| IPR000477 | RNA-directed DNA polymerase (reverse transcriptase) |
| IPR000504 | RNA recognition motif, RNP-1 |
| IPR000534 | Semialdehyde dehydrogenase, NAD-binding |
| IPR000571 | Zinc finger, CCCH-type |
| IPR000572 | Oxidoreductase, molybdopterin binding |
| IPR000573 | Aconitase A/isopropylmalate dehydratase small subunit, swivel |
| IPR000594 | UBA/THIF-type NAD/FAD binding fold |
| IPR000595 | Cyclic nucleotide-binding |
| IPR000608 | Ubiquitin-conjugating enzyme, E2 |
| IPR000623 | Shikimate kinase |
| IPR000626 | Ubiquitin |
| IPR000629 | RNA helicase, ATP-dependent, DEAD-box, conserved site |
| IPR000634 | Serine/threonine dehydratase, pyridoxal-phosphate-binding site |
| IPR000674 | Aldehyde oxidase and xanthine dehydrogenase, a/b hammerhead |
| IPR000683 | Oxidoreductase, N -terminal |
| IPR000700 | PAS-associated, C-terminal |
| IPR000717 | Proteasome component region PCI |
| IPR000719 | Protein kinase, core |
| IPR000727 | Target SNARE coiled-coil region |
| IPR000756 | Diacylglycerol kinase accessory region |
| IPR000793 | ATPase, F1/V1/A1 complex, alpha/beta subunit, C-terminal |
| IPR000795 | Protein synthesis factor, GTP-binding |
| IPR000836 | Phosphoribosyltransferase |
| IPR000857 | Unconventional myosin/plant kinesin-like protein/non-motor protein conserved region MyTH4 |
| IPR000873 | AMP-dependent synthetase and ligase |
| IPR000904 | SEC7-like |
| IPR000928 | SNAP-25 |
| IPR000953 | Chromo domain |
| IPR000961 | AGC-kinase, C-terminal |
| IPR000999 | Ribonuclease III |
| IPR001005 | SANT, DNA-binding |
| IPR001025 | Bromo adjacent region |
| IPR001026 | Epsin, N-terminal |
| IPR001030 | Aconitase/3-isopropylmalate dehydratase large subunit, alpha/beta/alpha |
| IPR001041 | Ferredoxin |
| IPR001048 | Aspartate/glutamate/uridylate kinase |
| IPR001092 | Basic helix-loop-helix dimerisation region bHLH |
| IPR001159 | Double-stranded RNA binding |
| IPR001179 | Peptidyl-prolyl cis-trans isomerase, FKBP-type |
| IPR001206 | Diacylglycerol kinase, catalytic region |

IPR001214
IPR001216
IPR001220
IPR001229
IPR001247
IPR001252
IPR001279
IPR001294
IPR001296
IPR001296
IPR001327
IPR001357
IPR001431
IPR001440
IPR001450
IPR001451
IPR001453
IPR001471
IPR001487
IPR001494
IPR001589
IPR001589
IPR001594
IPR001610
IPR001611
IPR001623
IPR001646
IPR001650
IPR001680
IPR001683
IPR001715
IPR001736
IPR001736
IPR001752
IPR001765
IPR001772
IPR001810
IPR001841
IPR001849
IPR001876
IPR001878
IPR001926
IPR001932
IPR001969
IPR001969
IPR002015
IPR002035
IPR002048
IPR002068
IPR002093
IPR002100
IPR002109
IPR002110
IPR002121
IPR002130
IPR002130
TPR002182
IPR002219
IPR002346
IPR002482
IPR002487
IPR002553
IPR002562
IPR002569
IPR002586
IPR002652
IPR002655
IPR002655
IPR002719
IPR002720
IPR002735
IPR002885
IPR002888
IPR002893
IPR002902
IPR002913
IPR002931
IPR002934
IPR002934
IPR003016
IPR003018
IPR003034
IPR003088
IPR003100
IPR003106
IPR003124
IPR003126
IPR003130
IPR003137

ST
Cysteine synthase/cystathionine beta-synthase P-phosphate-binding site
Legume lectin, beta domain
Mannose-binding lectin
Exoribonuclease, phosphorolytic domain 1
Malate dehydrogenase, active site
Beta-lactamase-like
Phytochrome
Glycosyl transferase, group 1
Pyridine nucleotide-disulphide oxidoreductase, NAD-binding region
BRCT
Peptidase M16, zinc-binding site
Tetratricopeptide TPR-1
$4 \mathrm{Fe}-4 \mathrm{~S}$ ferredoxin, iron-sulphur binding, subgroup
Bacterial transferase hexapeptide repeat
Molybdopterin binding
Pathogenesis-related transcriptional factor and ERF, DNA-binding
Bromodomain
Importin-beta, N -terminal
Actinin-type, actin-binding, conserved site
Zinc finger, DHHC-type
PAC motif
Leucine-rich repeat
Heat shock protein DnaJ, N-terminal
Pentapeptide repeat
DNA/RNA helicase, C-terminal
WD40 repeat
Phox-like
Calponin-like actin-binding
Phospholipase D/Transphosphatidylase
Kinesin, motor region
Carbonic anhydrase
Kinase-associated KA1
Cyclin-like F-box
Zinc finger, RING-type
Pleckstrin homology
Zinc finger, RanBP2-type
Zinc finger, CCHC-type
Pyridoxal phosphate-dependent enzyme, beta subunit
Protein phosphatase 2C-related
Peptidase aspartic, active site
Proteasome/cyclosome, regulatory subunit
von Willebrand factor, type A
Calcium-binding EF-hand
Heat shock protein Hsp20
BRCA2 repeat
Transcription factor, MADS-box
Glutaredoxin
Ankyrin
Helicase and RNase D C-terminal, HRDC
Peptidyl-prolyl cis-trans isomerase, cyclophilin-type
NB-ARC
Protein kinase C, phorbol ester/diacylglycerol binding
Molybdopterin dehydrogenase, FAD-binding
Peptidoglycan-binding Lysin subgroup
Transcription factor, K-box
Clathrin/coatomer adaptor, adaptin-like, N-terminal
3\'-5\' exonuclease
Methionine sulphoxide reductase A
Cobyrinic acid a, c-diamide synthase
Importin-alpha-like, importin-beta-binding region
Acyl-CoA oxidase, C-terminal
Retinoblastoma-associated protein, B-box
Retinoblastoma-associated protein, A-box
Translation initiation factor IF2/IF5
Pentatricopeptide repeat
[ $2 \mathrm{Fe}-2 \mathrm{~S}$ ]-binding
Zinc finger, MYND-type
Protein of unknown function DUF26
Lipid-binding START
Transglutaminase-like
Nucleotidyltransferas
Tubulin/FtsZ, GTPase
2-oxo acid dehydrogenase, lipoyl-binding site
GAF
DNA-binding SAP
Cytochrome c, class I
Argonaute and Dicer protein, PAZ
Leucine zipper, homeobox-associated
Actin-binding WH2
Zinc finger, N -recognin
Dynamin GTPase effector
Protease-associated PA

IPR003151
IPR003152
IPR003169
IPR003395
IPR003409
IPR003441
IPR003583
IPR003593
IPR003594
IPR003594
IPR003612
IPR003613
IPR003657
IPR003823
IPR003851
IPR003903
IPR003952
IPR003953
IPR004014
IPR004041
IPR004045
IPR004045
IPR004099
IPR004100
IPR004104
IPR004112
IPR004125
IPR004148
IPR004154
IPR004176
IPR004192
IPR004201
IPR004201
IPR004274
IPR004330
IPR004367
IPR004827
IPR004837
IPR004838
IPR004839
IPR004843
IPR004853
IPR004871
IPR004883
IPR004883
IPR005034
IPR005066
IPR005101
IPR005107
IPR005110
IPR005111
IPR005123
IPR005160
IPR005172
IPR005202
IPR005225
IPR005225
IPR005467
IPR005474
IPR005475
IPR005512
IPR005539
IPR005540
IPR005541
IPR005559
IPR005690
IPR005735
IPR005735
IPR005806
IPR005834
IPR005843
IPR005844
IPR005844
IPR005845
IPR005845
IPR005846
IPR006016
IPR006050
IPR006058
IPR006067
IPR006067
IPR006094
IPR006121
IPR006137
IPR006151
IPR006157
IPR006162
IPR006164
IPR006189
IPR006195
IPR006447

PIK-related kinase, FAT
PIK-related kinase, FATC
GYF
RecF/RecN/SMC protein, N-terminal
MORN motif
No apical meristem (NAM) protein
Helix-hairpin-helix DNA-binding motif, class 1
ATPase, AAA+ type, core
ATP-binding region, ATPase-like
Plant lipid transfer protein/seed storage/trypsin-alpha amylase inhibitor
U box
DNA-binding WRKY
Protein of unknown function CP12
Zinc finger, Dof-type
Ubiquitin interacting motif
Fumarate reductase/succinate dehydrogenase, FAD-binding site
Fumarate reductase/succinate dehydrogenase flavoprotein, N -terminal
ATPase, P-type cation-transporter, N-terminal
NAF
Glutathione S-transferase, N -terminal
Pyridine nucleotide-disulphide oxidoreductase, dimerisation
ATPase, F1/V1/A1 complex, alpha/beta subunit, N -terminal
Oxidore,
Fumarate reductase/succinate dehydrogenase flavoprotein, C-terminal
Signal recognition particle, SRP54 subunit, M-domain
BAR
Anticodon-binding
Clp, N -terminal
Ubiquinol cytochrome reductase transmembrane region
Cell division protein $48, \mathrm{CDC} 48$, domain 2
NLI interacting factor
Transcription factor, FAR1-related
Cyclin, C-terminal
Basic-leucine zipper (bZIP) transcription factor
Sodium/calcium exchanger membrane region
Aminotransferases, class-I, pyridoxal-phosphate-binding site
Aminotransferase, class I and II
Metallophosphoesterase
Protein of unknown function DUF250
Cleavage and polyadenylation specificity factor, A subunit, C-terminal
Lateral organ boundaries, LOB
Dicer double-stranded RNA-binding fold
Moybdenum cofactor oxidoreductase, dimerisation
DNA photolyase, FAD-binding/Cryptochrome, C-terminal
CO dehydrogenase flavoprotein, C-terminal
MoeA, N-terminal, domain I and II
MoeA, C-terminal, domain IV
2OG-Fe(II) oxygenase
Ku70/Ku80 C-terminal arm
Tesmin/TSO1-like, CXC
GRAS transcription factor
Small GTP-binding protein
Signal transduction histidine kinase, core
Transketolase, N -terminal
Transketolase, central region
Rop nucleotide exchanger, PRONE
ELK
KNOX1
KNOX2
CG-1
Chloroplast protein import component Toc86/159
Zinc finger, LSD1-type
Rieske [2Fe-2S] region
Haloacid dehalogenase-like hydrolase
Alpha-D-phosphohexomutase, C-terminal
Alpha-D-phosphohexomutase, alpha/beta/alpha domain I
Alpha-D-phosphohexomutase, alpha/beta/alpha domain II
Alpha-D-phosphohexomutase, alpha/beta/alpha domain III
UspA
DNA photolyase, N-terminal
$2 \mathrm{Fe}-2 \mathrm{~S}$ ferredoxin, iron-sulphur binding site
Nitrite and sulphite reductase $4 \mathrm{Fe}-4 \mathrm{~S}$ region
Acyl-CoA dehydrogenase/oxidase, central region
FAD linked oxidase, N -terminal
Heavy metal transport/detoxification protein
NADH ubiquinone oxidoreductase, 20 kDa subunit
Quinate/shikimate 5-dehydrogenase/glutamyl-tRNA reductase
Dihydroneopterin aldolase
Phosphopantetheine attachment site
DNA helicase, ATP-dependent, Ku type
CHASE
Aminoacyl-tRNA synthetase, class II, conserved region
Myb-like DNA-binding region, SHAQKYF class

IPR006455
IPR006456
IPR006458
IPR006502
IPR006527
IPR006562
IPR006563
IPR006565
IPR006566
IPR006566
IPR006594
IPR006597
IPR006652
IPR006656
IPR006671
IPR006703
IPR006751
IPR006785
IPR006845
IPR007012
IPR007051
IPR007051
IPR007087
IPR007192
IPR007201
IPR007281
IPR007365
IPR007484
IPR007517
IPR007526
IPR007527
IPR007529
IPR007529
IPR007582
IPR007699
IPR007854
IPR007856
IPR007863
IPR007863
IPR007900
IPR007900
IPR008138
IPR008139
IPR008146
IPR008147
IPR008152
IPR008152
IPR008207
IPR008210
IPR008250
IPR008266
IPR008271
IPR008274
IPR008280
IPR008284
IPR008778
IPR008889
IPR008905
TPR008905
IPR008928
IPR008942
IPR008973
IPR008974
IPR008978
IPR008978
IPR008984
IPR008984
IPR008985
IPR008991
IPR009000
IPR009010
IPR009014
IPR009025
IPR009028
IPR009030
IPR009036
IPR009056
IPR009057
IPR009065
IPR009067
IPR009071
IPR009072
IPR009075
IPR009076
IPR009082
IPR009097
IPR009100
IPR009743
IPR009869
IPR010399
IPR010402
IPR010493

Homeobox domain, ZF-HD class
ZF-HD homeobox protein Cys/His-rich dimerisation region
Protein of unknown function DUF623, plant
Protein of unknown function DUF506, plant
F-box associated
HSA
POX
Bromodomain transcription factor
FBD-like
LisH dimerisation motif
Sel1-like
Kelch repeat type 1
Molybdopterin oxidoreductase
Cyclin, N -terminal
AIG1
TAFII55 protein conserved region
Peroxisome membrane anchor protein Pex14p, N-terminal
Pex, N-terminal
Poly(A) polymerase, central region
CHORD
Zinc finger, C2H2-type
Cde 23
RNA recognition motif 2
Mre11, DNA-binding
Transferrin receptor-like, dimerisation
Peptidase M28
Rad50 zinc hook
SWIRM
Zinc finger, SWIM-type
Zinc finger, HIT-type
WD40 associated region in TFIID subunit
SGS
Fip1
Saposin-like type B, 1
Peptidase M16, C-terminal
Transcription initiation factor TFIID component TAF4
Saposin-like type B, 2
Saposin B
Glutamine synthetase, catalytic region
Glutamine synthetase, beta-Grasp
Clathrin adaptor, alpha/beta/gamma-adaptin, appendage, Ig-like subdomain
Signal transduction histidine kinase, phosphotransfer (Hpt) region
Phosphoenolpyruvate carboxykinase, N-terminal
ATPase, P-type, ATPase-associated region
Tyrosine protein kinase, active site
Serine/threonine protein kinase, active site
Aldehyde oxidase and xanthine dehydrogenase, molybdopterin binding
Tubulin/FtsZ, C-terminal
Molybdenum cofactor biosynthesis, conserved site
Pirin, C-terminal
VQ
Eukaryotic translation initiation factor 3 subunit 8, N -terminal
Six-hairpin glycosidase-like
ENTH/VHS
C2 calcium/lipid-binding region, CaLB
TRAF-like
HSP20-like chaperone
SMAD/FHA domain
Concanavalin A-like lectin/glucanase
Translation protein SH3-like
Translation elongation and initiation factors/Ribosomal, beta-barrel
Aspartate decarboxylase-like fold
Transketolase, C-terminal/Pyruvate-ferredoxin oxidoreductase, domain II
DNA-directed RNA polymerase, RBP11-like
Clathrin/coatomer adaptor, adaptin-like, appendage, C-terminal subdomain
Growth factor, receptor
Molybdenum cofactor biosynthesis, MoeB
Cytochrome c, monohaem
Homeodomain-like
FERM, 3-helical bundle
TAFII-230 TBP-binding
High mobility group, superfamily
Histone-fold
Acyl-CoA dehydrogenase/oxidase C-terminal
FKBP12-rapamycin-associated protein, FKBP12-rapamycin-binding
Signal transduction histidine kinase, homodimeric
RNA ligase/cyclic nucleotide phosphodiesterase
Acyl-CoA dehydrogenase/oxidase, middle and N -terminal
Hs 1 pro-1, C-terminal
Hs 1 pro-1, N-terminal
Tify
CCT domain
Serine acetyltransferase, N-terminal

IPR010525
IPR010544
IPR010625
IPR010827
IPR010916
IPR010987
IPR010989
IPR010995
IPR011001
IPR011001
IPR011004
IPR011006
IPR011008
IPR011009
IPR011011
IPR011013
IPR011025
IPR011028
IPR011032
IPR011037
IPR011039
IPR011039
IPR011042
IPR011043
IPR011046
IPR011053
IPR011060
IPR011068
IPR011108
IPR011124
IPR011261
IPR011262
IPR011262
IPR011442
IPR011498
IPR011525
IPR011527
IPR011538
IPR011545
IPR011989
IPR011991
IPR011992
IPR011993

Auxin response factor
Kinesin-related
CHCH
Surface antigen variable number
TonB box, conserved site
Glutathione S-transferase, C-terminal-like
t-SNARE
DNA repair Rad51/transcription factor NusA, alpha-helical
Saposin-like
Trimeric LpxA-like
CheY-like
Dimeric alpha-beta barrel
Protein kinase-like
Zinc finger, FYVE/PHD-type
Glycoside hydrolase-type carbohydrate-binding
G protein alpha subunit, helical insertion
Cyclin-like
GroES-like
Pyruvate kinase, beta-barrel-like
Transcription Factor IIF, Rap30/Rap74, interaction
Six-bladed beta-propeller, TolB-like
Galactose oxidase/kelch, beta-propeller
WD40 repeat-like
Single hybrid motif
Ribulose-phosphate binding barrel
Nucleotidyltransferase, class I, C-terminal-like
RNA-metabolising metallo-beta-lactamase
Zinc finger, CW-type
DNA-directed RNA polymerase, dimerisation
DNA-directed RNA polymerase, insert
Protein of unknown function DUF1546
Kelch repeat type 2
Aux/IAA-ARF-dimerisation
ABC transporter, transmembrane region, type 1
NADH-ubiquinone oxidoreductase, 51 kDa subunit
DNA/RNA helicase, DEAD/DEAH box type, N -terminal
Armadillo-like helical
Winged helix repressor DNA-binding
EF-Hand type
Pleckstrin homology-type

## Appendix Table 2

Predicted protein complexes with statistically significant domain-domain interaction information.
cluster number (number of members:)
members member proteins in the cluster
InterPro IDs in the protein (description of InterPro)[number of InterPro 'domains' in the protein: location of the
'domain'(',' delimited)]
PPIs list of PPIs in the cluster: (*) indicate the PPI is supported by statistically extracted DDIs
$\begin{array}{ll}\text { PPIs } & \text { list of PPIs in the cluster: }{ }^{(*)} \text { indicate the PPI is } \\ \text { DDIs list of DDIs, which support PPIs in the cluster. }\end{array}$
cluster 1 (number of members: 14)
members At1g59530
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1:48-97]
At2g41090
IPR002048(Calcium-binding EF-hand)[4: 12-40, 48-76, 82-110, 118-146]
IPR011992(EF-Hand type)[1:2-143]
At3g43810
IPR002048(Calcium-binding EF-hand)[4: 12-40, 48-76, 85-113, 121-149]
IPR011992(EF-Hand type)[1:2-146]
At3g51920
IPR002048(Calcium-binding EF-hand)[4: 12-40, 48-76, 85-113, 121-149]
IPR011992(EF-Hand type)[1:2-146]
At3g56800
IPR002048(Calcium-binding EF-hand)[4: 12-40, 48-76, 85-113, 121-149] IPR011992(EF-Hand type)[1: 2-146]
At3g58780
IPR002100(Transcription factor, MADS-box)[1: 17-114]
IPR002487(Transcription factor, K-box)[1: 102-192]
At3g62420
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1: 23-86]
At4g02640
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1: 221-276]
At4g37940
IPR002100(Transcription factor, MADS-box)[1: 1-61]

IPR002487(Transcription factor, K-box)[1: 86-176]
At5g 13790
IPR002100(Transcription factor, MADS-box) [1: 1-60]
IPR002487(Transcription factor, K -box) [1: 80-170]
At5g21274
IPR002048(Calcium-binding EF-hand)[4: 12-40, 48-76, 85-113, 121-149] IPR011992(EF-Hand type)[1: 2-146]
At5g24800
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1: 120-166]
At5g37780
IPR002048(Calcium-binding EF-hand)[4: 12-40, 48-76, 85-113, 121-149]
IPR011992(EF-Hand type)[1: 2-146]

## At5g49450

PPIs At1g59530-At2g41090 (*), At1g59530-At3g43810 (*), At1g59530-At3g51920 (*), At1g59530-At3g56800 (*),
At1g59530-At5g21274 (*), At1g59530-At5g37780 (*), At1g59530-At5g49450, At2g41090-At3g58780, At2g41090-At3g62420 (*), At2g41090-At4g02640 (*), At2g41090-At4g37940, At2g41090-At5g13790, At2g41090-At5g24800 (*), At2g41090-At5g49450, At3g43810-At3g43810, At3g43810-At3g51920, At3g43810-At3g56800, At3g43810-At3g58780, At3g43810-At3g62420 (*), At3g43810-At4g02640 (*), At3g43810-At4g37940, At3g43810-At5g13790, At3g43810-At5g21274, At3g43810-At5g24800 (*), At3g43810-At5g49450, At3g51920-At3g56800, At3g51920-At3g58780, At3g51920-At3g62420 (*), At3g51920-At4g02640 (*), At3g51920-At4g37940, At3g51920-At5g13790, At3g51920-At5g21274, At3g51920-At5g24800 (*), At3g51920-At5g49450, At3g56800-At3g56800, At3g56800-At3g58780, At3g56800-At3g62420 (*), At3g56800-At4g02640 (*), At3g56800-At4g37940, At3g56800-At5g13790, At3g56800-At5g21274, At3g56800-At5g24800 (*), At3g56800-At5g49450, At3g58780-At3g58780 (*), At3g58780-At4g37940 (*), At3g58780-At5g13790 (*), At3g58780-At5g21274, At3g58780-At5g37780, At3g62420-At3g62420 (*), At3g62420-At4g02640 (*), At3g62420-At5g21274 (*), At3g62420-At5g24800 (*), At3g62420-At5g37780 (*), At3g62420-At5g49450, At4g02640-At4g02640 (*), At4g02640-At5g21274 (*), At4g02640-At5g24800 (*), At4g02640-At5g37780 (*), At4g02640-At5g49450, At4g37940-At4g37940 (*), At4g37940-At5g13790 (*), At4g37940-At5g21274, At4g37940-At5g37780, At5g13790-At5g13790 (*) At5g13790-At5g21274, At5g13790-At5g37780, At5g21274-At5g21274, At5g21274-At5g24800 (*), At5g21274-At5g49450, At5g24800-At5g24800 (*), At5g24800-At5g37780 (*), At5g24800-At5g49450, At5g37780-At5g49450, At5g49450-At5g49450
DDIs IPR002048-IPR004827, IPR002100-IPR002100, IPR002100-IPR002487, IPR002487-IPR002487,
IPR004827-IPR004827, IPR004827-IPR011992
cluster 2 (number of members: 14)
members At1g06920
IPR006458(Protein of unknown function DUF623, plant)[1: 245-309]
At1g19700
IPR006563(POX) [1: 164-300]
IPR009057(Homeodomain-like)[1: 350-427]
At1g23380
IPR005539(ELK)[1: 226-247]
IPR005540(KNOX1)[1: 83-129]
IPR005541(KNOX2)[1: 136-187]
IPR009057(Homeodomain-like)[1: 247-324]
At1g26260
IPR001092(Basic helix-loop-helix dimerisation region bHLH)[1: 223-280]
At1g75410
IPR006563(POX)[1: 166-298]
IPR009057(Homeodomain-like)[1: 345-422]
At2g23760
IPR006563(POX)[1: 236-371]
IPR009057(Homeodomain-like)[1: 423-500]
At2g30400
IPR006458(Protein of unknown function DUF623, plant)[1: 260-319]
At2g35940
IPR006563(POX)[1: 188-333]
IPR009057(Homeodomain-like) [1: 384-461]
At4g32040
IPR005539(ELK)[1: 281-302]
IPR005540(KNOX1)[1: 116-160]
IPR005541(KNOX2)[1: 167-222]
IPR009057(Homeodomain-like)[1: 302-379]
At4g34610
IPR006563(POX)[1: 139-269]
IPR009057(Homeodomain-like)[1:313-390]
At4g36870
IPR006563(POX)[1: 311-445]
IPR009057(Homeodomain-like)[1: 496-573]
At5g01840
IPR006458(Protein of unknown function DUF623, plant)[1: 206-265]
At5g25220
IPR005539(ELK)[1: 322-343]
IPR005540(KNOX1)[1: 157-201]
IPR005541(KNOX2)[1: 213-268]
At5g41410
IPR006563(POX)[1: 192-338]
IPR009057(Homeodomain-like)[1: 390-467]
PPIs At1g06920-At1g19700 (*), At1g06920-At1g23380 (*), At1g06920-At1g75410 (*), At1g06920-At2g35940 (*),
At1g06920-At4g32040 (*), At1g06920-At4g34610 (*), At1g06920-At4g36870 (*), At1g06920-At5g25220 (*)



At2g45660-At3g58780 (*),
At2g45660-At5g15800 (*),
At3g57230-At4g09960 (*),
At3g57230-At4g37940 (*),
At3g58780-At4g11880 (*),
At3g58780-At5g15800 (*),
At4g09960-At4g18960 (*),
At4g11880-At4g18960 (*),
At4g 18960-At4g24540 (*)
At4g18960-At5g60910 (*)
At4g24540-At5g60910 (*),
At4g24540-At5g60910 (*),
At5g15800-At5g23260 (*),
At5g15800-At5g23260 (*),

At2g45660-At4g11880 (*), At2g45660-At5g60910 (*) At3g57230-At4g 11880 (*) At3g57230-At5g15800 (*), At3g58780-At4g18960 (*), At3g58780-At5g23260 (*), At4g09960-At4g37940 (*), At4g11880-At4g24540 (*) At 4 g 18960-At4g37940 (*) At4g24540-At4g24540 (*) At4g37940-At4g $47940^{(*)}$

At2g45660-At4g24540 (*), At3g57230-At3g57230 (*), At3g57230-At4g 18960 (*), At3g58780-At3g58780 (*), At3g58780-At4g24540 (*), At4g09960-At4g09960 (*), At4g09960-At5g15800 (*), At4g11880-At5g60910 (*), At4g18960-At5g15800 (*) At4g24540-At4g37940 (*) At4g24540-At4g37940 (*), At4g37940-At5g15800 (*)

At2g45660-At4g37940 (*) At3g57230-At3g58780 (*) At3g57230-At4g24540 (*) At3g58780-At4g09960 (*) At3g58780-At4g37940 (*) At4g09960-At4g11880 (*) At4g09960-At5g23260 (*) At 4 g 18960 -At 4 g 18960 (*) At4g18960-At5g23260 (*) At4924540-At5g15800 (*) At4g24540-At5g15800 (*)

| DDIs | IPR002100-IPR002100, IPR002100-IPR002487, IPR002487-IPR002487 |
| :---: | :---: |
| cluster 6 (number of members: 12) |  |
| members | At1g02680 <br> IPR009072 (Histone-fold) [1: 30-74] <br> At1g 17440 <br> IPR009072(Histone-fold)[1:525-598] <br> At1g27720 <br> IPR007900(Transcription initiation factor TFIID component TAF4)[1: 463-714] <br> At2g 18000 <br> At3g 10070 <br> IPR009072(Histone-fold) [1: 397-470] <br> At4g20280 <br> IPR009072(Histone-fold) [1: 112-201] <br> At4g31720 <br> At4g34340 <br> IPR006565(Bromodomain transcription factor)[1: 24-100] <br> IPR009072(Histone-fold)[1: 29-96] <br> At5g25150 <br> IPR001680(WD40 repeat)[6: 341-380, 406-446, 449-488, 491-530, 533-572, 575-614] <br> IPR007582(WD40 associated region in TFIID subunit)[1: 44-193] <br> IPR011046(WD40 repeat-like)[1: 42-666] <br> At5g 43130 <br> IPR007900(Transcription initiation factor TFIID component TAF4)[1:537-823] <br> At5g 45600 <br> At5g58470 <br> IPR000504(RNA recognition motif, RNP-1)[1: 280-371] <br> IPR001876(Zinc finger, RanBP2-type)[1: 84-115] |
| PPIs |  |
| DDIs | IPR001680-IPR001680, IPR001680-IPR011046, IPR001876-IPR007900, IPR006565-IPR009072, <br> IPR007582-IPR009072, IPR007900-IPR007900, IPR007900-IPR009072, IPR009072-IPR009072, <br> IPR011046-IPR011046    |
| cluster 7 (number of members: 10 ) |  |
| members | At1g 14440 <br> IPR006455(Homeobox domain, ZF-HD class)[1: 218-275] <br> IPR006456(ZF-HD homeobox protein Cys/His-rich dimerisation region)[1: 89-140] <br> At1g26920 <br> At1g69600 <br> IPR006455(Homeobox domain, ZF-HD class)[1: 156-213] <br> IPR006456(ZF-HD homeobox protein Cys/His-rich dimerisation region)[1:30-83] <br> At1g75240 <br> IPR006455(Homeobox domain, ZF-HD class)[1: 240-297] <br> IPR006456(ZF-HD homeobox protein Cys/His-rich dimerisation region)[1: 75-126] <br> At2g 18350 <br> IPR006455(Homeobox domain, ZF-HD class)[1: 198-255] <br> IPR006456(ZF-HD homeobox protein Cys/His-rich dimerisation region)[1: 81-132] <br> At2g36610 <br> IPR009057(Homeodomain-like)[1: 56-140] <br> At3g28920 <br> IPR006455(Homeobox domain, ZF-HD class)[1: 192-249] <br> IPR006456(ZF-HD homeobox protein Cys/His-rich dimerisation region)[1:51-104] At3g50890 <br> IPR006455(Homeobox domain, ZF-HD class)[1: 181-238] <br> IPR006456(ZF-HD homeobox protein Cys/His-rich dimerisation region)[1: 59-110] At5g 15210 <br> IPR006455(Homeobox domain, ZF-HD class)[1: 179-236] |

IPR006456(ZF-HD homeobox protein Cys/His-rich dimerisation region)[1:55-108]

## At5965410

IPR006455(Homeobox domain, ZF-HD class)[1:191-248]
IPR006456(ZF-HD homeobox protein Cys/His-rich dimerisation region) [1: 74-125]

| PPIs |  |
| :---: | :---: |
| DDIs | IPR006455-IPR006455, IPR006455-IPR006456, IPR006456-IPR006456, IPR009057-IPR009057 |
| cluster 8 (number of members: 11) |  |
| members | ```At1g02090 IPR000717(Proteasome component region PCI)[1: 86-177] At1g22920 At1g71230 At2g26930 At3g61140 IPR000717(Proteasome component region PCI)[1: 329-412] At4g 14110 At4g26430 At5g 14250 IPR000717(Proteasome component region PCI)[1: 293-383] IPR011991(Winged helix repressor DNA-binding) [1: 287-371] At5g 42970 IPR000717(Proteasome component region PCI)[1: 263-362] IPR011991(Winged helix repressor DNA-binding)[1: 288-371] At5g 46210 IPR011991(Winged helix repressor DNA-binding) [1: 690-792] At5g56280``` |
| PPIs |  |
| DDIs | IPR000717-IPR000717, IPR000717-IPR011991, IPR011991-IPR011991 |
| cluster 9 (number of members: 9) |  |
| members | At1g13180 <br> At1g 17550 <br> IPR001932(Protein phosphatase 2C-related)[1: 222-501] <br> At1g29170 <br> At1g60430 <br> At2g22640 <br> At2g34150 <br> At2g38440 <br> IPR003124(Actin-binding WH2)[1: 1335-1353] <br> At4g 16340 <br> At4g26080 <br> IPR001932(Protein phosphatase 2C-related)[1: 143-422] |
| PPIs | At1g13180-At1g17550, At1g13180-At1g29170, At1g13180-At2g22640, At1g13180-At4g26080, <br> At1g17550-At1g29170, At1g17550-At1g60430, At1g17550-At2g22640, At1g17550-At2g34150, <br> At1g17550-At2g38440, At1g17550-At4g16340, At1g29170-At1g60430, At1g29170-At2g22640, <br> At1g29170-At4g16340, At1g29170-At4g26080, At1g60430-At2g22640, At1g60430-At2g34150, <br> At1g60430-At2g38440, At1g60430-At4g26080, At2g22640-At2g22640, At2g22640-At2g34150, <br> At2g22640-At2g38440, At2g22640-At4g26080, At2g34150-At4g16340, At2g34150-At4g26080, <br> At2g38440-At4g16340, At2g38440-At4g26080, At4g16340-At4g16340, At4g16340-At4g26080 |
| DDIs |  |
| cluster 10 (number of members: 6) |  |
| members | ```At1g02340 IPR001092(Basic helix-loop-helix dimerisation region bHLH)[1: 139-189] At1g09530 IPR001092(Basic helix-loop-helix dimerisation region bHLH)[1: 340-393] At1g09570 IPR000014(PAS) [2: 618-688, 748-803] IPR001680(WD40 repeat)[1:633-647]``` |

IPR003018(GAF)[1: 218-412]
IPR003594(ATP-binding region, ATPase-like)[1: 1007-1119]
IPR005467(Signal transduction histidine kinase, core)[1: 902-1119]

## At2g18790

IPR000014(PAS)[2: 654-721, 785-855]
IPR001294(Phytochrome)[10: 167-189, 267-286, 352-373, 463-483, 548-567, 581-599, 655-671, 674-689, 748-765,
768-788]
IPR003018(GAF)[1: 252-433]
IPR003594(ATP-binding region, ATPase-like)[1: 1039-1152]
IPR005467(Signal transduction histidine kinase, core)[1: 934-1153]
At2g32950
IPR001680(WD40 repeat) [6: 355-399, 410-449, 459-491, 493-534, 545-585, 633-672]
IPR001841(Zinc finger, RING-type)[1:52-90]
IPR011046(WD40 repeat-like) [1: 1-670]
At2g43010
IPR001092(Basic helix-loop-helix dimerisation region bHLH)[1: 263-312]

| PPIs |  |
| :---: | :---: |
| DDIs | IPR000014-IPR000014, IPR000014-IPR001092, IPR000014-IPR001294, IPR000014-IPR003018, <br> IPR000014-IPR003594, IPR000014-IPR005467, IPR001092-IPR001092, IPR001092-IPR001294, <br> IPR001092-IPR003018, IPR001092-IPR003594, IPR001092-IPR005467, IPR001294-IPR001294, <br> IPR001294-IPR003018, IPR001294-IPR003594, IPR001294-IPR005467, IPR001680-IPR001680, <br> IPR001680-IPR001841, IPR001680-IPR011046, IPR001841-IPR001841, IPR003018-IPR003018, <br> IPR003018-IPR003594, IPR003018-IPR005467, IPR003594-IPR003594, IPR003594-IPR005467, <br> IPR005467-IPR005467, IPR011046-IPR011046,   |
| cluster 11 (number of members: 7) |  |
| members | At1g27320 <br> IPR003594(ATP-binding region, ATPase-like)[1: 562-722] <br> IPR005467(Signal transduction histidine kinase, core)[1: 457-723] <br> IPR006189(CHASE)[1: 163-389] <br> IPR009082(Signal transduction histidine kinase, homodimeric)[1: 440-511] <br> IPR011006(CheY-like)[2: 738-872, 889-1029] <br> At2g01760 <br> IPR006447(Myb-like DNA-binding region, SHAQKYF class)[1: 200-255] <br> IPR009057(Homeodomain-like)[1: 196-259] <br> IPR011006(CheY-like)[1: 8-145] <br> At3g21510 <br> IPR008207(Signal transduction histidine kinase, phosphotransfer (Hpt) region)[1: 16-138] <br> At3g29350 <br> IPR008207(Signal transduction histidine kinase, phosphotransfer (Hpt) region)[1: 40-147] <br> At4g 16110 <br> IPR009057(Homeodomain-like)[1: 217-276] <br> IPR011006(CheY-like)[1: 25-153] <br> At5g 35750 <br> IPR003594(ATP-binding region, ATPase-like)[1: 699-866] <br> IPR005467(Signal transduction histidine kinase, core) [1: 594-862] <br> IPR006189(CHASE)[1: 302-526] <br> IPR009082(Signal transduction histidine kinase, homodimeric)[1: 577-648] <br> IPR011006(CheY-like)[2: 883-1013, 1032-1174] <br> At5g39340 <br> IPR008207(Signal transduction histidine kinase, phosphotransfer (Hpt) region) [1: 18-141] |
| PPIs | At1g27320-At3g21510 (*), At1g27320-At3g29350 (*), At1g27320-At5g35750 (*), At1g27320-At5g39340 (*),  <br> At2g01760-At2g01760 (*), At2g01760-At3g21510 (*), At2g01760-At3g29350 (*), At2g01760-At4g16110 (*),  <br> At2g01760-At5g35750 (*), At2g01760-At5g39340 (*), At3g21510-At4g16110 (*), At3g21510-At5g35750 (*),  <br> At3g29350-At4g16110 (*), At3g29350-At5g35750 (*), At4g16110-At5g39340 (*), At5g35750-At5g35750 (*),  <br> At5g35750-At5g39340 (*)     <br>      |
| DDIs | IPR003594-IPR003594, IPR003594-IPR005467, IPR003594-IPR006189, IPR003594-IPR008207, <br> IPR003594-IPR009082, IPR005467-IPR005467, IPR005467-IPR006189, IPR005467-IPR008207, <br> IPR005467-IPR009082, IPR005467-IPR011006, IPR006189-IPR006189, IPR006189-IPR008207, <br> IPR006189-IPR009082, IPR006189-IPR011006, IPR006447-IPR006447, IPR006447-IPR008207, <br> IPR006447-IPR011006, IPR008207-IPR009057, IPR008207-IPR009082, IPR008207-IPR011006, <br> IPR009057-IPR009057, IPR009082-IPR009082, IPR009082-IPR011006, IPR011006-IPR011006 |
| cluster 12 (number of members: 5) |  |
| members | At1g53720 <br> IPR000504(RNA recognition motif, RNP-1)[1: 243-321] <br> IPR001878(Zinc finger, CCHC-type)[1:342-355] <br> IPR002130(Peptidyl-prolyl cis-trans isomerase, cyclophilin-type)[1: 3-162] <br> At1g55310 <br> IPR000504(RNA recognition motif, RNP-1)[1: 37-110] <br> At2g37340 <br> IPR000504(RNA recognition motif, RNP-1)[1: 11-81] <br> IPR001878(Zinc finger, CCHC-type)[2:100-114, 122-138] <br> At3g 54460 <br> IPR000330(SNF2-related) [1: 756-867] <br> IPR001650(DNA/RNA helicase, C-terminal)[1: 1212-1289] <br> IPR001810(Cyclin-like F-box)[1: 288-328] <br> IPR001841(Zinc finger, RING-type)[1: 1048-1082] |

IPR011124(Zinc finger, CW-type)[1:575-621]
At5g 18810
IPR000504(RNA recognition motif, RNP-1)[1:47-125]

cluster 15 (number of members: 5)

PR001005(SANT, DNA-binding) 2 : 9-59, 62-110]
IPR009057(Homeodomain-like)[2: 10-79, 82-125]
At1g63650
IPR001092(Basic helix-loop-helix dimerisation region bHLH)[1: 407-456]
At5g24520
IPR001680(WD40 repeat)[4: 64-109, 116-161, 164-202, 253-293]
IPR011046(WD40 repeat-like)[1:21-335]
At5g35550
IPR001005(SANT, DNA-binding) [2: 15-65, 68-116]
IPR009057(Homeodomain-like)[1: 16-85]
At5g 41315
IPR001092(Basic helix-loop-helix dimerisation region bHLH)[1: 443-492]
PPIs At1g56650-At1g63650 (*), At1g56650-At5g41315 (*), At1g63650-At1g63650 (*), At1g63650-At5g24520,
At1g63650-At5g35550 (*), At1g63650-At5g41315 (*), At5g24520-At5g35550, At5g24520-At5g41315,
At5g35550-At5g35550 (*), At5g41315-At5g41315 (*)
DDIs IPR001005-IPR001005, IPR001005-IPR001092, IPR001005-IPR009057, IPR001092-IPR001092,
IPR001092-IPR009057, IPR009057-IPR009057
cluster 16 (number of members: 6)
members At2g22810
IPR004838(Aminotransferases, class-I, pyridoxal-phosphate-binding site)[1:270-283] IPR004839(Aminotransferase, class I and II)[1: 40-425]
At3g49700
IPR004838(Aminotransferases, class-I, pyridoxal-phosphate-binding site)[1: 269-282]
IPR004839(Aminotransferase, class I and II) [1: 40-424]
At4g08040
IPR004838(Aminotransferases, class-I, pyridoxal-phosphate-binding site)[1: 264-277] IPR004839(Aminotransferase, class I and II) [1: 38-419]
At4g26200
IPR004838(Aminotransferases, class-I, pyridoxal-phosphate-binding site)[1:282-295]
IPR004839(Aminotransferase, class I and II)[1: 54-437]
At4g37770
IPR004838(Aminotransferases, class-I, pyridoxal-phosphate-binding site)[1: 269-282]
IPR004839(Aminotransferase, class I and II)[1: 40-424]
At5g65800
IPR004838(Aminotransferases, class-I, pyridoxal-phosphate-binding site)[1:269-282]
IPR004839(Aminotransferase, class I and II)[1: 40-424]
PPIs At2g22810-At2g22810 (*), At2g22810-At3g49700 (*), At2g22810-At4g08040 (*), At2g22810-At4g26200 (*),
At2g22810-At4g37770 (*), At2g22810-At5g65800 (*), At3g49700-At3g49700 (*), At3g49700-At4g08040 (*),
At3g49700-At4g26200 (*), At3g49700-At4g37770 (*), At3g49700-At5g65800 (*), At4g08040-At4g08040 (*)
At4g08040-At4g37770 (*), At4g08040-At5g65800 (*), At4g26200-At4g26200 (*), At4g26200-At4g37770 (*),
At4g37770-At4g37770 (*), At4g37770-At5g65800 (*), At5g65800-At5g65800 (*)
DDIs IPR004838-IPR004838, IPR004838-IPR004839, IPR004839-IPR004839
cluster 17 (number of members: 5)
members At1g48150
IPR002100(Transcription factor, MADS-box)[1: 15-63]
At1g60880
IPR002100(Transcription factor, MADS-box)[1: 3-51]
At1g60920
IPR002100(Transcription factor, MADS-box)[1: 3-40]
At5g38740
IPR002100(Transcription factor, MADS-box)[1: 27-69]
At5g65330
IPR002100(Transcription factor, MADS-box)[1: 1-56]

| PPIs | $\begin{aligned} & \text { At1g48150-At1g60880 (*), } \\ & \text { At1g60880-At5g65330 (*), } \end{aligned}$ | At1g48150-At1g60920 (*), At1g60920-At5g65330 (*), | At1g48150-At5g38740 (*), At5g38740-At5g65330 (*) | At1g48150-At5g65330 (*), |
| :---: | :---: | :---: | :---: | :---: |
| DDIs | IPR002100-IPR002100 |  |  |  |

cluster 18 (number of members: 9)
members At1g75390
At2g 18160
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1: 29-92]
At3g 54620
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1: 229-283]
At3g62420
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1:23-86]
At4g02640
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1: 221-276]
At4g 34590
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1: 25-88]
At5g24800
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1: 120-166]
At5g28770
IPR004827(Basic-leucine zipper (bZIP) transcription factor)[1: 156-171]
At5g49450
PPIs At1g75390-At1g75390, At1g75390-At2g18160, At1g75390-At3g54620, At1g75390-At3g62420,
At1g75390-At4g02640, At1g75390-At4g34590, At1g75390-At5g24800, At1g75390-At5g28770,
At1g75390-At5g49450, At2g18160-At2g18160 (*), At2g18160-At3g54620 (*), At2g18160-At3g62420 (*),
At2g18160-At4g02640 (*), At2g18160-At4g34590 (*), At2g18160-At5g24800 (*), At2g18160-At5g28770 (*), At2g18160-At5g49450, At3g54620-At3g54620 (*), At3g54620-At3g62420 (*), At3g54620-At4g02640 (*),





```
cluster 39 (number of members: 3)
    members At1g09130
        At4g25370
                            IPR004176(Clp, N-terminal)[2:97-149, 184-226]
                    ATCG00670
    PPIs At1g09130-At4g25370, At1g09130-ATCG00670, At4g25370-ATCG00670
cluster 40 (number of members: 3)
    members At1g09960
            At1g22710
            At2g02860
            PPIs At1g09960-At1g22710, At1g09960-At2g02860, At1g22710-At2g02860
            DDIs
cluster 41 (number of members: 3)
    members At1g16970
                IPR003034(DNA-binding SAP)[1: 585-619]
                IPR005160(Ku70/Ku80 C-terminal arm)[1: 453-548]
                    IPR006164(DNA helicase, ATP-dependent, Ku type)[1: 272-475]
                    At1g48050
                            IPR005160(Ku70/Ku80 C-terminal arm)[1: 427-513]
                            IPR006164(DNA helicase, ATP-dependent, Ku type)[1: 278-419]
            At4g13870
                IPR002562(3&apos;-5&apos; exonuclease)[1: 105-283]
            PPIs At1g16970-At1g48050, At1g16970-At4g13870 (*), At1g48050-At4g13870 (*)
            DDIs IPR002562-IPR005160, IPR002562-IPR006164
cluster 42 (number of members: 3)
    members At1g21700
                IPR001005(SANT, DNA-binding)[1: 399-447]
                    IPR007526(SWIRM)[1:176-274]
                    IPR009057(Homeodomain-like)[1: 394-447]
            At2g33610
                    IPR001005(SANT, DNA-binding)[1: 224-272]
                    IPR007526(SWIRM)[1:48-145]
                    IPR009057(Homeodomain-like)[1: 221-274]
            At2g46020
                    IPR000330(SNF2-related)[1: 984-1291]
                    IPR001487(Bromodomain)[1: 1900-2007]
            IPR001650(DNA/RNA helicase, C-terminal)[1: 1338-1422]
    PPIs At1g21700-At2g33610 (*), At1g21700-At2g46020 (*), At2g33610-At2g33610 (*), At2g33610-At2g46020 (*)
        DDIs IPR001005-IPR001005, IPR001005-IPR007526, IPR001005-IPR009057, IPR001487-IPR007526,
        IPR007526-IPR007526, IPR009057-IPR009057
cluster 43 (number of members: 3)
    members At1g26310
                            IPR002100(Transcription factor, MADS-box)[1: 2-95]
                            IPR002487(Transcription factor, K-box)[1: 90-180]
                    At5g03790
                            IPR009057(Homeodomain-like)[1:48-135]
            At5g61850
        PPIs At1g26310-At5g03790, At1g26310-At5g61850, At5g03790-At5g61850
            DDIs
cluster 44 (number of members: 3)
    members At1g49950
                            IPR001005(SANT, DNA-binding)[1:4-59]
                            IPR009057(Homeodomain-like)[1: 1-68]
                            IPR011991(Winged helix repressor DNA-binding)[1: 110-196]
            At3g49850
                    IPR001005(SANT, DNA-binding)[1:4-59]
                            IPR009057(Homeodomain-like)[1: 2-62]
                            IPR011991(Winged helix repressor DNA-binding)[1: 114-198]
            At5g67580
                            IPR001005(SANT, DNA-binding) [1: 4-59]
                            IPR009057(Homeodomain-like)[1: 1-68]
    IPR011991(Winged helix repressor DNA-binding)[1: 116-200] 
    At3g49850-At5g67580 (*), At5g67580-At5g67580 (*)
        DDIs IPR001005-IPR001005, IPR001005-IPR009057, IPR001005-IPR011991, IPR009057-IPR009057,
    IPR011991-IPR011991
cluster 45 (number of members: 3)
    members At1g50640
                IPR001471(Pathogenesis-related transcriptional factor and ERF, DNA-binding)[1: 26-77]
                    At2g45640
                    At2g45640
```

```
    PPIs
    At1g50640-At2g45640, At1g50640-At4g38130, At2g45640-At4g38130
cluster 46 (number of members: 3)
    members At2g06210
        IPR001440(Tetratricopeptide TPR-1)[8:128-161, 162-195, 278-311, 316-349, 422-449, 531-564, 686-714, 722-755]
    At4g29830
            IPR001680(WD40 repeat)[7:2-46, 49-88, 91-131, 134-186, 189-229, 232-271, 274-319]
            IPR011046(WD40 repeat-like)[1: 1-320]
            At5g61150
            PP1s At2g06210-At4g29830, At2g06210-At5g61150, At4g29830-At5g61150
            DDIs
cluster 47 (number of members: 3)
    members At3g12810
                            IPR000330(SNF2-related)[1: 539-825]
                            IPR001005(SANT, DNA-binding)[1: 1670-1729]
                            IPR001650(DNA/RNA helicase, C-terminal)[1: 1110-1188]
            IPR006562(HSA)[1: 36-107]
            At3g33520
            At5g37055
            IPR007529(Zinc finger, HIT-type)[1: 130-159]
            At3g12810-At3g33520, At3g12810-At5g37055, At3g33520-At5g37055
        DDIs
cluster 48 (number of members: 3)
    members At4g05420
                IPR000408(Regulator of chromosome condensation, RCC1)[1: 503-513]
                    IPR004871(Cleavage and polyadenylation specificity factor, A subunit, C-terminal)[1: 740-1027]
                            IPR011046(WD40 repeat-like)[1: 287-922]
            At4g10180
                IPR008928(Six-hairpin glycosidase-like)[1:319-467]
            At5g58760
                            IPR001680(WD40 repeat)[5: 160-200, 203-244, 302-341, 344-386, 459-491]
                            IPR001878(Zinc finger, CCHC-type)[3:0-0, 76-99, 101-117]
                            IPR001878(Zinc finger, CCHC-type)[3:0-0,
        PPIs At4g05420-At4g10180, At4g05420-At5g58760 (*), At4g10180-At5g58760
        DDIs IPR001680-IPR011046, IPR011046-IPR011046
cluster 49 (number of members: 3)
    members At4g18290
            IPR000595(Cyclic nucleotide-binding)[1: 395-484]
            At4g22200
                    IPR000595(Cyclic nucleotide-binding)[1:394-513]
                            IPR002110(Ankyrin)[1: 540-709]
            At5g46240
                IPR000595(Cyclic nucleotide-binding)[1:377-496]
            PPIs At4g18290-At4g22200(*), At4g18290-At5g46240 (*), At4g22200-At5g46240 (*)
            DDIs IPR000595-IPR000595, IPR000595-IPR002110
cluster 50 (number of members: 3)
    members At1g23860
                    IPR000504(RNA recognition motif, RNP-1)[1: 2-73]
                    IPR001878(Zinc finger, CCHC-type)[1:90-106]
            At1g55310
            IPR000504(RNA recognition motif, RNP-1)[1: 37-110]
            At3g50670
                IPR000504(RNA recognition motif, RNP-1)[1: 138-216]
    PPIs At1g23860-At1g55310 (*), At1g23860-At3g50670 (*), At1g55310-At1g55310 (*), At1g55310-At3g50670 (*)
                    PR000504-IPR000504, IPR000504-IPR001878
cluster 51 (number of members: 4)
    members At3g56150
                            IPR000717(Proteasome component region PCI)[1: 705-793]
                            IPR008905(Eukaryotic translation initiation factor 3 subunit 8,N-terminal)[1:34-636]
                            IPR011991(Winged helix repressor DNA-binding)[1: 699-782]
            At3g57290
                IPR000717(Proteasome component region PCI)[1: 336-422]
            At3g61140
                IPR000717(Proteasome component region PCI)[1: 329-412]
            At4g14110
                            At3g56150-At3g57290(*), At3g56150-At3g61140(*), At3g56150-At4g14110, At3g57290-At3g61140 (*),
                            At3g57290-At4g14110, At3g61140-At3g61140(*), At3g61140-At4g14110, At4g14110-At4g14110
                    IPR000717-IPR000717, IPR000717-IPR011991
cluster 52 (number of members: 3)
    members At2g32950
                            IPR001680(WD40 repeat)[6: 355-399, 410-449, 459-491, 493-534, 545-585, 633-672]
                                    IPR001841(Zinc finger, RING-type)[1:52-90]
```




IPR002487(Transcription factor, K-box)[1: 75-174]
At3g54340
IPR002100(Transcription factor, MADS-box)[1: 1-60]
IPR002487(Transcription factor, K-box) [1: 84-174]
At5g20240
IPR002100(Transcription factor, MADS-box)[1: 1-60]
PR002487(Transcription factor, K-box)[1: 84-170]



[^0]:    *Doctoral Dissertation, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0661206, February 4, 2010

[^1]:    NA/unknown.

